

Сравнительный анализ разных подходов к оценке параметров регрессионных моделей с помощью метода наименьших модулей на примере моделирования стоимости домов по выборке большого объема

М. П. Базилевский

Иркутский государственный университет путей сообщения, Иркутск

Аннотация: Статья посвящена исследованию проблемы оценки неизвестных параметров линейных регрессий с помощью метода наименьших модулей. Рассмотрено два известных подхода к идентификации регрессионной модели. Первый из них основан на решении задачи линейного программирования, второй, называемый итерационным методом наименьших квадратов, позволяет получить приближенное решение задачи. Для возможности тестирования итеративного метода в пакете Gretl разработана специальная программа. Для проведения вычислительных экспериментов использована выборка о стоимости домов и влияющих на неё факторах, содержащая 20640 наблюдений. В итоге наилучшие результаты были показаны при использовании встроенной в Gretl команды quantreg, реализующей алгоритм Фриша-Ньютона. Второй результат показал итеративный метод, а третий – решенная в пакете LPSolve задача линейного программирования.

Ключевые слова: регрессионный анализ, метод наименьших модулей, линейное программирование, итеративный метод наименьших квадратов, метод вариационно-взвешенных квадратичных приближений.

Инструменты регрессионного анализа на сегодняшний день находят широкое применение при решении инженерных задач. Так, например, в [1] регрессионный анализ применен для моделирования погрешностей кривой тока при измерении магнитных характеристик, а в [2] – для моделирования качества нефтепродуктов в зависимости от технологических параметров.

Для оценки неизвестных параметров линейных регрессионных моделей к настоящему времени разработано множество различных методов. Пожалуй, самым известным и простым в применении считается метод наименьших квадратов (МНК) [3,4], для реализации которого достаточно сформировать и решить систему линейных алгебраических уравнений. Однако, как отмечено в [5], оценка МНК резко реагирует на наличие выбросов в исходных данных. Поэтому возникает необходимость пользоваться на практике робастными методами, которые малочувствительны (устойчивы) к выбросам. Одним из

таких методов является метод наименьших модулей (МНМ) [6,7], состоящий в решении задачи минимизации суммы модулей ошибок регрессии. Отметим, что в [8,9] на основе МНК и МНМ предложен метод двухкритериального оценивания регрессионных моделей.

Реализация МНМ возможна следующими способами.

1. С помощью симплекс-метода, поскольку, как известно [10], задача МНМ-оценивания линейной регрессии сводится к задаче линейного программирования. Благодаря этому приёму относительно недавно появились новые нелинейные спецификации регрессионных моделей [11,12], для оценки неизвестных параметров которых с помощью МНМ применяется аппарат частично целочисленного линейного программирования.

2. С помощью итеративного МНК [5]. Этот способ предполагает вычисление весов на каждой итерации и применение взвешенного МНК до тех пор, пока не будет достигнута требуемая точность МНМ-оценок. В [5] отмечено, что итеративный МНК в книге [13] назван методом вариационно-взвешенных квадратичных приближений. И в той же книге [13] доказана сходимость итеративного МНК в зависимости от выбранной метрики. В [14] этот метод называют алгоритмом Вейсфельда.

3. С помощью алгоритма спуска по узловым прямым [15]. Как резюмируют авторы работы [15], в ходе проведенных экспериментов алгоритм спуска по узловым прямым значительно выиграл у алгоритмов прямой и двойственной задач линейного программирования.

Цель настоящей работы состоит в проведении сравнительного анализа эффективности симплекс-метода и итеративного МНК при МНМ-оценке параметров линейной регрессии. Причём, для реализации указанных приёмов требуется разработать соответствующие алгоритмы и компьютерные программы.

Пусть по выборочной совокупности объема n строится следующая модель множественной линейной регрессии зависимости объясняемой (выходной) переменной y , значений которой y_i , $i = \overline{1, n}$, от l объясняющих (входных) переменных $x_1, x_2, \dots, x_l$, значения которых $x_{i1}, x_{i2}, \dots, x_{il}$, $i = \overline{1, n}$:

$$y_i = \alpha_0 + \sum_{j=1}^l \alpha_j x_{ij} + \varepsilon_i, \quad i = \overline{1, n}, \quad (1)$$

где α_j , $j = \overline{0, l}$ – неизвестные параметры; ε_i , $i = \overline{1, n}$ – ошибки регрессии.

Для нахождения МНМ-оценок неизвестных параметров линейной регрессии (1) требуется решить задачу:

$$\sum_{i=1}^n |\varepsilon_i| \rightarrow \min. \quad (2)$$

1. Задача линейного программирования.

