

С. И. Носков, С. В. Беляев

ОЦЕНКА НЕПРОТИВОРЕЧИВОСТИ КЛАСТЕРНОЙ ЛИНЕЙНОЙ РЕГРЕССИОННОЙ МОДЕЛИ

Ключевые слова: кластерная линейная регрессионная модель, непротиворечивость, метод наименьших модулей, задача линейно-булева программирования, химическая промышленность, средняя и максимальная относительные ошибки аппроксимации, критерий смещения.

В работе дан краткий обзор публикаций по различным аспектам построения, исследования и применения кластерной линейной регрессии в рамках прикладного регрессионного анализа. В частности, рассмотрены: процедура тестирования данных для определения приемлемого уровня их кластеризации; семейство моделей смесей со случайными ковариатами, вложенных в линейную кластерно-взвешенную модель; процедура функциональной кластеризации для классификации суточных кривых нагрузки; анализ кластеризованных данных с помощью моделей частичной линейной регрессии; адаптивный выпуклый метод кластеризации для одновременного выполнения сегментации данных и подгонки модели для обобщенных линейных моделей; эффективные вычислительные алгоритмы для бутстрэпа линейных регрессионных моделей с кластеризованными данными. В статье приведена форма задачи линейно-булева программирования, к которой может быть сведена задача оценивания параметров кластерной линейной регрессии, если за расстояние между расчетными и фактическими значениями зависимой переменной принято манхэттенское расстояние. В качестве основного критерия оценки непротиворечивости кластерной линейной регрессионной модели предложено использовать критерий смещения, применяемый в качестве внешнего в методе группового учета аргументов. При оценке непротиворечивости можно применять также среднюю и максимальную относительные ошибки аппроксимации, вычисленные для каждого кластера. Отмечено, что использование кластерной линейной регрессии для пространственных данных имеет большее прикладное значение, чем для временных рядов, как это сделано в работе, поскольку кластеризация таких данных открывает большие возможности в решении широкого круга практических задач анализа и прогнозирования, а ее результаты хорошо интерпретируемы. Построена кластерная линейная регрессионная модель развития химической промышленности Российской Федерации, произведена оценка ее непротиворечивости.

S. I. Noskov, S. V. Belyaev

ASSESSMENT OF THE CONSISTENCY OF A CLUSTER LINEAR REGRESSION MODEL

Keywords: cluster linear regression model, consistency, least absolute value method, linear Boolean programming problem, chemical industry, average and maximum relative approximation errors, bias criterion.

The paper provides a brief overview of publications on various aspects of construction, research and application of cluster linear regression in the framework of applied regression analysis. In particular, the following are considered: a procedure for testing data to determine an acceptable level of their clustering; a family of mixture models with random covariates embedded in a linear cluster-weighted model; a procedure for functional clustering for classifying daily load curves; analysis of clustered data using partial linear regression models; an adaptive convex clustering method for simultaneously performing data segmentation and model fitting for generalized linear models; efficient computational algorithms for bootstrapping linear regression models with clustered data. The paper presents a form of the linear Boolean programming problem to which the problem of estimating the parameters of cluster linear regression can be reduced if the Manhattan distance is taken as the distance between the calculated and actual values of the dependent variable. It is proposed to use the bias criterion used as an external one in the group argument accounting method as the main criterion for assessing the consistency of the cluster linear model. When assessing consistency, it is also possible to use the average and maximum relative approximation errors calculated for each cluster. It is noted that the use of cluster linear regression for spatial data has a greater applied significance than for time series, as is done in the work, since clustering of such data opens up wide opportunities in solving a wide range of practical problems of analysis and forecasting, and its results are well interpretable. A cluster linear model of the development of the chemical industry of the Russian Federation is constructed, and its consistency is assessed.

