Министерство науки и высшего образования Российской Федерации
федеральное государственное бюджетное образовательное учреждение высшего образования
«Самарский государственный технический университет»
(ФГБОУ ВО «СамГТУ»)

Кафедра «Высшая математика»

Модификации метода наименьших квадратов

название работы

Курсовая работа

Выполнил студент
__1_ курса, 113М группы
Бекиш Егор Павлович

_____
подпись

Руководитель
Доцент, к. ф.-м. н.
Гурская Альбина Валентиновна

_____
подпись

Работа защищена
«____» _____ 202__ г.

Оценка _____

Самара, 202__ г.

Оглавление

Введение

В нынешнее время цифровизации почти не одна сфера, профессия, деятельность не обходится без анализа. Следовательно, поступает большой объем данных, который необходимо обработать и провести оценку. Для этого используется регрессионный анализ, в котором находятся статистические методы, прогнозирующие значение измеряемых величин и их точность.

Одной из особенностей анализа поступающих данных – является погрешность. Так как данные в основном бывают некорректными, т.е. появляется погрешность из-за множества факторов, для учета производится несколько измерения и вычисляется среднее значение.

Хотя, проводя ряд экспериментов, измерений сложно определить значения, которые будут более вероятны. Для решение данной задачи используется метод наименьших квадратов.

Метод наименьших квадратов (МНК) [1] – статистический метод, который используется для нахождения линейной аппроксимации данных. Стоит обратить внимание на то, что, как и было сказано, факторов довольно много, тем самым стандартный метод не может всегда давать наилучшую оценку, поэтому существую модификации, которые могут учитывать условия, в которых находятся данные. Будут рассмотрены следующие методы, позволяющею учитывать специфику:

1. Lasso регрессия [2].
2. Ridge регрессия [3].

1. ОПИСАНИЕ МЕТОДОВ

1.1. Метод наименьших квадратов

В общем случае всегда можно применить стандартную форму метода наименьших квадратов. Суть в том, чтобы найти лучшею аппроксимацию линейной функции вида:

$$y = kx + b \tag{1.1}$$

где k – угловой коэффициент;

x – переменная, принимающая произвольное значение;

b – свободный коэффициент, определяющий пересечение с осью OY.

Для нахождения коэффициентов, которые будут минимизировать минимальное отклонение прямой от исходных данных, необходимо просуммировать квадраты каждого y от x. Алгебраический вид, следующая:

$$MSS = \sum_{i=1}^{n} \left(y_i - P(x_i)\right)^2 \tag{1.2}$$

где P – функция.

После того, как составили сумму, необходимо дифференцировать по каждому коэффициенту функции, чтобы перейти к системе линейных алгебраических уравнений (СЛАУ). Алгебраический вид, следующая:

$$\begin{cases} \dfrac{\partial}{\partial p_1}(MMS) = 0 \\ \dfrac{\partial}{\partial p_2}(MMS) = 0 \\ \cdots \\ \dfrac{\partial}{\partial p_n}(MMS) = 0 \end{cases} \tag{1.3}$$

где p – коэффициенты исходной функции.

Далее, необходимо решить СЛАУ, для этого можно воспользоваться методами их решения: Гаусса, матричный и т.п. В конечном итоге получим коэффициенты, которые необходимо подставить в исходную функцию и отобразить данную прямую.

Для наглядности разберем небольшой пример. Исходные данные будут представлены в таблице 1.1.

Таблица 1.1 – Исходные данные

x	0	1	1	4	7
y	-1	1	2	3	4

Исходная функция: $y = kx + b$.

Общий вид МНК: $f = (-1 - b)^2 + (1 - k - b)^2 + (2 - k - b)^2 + (3 - 4k - b)^2 + (4 - 7k - b)^2$.

Дифференцируем по k и b:

$$\begin{cases} \frac{\partial f}{\partial k} = 134k + 26b - 86 \\ \frac{\partial f}{\partial b} = 26k + 10b - 18 \end{cases} = \begin{cases} 67k + 13b = 43 \\ 13k + 5b = 9 \end{cases}$$

