Инженерный вестник Дона, №5 (2024)
ivdon.ru/ru/magazine/archive/n5y2024/9188

Оценивание регрессионных моделей с мультиарной операцией модуль методом наименьших модулей

М. П. Базилевский

Иркутский государственный университет путей сообщения, Иркутск

Аннотация: В статье рассмотрена исследованная ранее линейная по факторам и нелинейная по параметрам модель модульной регрессии, содержащая унарные операции модуль. За счёт применения бинарных, тернарных, …, l-арных операций модуль впервые предложено обобщение модульной регрессии. Рассмотрен частный случай обобщения – регрессия с мультиарной операцией модуль. Задача точного оценивания такой модели с помощью метода наименьших модулей сведена к задаче частично-булевого линейного программирования. По встроенным в эконометрический пакет Gretl данным о производительности ферм построена классическая линейная регрессия и модульная регрессия с мультиарной операцией. Качество аппроксимации предложенной модульной регрессии оказалось выше, чем качество линейной модели.
Ключевые слова: регрессионный анализ, модульная регрессия, метод наименьших модулей, мультиарная операция модуль, задача частично-булевого линейного программирования.

Методы машинного обучения [1] активно развиваются в настоящее время. Эффективным инструментом обработки статистических данных считается регрессионный анализ [2]. Построению регрессионных моделей за последние годы посвящено много зарубежных научных работ. Так, например, в [3] строилась регрессионная модель для прогнозирования заболеваемости COVID-19 в Индии, в [4] – для прогнозирования уровня инфляции по данным Центрального банка Индонезии, в [5] – для прогнозирования выбросов углекислого газа, связанных с ископаемым топливом, в странах БРИКС, в [6] оценена модель полупараметрической регрессии для исследования распределенных энергетических ресурсов в Китае. Не отстают от зарубежных и отечественные исследования в области регрессионного анализа. Например, в [7] задача отбора информативных регрессоров в линейной регрессии, оцениваемой с помощью метода наименьших квадратов, формализована в виде задачи частично-булевого линейного программирования (далее ЧБЛП), что гораздо эффективнее на

практике метода «всех регрессий» [8]. В [9] исследованы модели полносвязной линейной регрессии, в которых все истинные исходные переменные связаны между собой линейными функциональными зависимостями, что позволяет проводить моделирование в условиях мультиколлинеарности [10]. В [11,12] с помощью регрессионного анализа решены конкретные прикладные задачи технического характера.

В работах [13,14] предложена и исследована линейная по факторам и нелинейная по параметрам модель модульной регрессии следующего вида:

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j \left| x_{ij} - \lambda_j \right| + \varepsilon_i, \qquad i = \overline{1,n}, \tag{1}$$

где y_i, x_{i1}, …, x_{il}, $i = \overline{1,n}$ – значения объясняемой переменной y и объясняющих переменных x_1, x_2, …, x_l; n – количество наблюдений; α_0, α_j, λ_j, $j = \overline{1,l}$ – неизвестные параметры; ε_i, $i = \overline{1,n}$ – ошибки аппроксимации.

В модели (1), например, операцию вида $\left|x_1 - \lambda_1\right|$ назовём унарной операцией модуль. Тогда можно ввести бинарную операцию модуль, например, $\left|x_1 - \lambda_1 - \lambda_2 x_2\right|$, тернарную операцию модуль, например, $\left|x_1 - \lambda_1 - \lambda_2 x_2 - \lambda_3 x_3\right|$ и т.д. Из этого следует, что модульная регрессия (1) является частным случаем более сложной конструкции следующего вида:

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j^{(1)} \left| x_{ij} - \lambda_{j0}^{(1)} \right| + \sum_{j=1}^{C_l^2} \alpha_j^{(2)} \left| x_{i,\mu_{j1}^{(1)}} - \lambda_{j0}^{(2)} - \lambda_{j1}^{(2)} \cdot x_{i,\mu_{j2}^{(1)}} \right| +$$

$$+ \sum_{j=1}^{C_l^3} \alpha_j^{(3)} \left| x_{i,\mu_{j1}^{(2)}} - \lambda_{j0}^{(3)} - \lambda_{j1}^{(3)} \cdot x_{i,\mu_{j2}^{(2)}} - \lambda_{j2}^{(3)} \cdot x_{i,\mu_{j3}^{(2)}} \right| + \ldots +$$

$$+ \alpha_1^{(l)} \left| x_{i1} - \lambda_0^{(l)} - \sum_{k=2}^{l} \lambda_{k-1}^{(l)} \cdot x_{ik} \right| + \varepsilon_i, \qquad i = \overline{1,n}, \tag{2}$$