Введение

Разработка кластерной линейной регрессии (КЛР) (англ. clustered linear regression (CLR)) — активно развиваемый в последние десятилетия подход к аппроксимации данных, основанный на использовании нескольких линейных функций, заданных на различных участках (частях, фрагментах) исходной информации, называемых кластерами (см., в частности, [1,2]). Различные аспекты построения, исследования и применения КЛР активно изучаются в рамках прикладного регрессионного анализа. Так, в работе [3] предлагается последовательная процедура тестирования данных для определения приемлемого

уровня их кластеризации. В [4] представлено новое семейство из двенадцати моделей смесей со случайными ковариатами, вложенных в линейную кластерно-взвешенную модель. В статье [5] используется процедура функциональной кластеризации для классификации суточных кривых нагрузки с использованием прошлых данных о спросе на отопление в системе централизованного теплоснабжения. Исследование [6] посвящено анализу кластеризованных данных с помощью моделей частичной линейной регрессии. При этом оценка параметра наклона производится методом наименьших квадратов с использованием кусочно-

локальной полиномиальной аппроксимации непараметрической компоненты.

В работе [7] изучаются методы профильного ядра и обратной подгонки в частично линейных моделях для кластеризованных данных. Показывается, что для коррелированных данных обратная подгонка часто дает большую асимптотическую дисперсию, чем метод профильного ядра. В [8] предлагается адаптивный выпуклый метод кластеризации для одновременного выполнения сегментации данных и подгонки модели для обобщенных линейных моделей. Этот метод сегментирует точки данных с использованием объединенного штрафа для учета сходства в структурах моделей. Он гарантирует, что точки данных, имеющие одну и ту же структуру модели, группируются в один сегмент. В статье [9] обсуждаются эффективные вычислительные алгоритмы для бутстрэпа линейных регрессионных моделей с кластеризованными данными. При этом все вычисления основаны на матрицах и векторах, которые содержат суммы квадратов и перекрестных произведений для наблюдений в каждом кластере, которые должны быть вычислены только один раз до начала цикла бутстрэпа.

Интересные результаты, связанные с кластерным анализом, получены в работах: [10] (выделение однородных по климатическим условиям групп регионов с использованием кластерного анализа), [11] (применение кластерного анализа при классификации моногородов по кадастровой стоимости земель), [12] (кластерный анализ инновационной деятельности субъектов РФ на основе выделенных критериев), [13] (решение проблемы прогнозирования стоимости жилой недвижимости в Российской Федерации на основе методов кластеризации), [14] (оценка масс элементов космических аппаратов с использованием кластерного анализа), [15] (кластеризация временных рядов при прогнозировании покупательского спроса).

Некоторые критерии непротиворечивости кластерной линейной регрессионной модели

Формально задача построения КЛР может быть поставлена следующим образом. Пусть задана выборка исходных данных (X, y) , где $X - (n \times m)$ – матрица значений независимых (входных) переменных с компонентами $X = \|x_{ki}\|$, $k = \overline{1, n}$, $i = \overline{1, m}$, $y = (y_1, \dots, y_n)^T$ – вектор значений зависимой (выходной) переменной. Предположим, у исследователя есть основания полагать, что закономерности влияния переменных x_i , $i = \overline{1, m}$ на выходную переменную y различаются на разных r подвыборках (кластерах) исходной выборки данных (X^j, y^j) , $j = \overline{1, r}$, где в матрицы X^j и векторы y^j будут включены соответственно строки и компоненты с номерами из индексных множеств $P^j \subset \{1, 2, \dots, n\}$:

$$\bigcup_{j=1}^r P^j = \{1, 2, \dots, n\}, \quad P^i \cap P^j = \emptyset, \quad i \neq j.$$

Тогда кластерная линейная регрессия имеет вид:

$$y_k = \alpha_0^j + \sum_{i=1}^m \alpha_i^j x_{ki} + \varepsilon_k^j, \quad j = \overline{1, r}, \quad k \in P^j, \quad (1)$$

где α_i^j , $j = \overline{1, r}$, $i = \overline{0, m}$ – подлежащие оцениванию параметры, ε_k^j , $j = \overline{1, r}$, $k \in P^j$ – ошибки аппроксимации. При этом число кластеров r задано, а их состав заранее неизвестен.