В [10] показано, что задача (2) сводится к следующей задаче линейного программирования:

$$\sum_{i=1}^n (u_i + v_i) \rightarrow \min, \quad (3)$$

$$y_i = \alpha_0 + \sum_{j=1}^l \alpha_j x_{ij} + \varepsilon_i + u_i - v_i, \quad i = \overline{1, n}, \quad (4)$$

$$u_i \geq 0, v_i \geq 0, \quad i = \overline{1, n}, \quad (5)$$

где переменная u_i равна остатку регрессии e_i , если он положителен, а v_i равна модулю остатка, если он отрицателен, причём, $u_i \cdot v_i = 0$ для любого i . Остаток регрессии e_i есть разность между фактическим значением выходной переменной y_i и её расчетным по модели значением $\tilde{y}_i$.

Задача (3) – (5) может быть решена с помощью симплекс-метода.

2. Итеративный МНК.

В [5] показано, что численное решение задачи (2) в общем случае может быть получено по следующему алгоритму. На нулевой итерации оцениваем регрессию (1), например, с помощью МНК, получаем вектор оценок $\tilde{\alpha}^{(0)} = (\tilde{\alpha}_0^{(0)}, \tilde{\alpha}_1^{(0)}, \dots, \tilde{\alpha}_l^{(0)})$. На первой итерации находим вектор весов $w_i^{(1)}$ по формуле:

$$w_i^{(1)} = |e_i^{(0)}|^{-1}, \quad (6)$$

где $e_i^{(0)}$ – остатки регрессии, оцененной на нулевой итерации.

Затем, реализуя взвешенный МНК с весами $w_i^{(1)}$, находим следующее значение вектора оценок $\tilde{\alpha}^{(1)}$ и т.д. В [13] доказана сходимость итеративного МНК.

Как справедливо замечено в [14], если один из остатков равен 0, то на этом шаге вычисление весового коэффициента по формуле (6) становится невозможным. Для решения этой проблемы используется регуляризация, при которой весовая функция строится следующим образом [13]:

$$w_i^{(k)} = \begin{cases} 1 / |e_i^{(k-1)}|, & \text{если } |e_i^{(k-1)}| \neq 0, \\ M, & \text{если } |e_i^{(k-1)}| = 0, \end{cases} \quad k = 1, 2, \dots, \quad (7)$$

где M – произвольное положительное число (желательно большое, поскольку это убыстряет процесс вычислений).

Для интеграции формулы (7) в программный код перепишем её, как это сделано в [14], в виде:

$$w_i^{(k)} = \begin{cases} 1 / |e_i^{(k-1)}|, & \text{если } |e_i^{(k-1)}| \geq \Delta, \\ M, & \text{если } |e_i^{(k-1)}| < \Delta, \end{cases} \quad k = 1, 2, \dots \quad (8)$$

где Δ – достаточно малое положительное число.

Рассмотренный итеративный МНК был реализован автором в виде программы для пакета Gretl. Предварительно необходимо задать 3 параметра:

точность ε , большое и малое числа M и Δ для (8). Итерационный процесс продолжается до тех пор, пока сумма модулей разностей оценок на двух соседних итерациях не станет меньше или равной точности ε . Программа работает по следующему алгоритму.

Итерация 0. С помощью команды *ols* находим вектор МНК-оценок $\tilde{\alpha}^{(0)}$ регрессии (1).

Итерация 1. Вычисляем остатки $e_i^{(0)}$, $i = \overline{1, n}$. По формуле (7) находим веса $w_i^{(1)}$, по которым с помощью команды *wls* через взвешенный МНК находим вектор оценок $\tilde{\alpha}^{(1)}$. Если $\sum |\tilde{\alpha}^{(1)} - \tilde{\alpha}^{(0)}| \leq \Delta$, то оценки найдены. Иначе переходим к следующей итерации.

Итерация 2. Вычисляем остатки $e_i^{(1)}$, $i = \overline{1, n}$. По формуле (7) находим веса $w_i^{(2)}$, по которым с помощью команды *wls* через взвешенный МНК находим вектор оценок $\tilde{\alpha}^{(2)}$. Если $\sum |\tilde{\alpha}^{(2)} - \tilde{\alpha}^{(1)}| \leq \Delta$, то оценки найдены. Иначе переходим к следующей итерации. И так далее.

Заметим, что разработанная программа может использоваться не только для нахождения МНМ-оценок линейной регрессии, но и для L_ν -оценивания [5], предполагающего решение задачи $\sum_{i=1}^n |\varepsilon_i|^\nu \rightarrow \min$ ($1 \leq \nu < 2$).