где $\mu_{j1}^{(1)}$, $\mu_{j2}^{(1)}$ ($j = \overline{1,C_l^2}$) – элементы j-й строки матрицы $\mathrm{M}^{(1)}$ размера $C_l^2 \times 2$, содержащей по строкам все сочетания индексов объясняющих переменных

по два; $\mu_{j1}^{(2)}$, $\mu_{j2}^{(2)}$, $\mu_{j3}^{(2)}$ ($j=\overline{1,C_l^3}$) – элементы j-й строки матрицы $\mathrm{M}^{(2)}$ размера $C_l^3 \times 3$, содержащей по строкам все сочетания индексов объясняющих переменных по три и т.д.; α_0, $\alpha_j^{(1)}$, $j=\overline{1,l}$, $\alpha_j^{(2)}$, $j=\overline{1,C_l^2}$, …, $\alpha_1^{(l)}$, $\lambda_{j0}^{(1)}$, $j=\overline{1,l}$, $\lambda_{j0}^{(2)}$, $\lambda_{j1}^{(2)}$, $j=\overline{1,C_l^2}$, …, $\lambda_j^{(l)}$, $j=\overline{0,l-1}$ – неизвестные параметры.

Назовём модель (2) модульной регрессией с унарными, бинарными, тернарными, …, (l-1)-арными, l-арной операциями модуль.

Рассмотрим упрощенную форму модели (2) – модульную регрессию с мультиарной (l-арной) операцией модуль следующего вида:

$$y_i = \alpha_0 + \alpha_1 \left| x_{i1} - \lambda_0 - \sum_{k=2}^{l} \lambda_{k-1} \cdot x_{ik} \right| + \varepsilon_i, \qquad i=\overline{1,n}. \tag{3}$$

Оценивание неизвестных параметров модульной регрессии (3) методом наименьших модулей (далее МНМ) предполагает решение оптимизационной задачи вида:

$$\sum_{i=1}^{n} \left| y_i - \alpha_0 - \alpha_1 \left| x_{i1} - \lambda_0 - \sum_{k=2}^{l} \lambda_{k-1} \cdot x_{ik} \right| \right| \to \min.$$

Пусть $|\alpha_1| = k \geq 0$. Тогда перепишем выражение (3) в виде:

$$y_i = \alpha_0 + (-1)^{\sigma} \left| k \cdot x_{i1} - \beta_0 - \sum_{k=2}^{l} \beta_{k-1} \cdot x_{ik} \right| + \varepsilon_i, \qquad i=\overline{1,n}, \tag{4}$$

где $\beta_0 = k \cdot \lambda_0$, $\beta_1 = k \cdot \lambda_1$, …, $\beta_{l-1} = k \cdot \lambda_{l-1}$; σ – бинарная переменная, которая равна 0, если $\alpha_1 \geq 0$, и 1, если $\alpha_1 < 0$.

В соответствии с приёмом, описанным в работе [13], точные МНМ-оценки модульной регрессии (4) можно получить, решив при $\sigma = 0$ и $\sigma = 1$ следующую задачу ЧБЛП:

$$\sum_{j=1}^{l} (g_i + h_i) \to \min, \tag{5}$$

$$k \cdot x_{i1} - \beta_0 - \sum_{k=2}^{l} \beta_{k-1} \cdot x_{ik} = \mu_i - v_i, \qquad i = \overline{1,n}, \tag{6}$$

$$\mu_i \le M \cdot \delta_i, \qquad i = \overline{1,n}, \tag{7}$$

$$v_i \le M \cdot (1 - \delta_i), \qquad i = \overline{1,n}, \tag{8}$$

$$y_i = \alpha_0 + (-1)^{\sigma} (\mu_i + v_i) + g_i - h_i, \qquad i = \overline{1,n}, \tag{9}$$

$$k \ge 0,\ \mu_i \ge 0,\ v_i \ge 0,\ g_i \ge 0,\ h_i \ge 0, \qquad i = \overline{1,n}, \tag{10}$$

$$\delta_i \in \{0,1\}, \qquad i = \overline{1,n}, \tag{11}$$

где M – большое положительное число;

$$\mu_i = \begin{cases} kx_{i1} - \beta_0 - \sum_{k=2}^{l} \beta_{k-1} \cdot x_{ik}, \text{ если } kx_{i1} - \beta_0 - \sum_{k=2}^{l} \beta_{k-1} \cdot x_{ik} \ge 0, \\ 0, \text{ если } kx_{i1} - \beta_0 - \sum_{k=2}^{l} \beta_{k-1} \cdot x_{ik} < 0, \end{cases} \qquad i = \overline{1,n},$$