Введем в рассмотрение матрицу оценок параметров $A = \|\alpha_i^j\|$, $j = \overline{1, r}$, $i = \overline{0, m}$ регрессии (1) и матрицу булевых индикаторных переменных $\Xi = \|\sigma_{kj}\|$, $k = \overline{1, n}$, $j = \overline{1, r}$:

$$\sigma_{kj} = \begin{cases} 1, & k \in P^j \\ 0, & \text{в противном случае.} \end{cases}$$

Тогда, если в качестве расстояния между вычисленными и реальными значениями зависимой переменной определено городское расстояние, оценка параметров и формирование составов индексных множеств P^j , $j = \overline{1, r}$ кластерной линейной регрессии (1) осуществляется посредством решения следующей задачи с булевыми переменными:

$$G(A, \Xi) = \sum_{k=1}^n \sum_{j=1}^r \sigma_{kj} |\varepsilon_k^j|. \quad (2)$$

Нетрудно видеть, что, воспользовавшись приемами сведения задачи оценивания параметров обычной линейной регрессии по методу наименьших модулей (МНМ) к задаче линейного программирования (см., в частности, [16, 17]), а также отражая присутствие в задаче (2) булевых переменных [17], она может быть сведена к следующей задаче линейно-булева программирования (ЛБП):

$$\alpha_0^j + \sum_{i=1}^m \alpha_i^j x_{ki} - M \sigma_{kj} + u_k \geq y_k - M, \quad j = \overline{1, r}, \quad k = \overline{1, n}, \quad (3)$$

$$\alpha_0^j + \sum_{i=1}^m \alpha_i^j x_{ki} + M \sigma_{kj} - u_k \leq y_k + M, \quad j = \overline{1, r}, \quad k = \overline{1, n}, \quad (4)$$

$$\sum_{j=1}^r \sigma_{kj} = 1, \quad k = \overline{1, n}, \quad (5)$$

$$\sigma_{kj} \in \{0, 1\}, \quad j = \overline{1, r}, \quad k = \overline{1, n}, \quad (6)$$

$$u_k \geq 0, \quad k = \overline{1, n}, \quad (7)$$

$$\sum_{i=1}^n u_k \rightarrow \min. \quad (8)$$

Здесь M – заранее заданное большое положительное число.

После решения задачи ЛБП (3) – (8) состав множеств P^j , $j = \overline{1, r}$ формируется по правилу:

$$\sigma_{kj} = 1 \Rightarrow k \in P^j.$$

После построения КЛР (1) возникает естественный вопрос – в какой степени она непротиворечива? Или, другими словами, в какой мере близки ее частные линейные модели, каждая из которых задана на «своем» кластере?

Для оценки непротиворечивости КЛР можно воспользоваться внешним критерием в рамках известного метода группового учета аргументов (МГУА), названным критерием смещения $n_{\text{см}}$ (см., в частности [18]).

Введем обозначения:

$$\tilde{y}_k^j = \alpha_0^j + \sum_{i=1}^m \alpha_i^j x_{ki}, \quad j = \overline{1, r}, \quad k = \overline{1, n}.$$

Тогда $n_{\text{см}}$ можно рассчитать по формуле:

$$n_{\text{см}} = 2 \sum_{i=1}^{r-1} \sum_{j=i+1}^r n_{\text{см}}^{ij} / (r(r-1)),$$

где

$$n_{\text{см}}^{ij} = \sum_{k=1}^n (\tilde{y}_k^i - \tilde{y}_k^j)^2 / \sum_{k=1}^n y_k^2.$$

Разумеется, чем меньше значение $n_{\text{см}}$, тем более непротиворечива КЛР (1).