Было решено протестировать следующие алгоритмы МНМ-оценки линейной регрессии (1): 1) с использованием встроенной команды *lad* пакета Gretl; 2) за счет решения задачи линейного программирования симплекс-методом в пакете LPSolve; 3) с использованием разработанной для пакета Gretl программы, реализующей итеративный МНК. Для тестирования были использованы статистические данные объема $n = 20640$ из источника [16] о стоимости домов в разных округах Калифорнии. Выходная переменная y – медианная стоимость домов; входная переменная x_2 – географическая широта

округа; x_3 – медианный возраст жилья; x_5 – общее количество спален; x_6 – число жителей; x_8 – медианный доход. Выборка содержит и другие входные переменные, но они были исключены из-за сильной корреляции с вышеперечисленными факторами. Для тестирования использовался обычный персональный компьютер с процессором AMD Ryzen 3 4300U (тактовая частота 2.7 ГГц) и 16 Гб оперативной памяти.

На оценку модели с помощью встроенной команды *lad* в пакете Gretl ушло 121,6 секунды. При этом уравнение регрессии выглядит следующим образом:

$$\begin{aligned}\tilde{y} = & 134908 - 5201,48x_2 + 1894,28x_3 + \\ & + 122,694x_5 - 36,299x_6 + 45739,1x_8.\end{aligned}\quad (9)$$

Сумма модулей остатков *SAE* модели (9) составляет 1150909771,015.

Как отмечено в описании команды *lad*, она реализует симплекс алгоритм Барродейла-Робертса. И там же указано, что этот метод может работать медленно на больших выборках данных, поэтому вместо *lad* лучше пользоваться командой *quantreg*, которая использует более быстрый алгоритм Фриша-Ньютона. С помощью команды *quantreg* была получена та же модель (9), но на этой ушло всего 0,076 секунды.

На решение задачи линейного программирования (3) – (5) в LPSolve ушло 66,2 секунды. Причём, на загрузку модели в память программы после запуска потребовалось 6,98 секунд, на предварительной решение – 0,05 секунды, а на симплекс-метод – 59,19 секунд. В результате была построена та же модель (9).

Результаты работы программы, реализующей в Gretl итеративный МНК, представлены в таблице 1. В ней в первом столбце приведен номер эксперимента, во втором – точность ε , в третьем – число итераций, обеспечивших точность, в четвертом – время решения задачи в секундах, в

пятом – значение критерия SAE. МНМ-оценивание проводилось при $M = 100000$ и $\Delta = 0,000001$.

Таблица 1

Результаты тестирования итеративного МНК в Gretl

№	ε	Число итераций	Время, с	Критерий SAE
1	10	55	1,565	1150910701,75
2	1	85	2,438	1150909836,61
3	0,1	106	3,032	1150909777,34
4	0,01	261	7,309	1150909771,36

По таблице 1 видно, что при уменьшении ε число итераций возрастает, поэтому увеличивается и время решения задачи. Но при этом значение критерия SAE становится всё ближе к его значению для модели (9). При $\varepsilon = 10$ разница между ними составляет 930,73, при $\varepsilon = 1$ – 65,59, при $\varepsilon = 0,1$ – 6,32, а при $\varepsilon = 0,01$ – 0,34. Оцененная при $\varepsilon = 0,01$ модель регрессии имеет следующий вид:

$$\begin{aligned}\tilde{y} = & 134988 - 5203,19x_2 + 1894,19x_3 + \\ & + 122,698x_5 - 36,307x_6 + 45738,0x_8.\end{aligned}\quad (10)$$

Как видно, оценки моделей (9) и (10) различаются незначительно.

Таким образом, наилучший результат (0,076 секунды) показан при использовании команды *quantreg* в Gretl, на втором месте (7,309 секунды) итеративный МНК, на третьем месте результат (66,2 секунды), полученный при решении задачи линейного программирования в LPSolve. Наихудший результат (121,6 секунды) показан при использовании команды *lad*.

Литература

1. 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.

2. Зиануров А.Ш., Шумихин А.Г. Аналитический обзор исследований по разработке и применению калибровочных моделей для поточного анализатора качества нефтепродуктов // Инженерный вестник Дона. 2021. № 10. URL: ivdon.ru/ru/magazine/archive/n10y2021/7217.

3. Cinelli C., Ferwerda J., Hazlett C. Sensemakr: Sensitivity analysis tools for OLS in R and Stata // Observational Studies. 2024. Vol. 10. No. 2. Pp. 93-127.

4. Merlin M.L., Chen Y. Analysis of the factors affecting electricity consumption in DR Congo using fully modified ordinary least square (FMOLS), dynamic ordinary least square (DOLS) and canonical cointegrating regression (CCR) estimation approach // Energy. 2021. Vol. 232. P. 121025.

5. Демиденко Е.З. Линейная и нелинейная регрессии. М.: Финансы и статистика, 1981. 303 с.

6. Liu Z., Yang Y. Least absolute deviations estimation for uncertain regression with imprecise observations // Fuzzy Optimization and Decision Making. 2020. Vol. 19. Pp. 33-52.