$$v_i = \begin{cases} 0, \text{ если } kx_{i1} - \beta_0 - \sum_{k=2}^{l} \beta_{k-1} \cdot x_{ik} \ge 0, \\ -kx_{i1} + \beta_0 + \sum_{k=2}^{l} \beta_{k-1} \cdot x_{ik}, \text{ если } kx_{i1} - \beta_0 - \sum_{k=2}^{l} \beta_{k-1} \cdot x_{ik} < 0, \end{cases}, \quad i = \overline{1,n},$$

$$g_i = \begin{cases} y_i - \alpha_0 - (-1)^{\sigma} (\mu_i + v_i), \text{ если } y_i - \alpha_0 - (-1)^{\sigma} (\mu_i + v_i) \ge 0, \\ 0, \text{ если } y_i - \alpha_0 - (-1)^{\sigma} (\mu_i + v_i) < 0, \end{cases} \qquad i = \overline{1,n},$$

$$h_i = \begin{cases} 0, \text{ если } y_i - \alpha_0 - (-1)^{\sigma} (\mu_i + v_i) \ge 0, \\ -y_i + \alpha_0 + (-1)^{\sigma} (\mu_i + v_i), \text{ если } y_i - \alpha_0 - (-1)^{\sigma} (\mu_i + v_i) < 0, \end{cases} \qquad i = \overline{1,n},$$

$$\delta_i = \begin{cases} 0, \text{ если } kx_{i1} - \beta_0 - \sum_{k=2}^{l} \beta_{k-1} \cdot x_{ik} < 0, \\ 1, \text{ если } kx_{i1} - \beta_0 - \sum_{k=2}^{l} \beta_{k-1} \cdot x_{ik} \ge 0. \end{cases}$$

Решив задачу ЧБЛП (5) – (11) при $\sigma = 0$ и $\sigma = 1$, из двух нужно выбрать регрессионную модуль с наименьшей величиной суммы модулей остатков.

Для демонстрации работоспособности предложенного способа точного оценивания модульной регрессии (3) были использованы встроенные в эконометрический пакет Gretl статистические данные (файл data9-5.gdt на вкладке Ramanathan) о затратах и объеме произведенной продукции сельского хозяйства в США с 1948 по 1993 годы, т.е. объем выборки n составил 46. В качестве выходной переменной y выступает производительность фермы (output), а входными переменными выбраны следующие:

x_1 – сельскохозяйственный труд (farm labor);

x_2 – оборудование длительного пользования (machines);

x_3 – используемые сельскохозяйственные химикаты (fert).

Сначала по этим данным была оценена с помощью МНМ классическая линейная регрессионная модель:

$$\tilde{y} = 91,684 - 0,143x_1 - 0,456x_2 + 0,521x_3. \quad (12)$$

Сумма модулей остатков регрессии (12) составляет 183,1978.

Затем при M=10000 и $\sigma = 0$ с использованием пакета LPSolve IDE была решена задача ЧБЛП (5) – (11). Результатом её решения является следующее оцененное уравнение модульной регрессии:

$$\tilde{y} = 51,578 + 0,1636 \cdot \left| x_1 - 287,063 + 3,512x_2 - 3,554x_3 \right|. \quad (13)$$

Сумма модулей остатков регрессии (13) равна 160,8055, что меньше, чем у модели (12). Другими словами, качество аппроксимации предложенной в статье модульной регрессии выше, чем качество классической линейной модели.

После чего задача ЧБЛП (5) – (11) была решена при M=10000 и $\sigma = 1$. Оцененное уравнение модульной регрессии в такой ситуации имело вид:

$$\tilde{y} = 108 - 0,1431 \cdot \left| x_1 + 113,975 + 3,1847x_2 - 3,642x_3 \right|. \quad (14)$$

Сумма модулей остатков регрессии (14) составила 183,1978, т.е. эта характеристика такая же, как и у линейной регрессии (12). Можно заметить, что это происходит потому, что выражение под знаком модуля в уравнении (14) для любого наблюдения выборки неотрицательно, поэтому знак модуля можно просто опустить и прийти к уравнению (12). В любом случае, оцененной с помощью МНМ модульной регрессией (3) признаётся модель (13).

Достоинство модульной регрессии (3) ещё и в том, что задача ЧБЛП для её оценивания с помощью МНМ (5) – (11) содержит существенно меньше булевых переменных, чем задача ЧБЛП для оценивания регрессии (1).

Литература

1. Molnar C. Interpretable machine learning. Lulu. com, 2020. 368 p.

2. Pardoe I. Applied regression modeling. John Wiley & Sons, 2020. 325 p.

3. Pandey G., Chaudhary P., Gupta R., Pal S. SEIR and regression model based COVID-19 outbreak predictions in India // arXiv preprint. arXiv:2004.00958. 2020. URL: arxiv.org/ftp/arxiv/papers/2004/2004.00958.pdf.