Степень непротиворечивости КЛР можно оценить также, если сравнить средние относительные ошибки аппроксимации E^j , $j = \overline{1, r}$ на каждом кластере. Введем обозначения:

$$\hat{y}_k^j = \alpha_0^j + \sum_{i=1}^m \alpha_i^j x_{ki}, j = \overline{1, r}, k \in P^j.$$

Тогда эти ошибки рассчитываются по формуле:

$$E^j = 100\% \sum_{k \in P^j} \left| \frac{\hat{y}_k^j - y_k}{y_k} \right| / |P^j|, j = \overline{1, r},$$

где $|P^j|$ - мощность множества P^j (число элементов в нем).

Наконец, эту степень можно оценить также путем сравнения максимальных относительных ошибок аппроксимации H^j , $j = \overline{1, r}$ на каждом кластере по следующей формуле:

$$H^j = 100\% \max_{k \in P^j} \left| \frac{\hat{y}_k^j - y_k}{y_k} \right|.$$

Оценка непротиворечивости кластерной линейной регрессионной модели развития химической промышленности Российской Федерации

В работе [19] представлена, в частности, линейная регрессионная модель развития химической промышленности Российской Федерации, построенная с помощью МНМ по статистической информации за 2003 – 2022 гг.:

$$y_k = 318742 + 6747.77 x_k + \varepsilon_k, k = \overline{1, 20}, \quad (9)$$

$$E = 8.7,$$

где E - средняя относительная ошибка аппроксимации, y - объем отгруженных товаров собственного производства, выполненных работ и услуг собственными силами по видам экономической деятельности для химической промышленности, млн. руб.; x - инвестиции в основной капитал по видам экономической деятельности для химической промышленности, млрд. руб.

Построим по тем же данным кластерную регрессионную модель (1), решив задачу ЛБП (3) – (8) при $r = 2$. В результате получим:

$$y_k = 219544 + 7853.05 x_k + \varepsilon_k, k \in P^1, \quad (10)$$

$$P^1 = \{1, 2, 3, 4, 5, 6, 7, 11, 19, 20\}, E^1 = 5.54, H^1 = 13.7,$$

$$y_k = 848294 + 5150.68 x_k + \varepsilon_k, k \in P^2, \quad (11)$$

$$P^2 = \{8, 9, 10, 12, 13, 14, 15, 16, 17, 18\}, E^2 = 3.35, H^2 = 10.7.$$

Значение критерия смещения $n_{см} = 0.04$. Оно указывает на выраженную близость расчетных значений зависимой переменной, полученных с помощью векторов оценок параметров α^1 и α^2 , а, следовательно, на непротиворечивость КЛР (10), (11). При этом оценки параметров частных линейных регрессий в ней существенно различаются, как и с построенной по всей выборке моделью (9). Следует также отметить, что модель (11), судя по значениям средней и максимальной относительных ошибок параметров, несколько точнее, чем модель (10). Обращает на себя внимание тот факт, что в состав кластера P^1 в основном вошли наблюдения из первой половины исходной выборки, а в состав кластера P^2 – из второй. Но, вместе с тем, эти кластеры не делят выборку ровно пополам.

Безусловно, использование кластерной линейной регрессии для пространственных данных имеет большее прикладное значение, чем для временных рядов, как это сделано выше, поскольку кластеризация таких данных открывает особенно большие возможности в решении широкого круга практических задач анализа и прогнозирования, а ее результаты хорошо интерпретируемы.

Заключение

В работе приведен вид задачи линейно-булева программирования, к которой сводится задача оценивания параметров кластерной линейной регрессионной модели, если за расстояние между вычисленными и реальными значениями зависимой переменной взято городское расстояние. В качестве основного критерия оценки непротиворечивости этой модели предложено использовать критерий смещения, применяемый в методе группового учета аргументов. При этой оценке можно использовать также среднюю и максимальную относительные ошибки аппроксимации, рассчитанные на каждом кластере. Построена кластерная линейная регрессия развития химической промышленности Российской Федерации, оценена степень ее непротиворечивости.