7. Yang X., Park G.K., Hu Y. Least absolute deviations estimation for uncertain autoregressive model // Soft Computing. 2020. Vol. 24. Pp. 18211-18217.

8. Базилевский М.П. Двухкритериальное оценивание линейных регрессионных моделей методами наименьших модулей и квадратов // International Journal of Open Information Technologies. 2024. Т. 12. № 6. С. 76-81.

9. Базилевский М.П. Исследование метода двухкритериального оценивания линейных регрессионных моделей // Цифровые модели и решения. 2024. Т. 3. № 4. С. 79-90.

10. Wagner H.M. Linear programming techniques for regression analysis // Journal of the American Statistical Association. 1959. Vol. 54. No. 285. Pp. 206-212.
11. Базилевский М.П. Оценивание регрессионных моделей с мультиарной операцией модуль методом наименьших модулей // Инженерный вестник Дона. 2024. № 5. URL: ivdon.ru/ru/magazine/archive/n5y2024/9188.
12. Базилевский М.П. Оценивание с помощью метода наименьших модулей регрессионных моделей с целочисленными функциями пол и потолок // International Journal of Open Information Technologies. 2024. Т. 12. № 10. С. 56-61.
13. Мудров В.И., Кушко В.Л. Метод наименьших модулей. М.: Знание, 1971. 64 с.
14. Андрейчик Н.Л., Горяинов В.Б. Сравнительный анализ устойчивости различных методов оценивания параметров билинейной авторегрессионной модели // Вестник Московского государственного технического университета им. Н.Э. Баумана. Серия «Естественные науки». 2022. № 6 (105). С. 4-16.
15. Голованов О.А., Тырсин А.Н. Спуск по узловым прямым и симплекс-алгоритм – два варианта регрессионного анализа на основе метода наименьших модулей // Заводская лаборатория. Диагностика материалов. 2024. Т. 90. № 5. С. 79-87.
16. Knowledge Extraction based on Evolutionary Learning. URL: sci2s.ugr.es/keel/dataset.php?cod=83.

References

1. Korol' V.I., Lankin M.V., Gorbatenko N.I. Inzhenernyj vestnik Dona. 2022. No. 7. URL: ivdon.ru/ru/magazine/archive/n7y2022/7825.
 2. Zianurov A.Sh., Shumikhin A.G. Inzhenernyj vestnik Dona. 2021. No. 10. URL: ivdon.ru/ru/magazine/archive/n10y2021/7217.
-

3. Cinelli C., Ferwerda J., Hazlett C. *Observational Studies*. 2024. Vol. 10. No. 2. Pp. 93-127.
4. Merlin M.L., Chen Y. *Energy*. 2021. Vol. 232. P. 121025.
5. Demidenko E.Z. *Lineynaya i nelineynaya regressii [Linear and nonlinear regressions]*. Moskva: Finansy i statistika, 1981. 303 p.
6. Liu Z., Yang Y. *Fuzzy Optimization and Decision Making*. 2020. Vol. 19. Pp. 33-52.
7. Yang X., Park G.K., Hu Y. *Soft Computing*. 2020. Vol. 24. Pp. 18211-18217.
8. Bazilevskiy M.P. *International Journal of Open Information Technologies*. 2024. Vol. 12. No. 6. Pp. 76-81.
9. Bazilevskiy M.P. *Tsifrovye modeli i resheniya*. 2024. Vol. 3. No. 4. Pp. 79-90.
10. Wagner H.M. *Journal of the American Statistical Association*. 1959. Vol. 54. No. 285. Pp. 206-212.
11. Bazilevskiy M.P. *Inzhenernyj vestnik Dona*. 2024. No. 5. URL: ivdon.ru/ru/magazine/archive/n5y2024/9188.
12. Bazilevskiy M.P. *International Journal of Open Information Technologies*. 2024. Vol. 12. No. 10. Pp. 56-61.
13. Mudrov V.I., Kushko V.L. *Metod naimen'shikh moduley [Least absolute deviations method]*. Moskva: Znanie, 1971. 64 p.
14. Andreychik N.L., Goryainov V.B. *Vestnik Moskovskogo gosudarstvennogo tekhnicheskogo universiteta im. N.E. Baumana. Seriya «Estestvennye nauki»*. 2022. No. 6 (105). Pp. 4-16.
15. Golovanov O.A., Tyrsin A.N. *Zavodskaya laboratoriya. Diagnostika materialov*. 2024. Vol. 90. No. 5. Pp. 79-87.
16. *Knowledge Extraction based on Evolutionary Learning*. URL: sci2s.ugr.es/keel/dataset.php?cod=83.

Дата поступления: 21.04.25

Дата публикации: 25.05.2025