4. Dharma F., Shabrina S., Noviana A., Tahir M., Hendrastuty N., Wahyono W. Prediction of Indonesian inflation rate using regression model based on genetic algorithms // Jurnal Online Informatika. 2020. No. 5(1). Pp. 45-52.

5. Karakurt I., Aydin G. Development of regression models to forecast the CO2 emissions from fossil fuels in the BRICS and MINT countries // Energy. 2023. Vol. 263. P. 125650.

6. Xu B., Luo Y., Xu R., Chen J. Exploring the driving forces of distributed energy resources in China: Using a semiparametric regression model // Energy. 2021. Vol. 236. P. 121452.

7. Базилевский М.П. Отбор информативных регрессоров с учётом мультиколлинеарности между ними в регрессионных моделях как задача

частично-булевого линейного программирования // Моделирование, оптимизация и информационные технологии. 2018. Т. 6. № 2 (21). С. 104-118.

8. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998. 1005 с.

9. Базилевский М.П. Исследование двухфакторной модели полносвязной линейной регрессии // Моделирование, оптимизация и информационные технологии. 2019. Т. 7. № 2 (25). С. 80-96.

10. Shrestha N. Detecting multicollinearity in regression analysis // American Journal of Applied Mathematics and Statistics. 2020. Vol. 8. No. 2. Pp. 39-42.

11. Король В.И., Ланкин М.В., Горбатенко Н.И. Регрессионная модель погрешностей аппроксимации кривой тока для измерения магнитных характеристик // Инженерный вестник Дона. 2022. № 7. URL: ivdon.ru/ru/magazine/archive/n7y2022/7825.

12. Баклагин В.Н. Регрессионная модель изменения ледовитости Белого моря // Инженерный вестник Дона. 2018. № 2. URL: ivdon.ru/ru/magazine/archive/N2y2018/4825.

13. Базилевский М.П., Ойдопова А.Б. Оценивание модульных линейных регрессионных моделей с помощью метода наименьших модулей // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. 2023. № 45. С. 130-146.

14. Базилевский М.П. Программное обеспечение для оценивания модульных линейных регрессий // Информационные и математические технологии в науке и управлении. 2023. № 3 (31). С. 136-146.

References

1. Molnar C. Interpretable machine learning. Lulu. com, 2020. 368 p.

2. Pardoe I. Applied regression modeling. John Wiley & Sons, 2020. 325 p.

3. Pandey G., Chaudhary P., Gupta R., Pal S. arXiv preprint arXiv:2004.00958. 2020. URL: arxiv.org/ftp/arxiv/papers/2004/2004.00958.pdf.

4. Dharma F., Shabrina S., Noviana A., Tahir M., Hendrastuty N., Wahyono W. Jurnal Online Informatika. 2020. № 5(1). Pp. 45-52.

5. Karakurt I., Aydin G. Energy. 2023. Vol. 263. P. 125650.

6. Xu B., Luo Y., Xu R., Chen J. Energy. 2021. Vol. 236. P. 121452.

7. Bazilevskiy M.P. Modelirovanie, optimizacija i informacionnye tehnologii. 2018. Vol. 6. No. 2 (21). Pp. 104-118.

8. Ajvazjan S.A., Mhitarjan V.S. Prikladnaja statistika i osnovy jekonometriki. Moscow: JuNITI, 1998. 1005 p.

9. Bazilevskiy M.P. Modelirovanie, optimizacija i informacionnye tehnologii. 2019. Vol. 7. No. 2 (25). Pp. 80-96.

10. Shrestha N. American Journal of Applied Mathematics and Statistics. 2020. Vol. 8. No. 2. Pp. 39-42.

11. Korol' V.I., Lankin M.V., Gorbatenko N.I. Inzhenernyj vestnik Dona. 2022. № 7. URL: ivdon.ru/ru/magazine/archive/n7y2022/7825.

12. Baklagin V.N. Inzhenernyj vestnik Dona. 2018. № 2. URL: ivdon.ru/ru/magazine/archive/N2y2018/4825.

13. Bazilevskiy M.P., Oydopova A.B. Vestnik Permskogo nacional'nogo issledovatel'skogo politehnicheskogo universiteta. Jelektrotehnika, informacionnye tehnologii, sistemy upravlenija. 2023. № 45. Pp. 130-146.

14. Bazilevskiy M.P. Informacionnye i matematicheskie tehnologii v nauke i upravlenii. 2023. № 3 (31). Pp. 136-146.

Дата поступления: 16.03.2024

Дата публикации: 22.04.2024