Литература

1. Helmuth Späth. *Computing*, **22**, 4, 367–373 (1979), DOI: 10.1007/BF02265317.
2. Qiang Long, Adil Bagirov, Sona Taheri, Nargiz Sultanova, and Xue Wu. *ACM Trans. Knowl. Discov. Data*, **17**, 3, 1–54 (2023), DOI: 10.1145/3550074.
3. James G. MacKinnon, Morten Ørregaard Nielsen, Matthew D. Webb. *Journal of Econometrics*, **235**, 2, 2027–2056 (2023), DOI: 10.1016/j.jeconom.2023.03.005.
4. Salvatore Ingrassia, Simona C. Minotti, Antonio Punzo. *Computational Statistics & Data Analysis*, **71**, 159–182 (2014), DOI: 10.1016/j.csda.2013.02.012.
5. Aldo Goia, Caterina May, Gianluca Fusai. *International Journal of Forecasting*, **26**, 4, 700–711 (2010), DOI: 10.1016/j.ijforecast.2009.05.015.
6. Kani Chen, Zhezhen Jin. *Journal of the American Statistical Association*, **101**, 473, 195–204 (2006).
7. Zonghui Hu, Naisyin Wang and Raymond J. Carroll. *Biometrika*, **91**, 2, 251–262 (2004), DOI: 10.1093/biomet/91.2.251.
8. Shuyu Chu, Huijing Jiang, Zhengliang Xue, Xinwei Deng. *Technometrics*, **63**, 2, 171–183 (2021), DOI: 10.1080/00401706.2020.1733094.
9. James G. MacKinnon. *Econometrics and Statistics*, **26**, 52–71 (2023), DOI: 10.1016/j.ecosta.2021.11.009.
10. Гордиенко А.С., Лозинская А.М., Тетерина Д.В., Шенкман Е.А. *Известия Российской академии наук. Энергетика*, **1**, 15–27 (2019), DOI: 10.1134/S0002331019010072.
11. Ковязин В.Ф., Лепихина О.Ю., Зимин В.П. *Известия Томского политехнического университета. Инжиниринг георесурсов*, **328**, 3, 6–13 (2017).
12. Дмитриев Ю., Фраймович Д., Мищенко З. *Вестник Института экономики Российской академии наук*, **3**, 79–87 (2013).
13. Леонов Ю.А., Филиппова Л.Б., Мартыненко А.А., Живодовский И.И. *Известия Тульского государственного университета. Технические науки*, **2**, 87–93 (2024), DOI: 10.24412/2071-6168-2024-2-87-88.
14. Куренков В.И., Кучеров А.С., Якищук А.А. *Известия Самарского научного центра Российской академии наук*,

- 20, 1, (81), 45-53 (2018), DOI: 10.24411/1990-5378-2018-00002.
15. Пивкин К.С. *Математическое и компьютерное моделирование в экономике, страховании и управлении рисками*, 2, 165-169 (2017).
 16. Носков С.И. *Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных*. Облформпечать, Иркутск, 1996. 320 с.
 17. Dimitris Bertsimas and Romy Shioda. *Operations Research*, 55, 2, 252–271 (2007).
 18. Ивахненко А.Г., Юрачковский Ю.П. *Моделирование сложных систем по экспериментальным данным*. Радио и связь, Москва, 1987. 120 с.
 19. Носков С.И., Беляев С.В. *Вестник Технологического университета*, 27, 11, 206-210 (2024), DOI: 10.55421/1998-7072_2024_27_11_206.

References

1. Helmuth Späth. *Computing*, 22, 4, 367–373 (1979), DOI: 10.1007/BF02265317.
2. Qiang Long, Adil Bagirov, Sona Taheri, Nargiz Sultanova, and Xue Wu. *ACM Trans. Knowl. Discov. Data*, 17, 3, 1–54 (2023), DOI: 10.1145/3550074.
3. James G. MacKinnon, Morten Ørregaard Nielsen, Matthew D. Webb. *Journal of Econometrics*, 235, 2, 2027-2056 (2023), DOI: 10.1016/j.jeconom.2023.03.005.
4. Salvatore Ingrassia, Simona C. Minotti, Antonio Punzo. *Computational Statistics & Data Analysis*, 71, 159-182 (2014), DOI: 10.1016/j.csda.2013.02.012.
5. Aldo Goia, Caterina May, Gianluca Fusai. *International Journal of Forecasting*, 26, 4, 700-711 (2010), DOI: 10.1016/j.ijforecast.2009.05.015.
6. Kani Chen, Zhezhen Jin. *Journal of the American Statistical Association*, 101, 473, 195-204 (2006).
7. Zonghui Hu, Naisyin Wang and Raymond J. Carroll. *Biometrika*, 91, 2, 251-262 (2004), DOI: 10.1093/biomet/91.2.251.
8. Shuyu Chu, Huijing Jiang, Zhengliang Xue, Xinwei Deng. *Technometrics*, 63, 2, 171-183 (2021), DOI: 10.1080/00401706.2020.1733094.
9. James G. MacKinnon. *Econometrics and Statistics*, 26, 52-71 (2023), DOI: 10.1016/j.ecosta.2021.11.009.
10. Gordienko A.S., Lozinskaya A.M., Teterina D.V., Shenkman E.A. *Bulletin of the Russian Academy of Sciences. Power Engineering*, 1, 15-27 (2019), DOI: 10.1134/S0002331019010072.
11. Kovyazin V.F., Lepikhina O.Yu., Zimin V.P. *Bulletin of Tomsk Polytechnic University. Georesources Engineering*, 328, 3, 6-13 (2017).
12. Dmitriev Yu., Fraimovich D., Mishchenko Z. *Bulletin of the Institute of Economics of the Russian Academy of Sciences*, 3, 79-87 (2013).
13. Leonov Yu.A., Filippova L. B., Martynenko A.A., Zhivodovsky I.I. *News of Tula State University. Engineering sciences*, 2, 87-93 (2024), DOI: 10.24412/2071-6168-2024-2-87-88.
14. Kurenkov V.I., Kucherov A.S., Yakishchik A.A. *News of the Samara Scientific center of the Russian Academy of Sciences*, 20, 1, (81), 45-53 (2018), DOI: 10.24411/1990-5378-2018-00002.
15. Pivkin K.S. *Mathematical and computer modeling in economics, insurance and risk management*, 2, 165-169 (2017).
16. Nосков S.I. *Technology of modeling objects with unstable functioning and uncertainty in data*. Oblinformpechat, Irkutsk, 1996. 320 p.
17. Dimitris Bertsimas and Romy Shioda. *Operations Research*, 55, 2, 252–271 (2007).
18. Ivakhnenko A.G., Yurachkovsky Yu.P. *Modeling of complex systems based on experimental data*. Radio and Communications, Moscow, 1987. 120 p.
19. Nосков S.I., Belyaev S.V. *Herald of Technological University*, 27, 11, 206-210 (2024), DOI: 10.55421/1998-7072_2024_27_11_206.

© С. И. Носков - д-р техн. наук, проф., профессор кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения (ИГУПС), Иркутск, Россия, sergey.noskov.57@mail.ru; С. В. Беляев - магистрант кафедры «Информационные системы и защита информации», ИГУПС, bsv2001@list.ru.

© S. I. Noskov – Doctor of Sciences (Technical Sci.), Professor of the department of Information Systems and Information Security (ISIS), Irkutsk State Transport University (ISTU), Irkutsk, Russia, sergey.noskov.57@mail.ru; S. V. Belyaev – Master-student of the ISIS department, ISTU, bsv2001@list.ru.