Решаем СЛАУ методом Гаусса:

$$\begin{pmatrix} 67 & 13 & | & 43 \\ 13 & 5 & | & 9 \end{pmatrix} \frac{1}{13} \sim \begin{pmatrix} 67 & 13 & | & 43 \\ 1 & \frac{5}{13} & | & \frac{9}{13} \end{pmatrix} -II * 67 \sim \begin{pmatrix} 0 & -\frac{166}{13} & | & -\frac{44}{13} \\ 1 & \frac{5}{13} & | & \frac{9}{13} \end{pmatrix} -\frac{13}{166} \sim$$

$$\sim \begin{pmatrix} 0 & 1 & | & \frac{22}{83} \\ 1 & \frac{5}{13} & | & \frac{9}{13} \end{pmatrix} -I * \frac{5}{13} \sim \begin{pmatrix} 0 & 1 & | & \frac{22}{83} \\ 1 & 0 & | & \frac{49}{83} \end{pmatrix} \rightarrow \begin{cases} k = \frac{49}{83} \\ b = \frac{22}{83} \end{cases}$$

Полученная прямая:

$$y = \frac{49}{83}x + \frac{22}{83} \quad (1.4)$$

Отобразим данную прямую (рисунок 1.1).

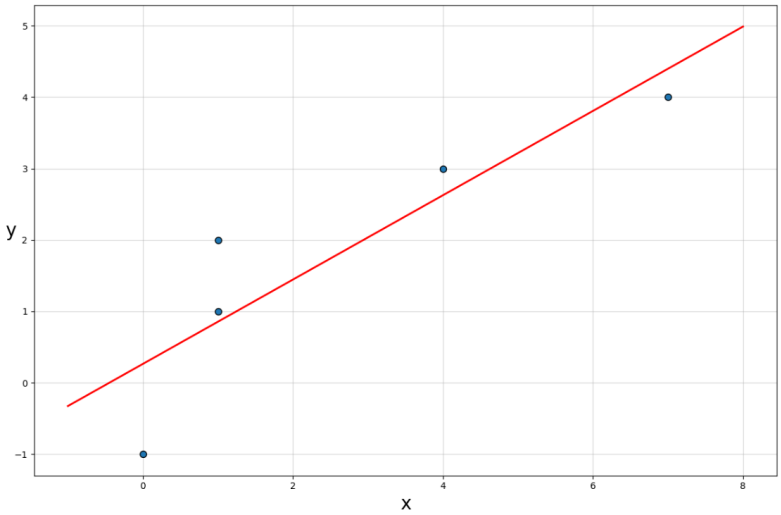

Рисунок 1.1 – аппроксимация $y = kx + b$

Стоит дополнить тем, что функция, т.е. полином может быть любого вида и степени. Выше был рассмотрен простейший вид полинома первой степени, где функция зависела только от одного аргумента. Полином можно задать следующим образом:

$$f(x) = k_1 + k_2x_2 + k_3x_3^2 + \cdots + k_nx^n, \quad (1.5)$$

$$f(x_1, x_2, \ldots, x_n) = k_1 + k_2x_2 + k_3x_3 + \cdots + k_nx_n, \quad (1.6)$$

$$f(x_1, x_2, \ldots, x_n) = k_1x_1^2 + k_2x_2^2 + k_1x_1x_2 + \cdots + k_nx_1^n + k_nx_2^n + \cdots \quad (1.7)$$

Так же рассмотрим примеры с исходными данными с таблицы 1.1 для полинома (1.4) и составим для (1.5) свои исходные данные, которые будут представлены в таблице 1.2.

Таблица 1.2 – Исходные данные

x_1	2	5	7	3	8	4	6
x_2	3	1	4	6	2	5	7
y	8	12	16	5	20	10	11

Исходная функция: $y = ax^2 + bx + c$.

Общий вид МНК: $f = (-1 - c)^2 + (1 - a - b - c)^2 + (2 - a - b - c)^2 + (3 - 16a - 4b - c)^2 + (4 - 49a - 7b - c)^2$.

Дифференцируем по a, b, c:

$$\begin{cases} \frac{\partial f}{\partial a} = 5318a + 818b + 134c - 494 \\ \frac{\partial f}{\partial b} = 818a + 134b + 26c - 86 \\ \frac{\partial f}{\partial c} = 134a + 26b + 10c - 18 \end{cases}$$

Решаем СЛАУ методом Гаусса и получаем следующие значение параметров: $a = -\frac{169}{1392}$; $b = \frac{2017}{1392}$; $c = -\frac{79}{232}$.

Полученная кривая:

$$y = -\frac{169}{1392}x^2 + \frac{2017}{1392}x - \frac{79}{232} \quad (1.8)$$

Полученная кривая отображена на рисунке 1.2.

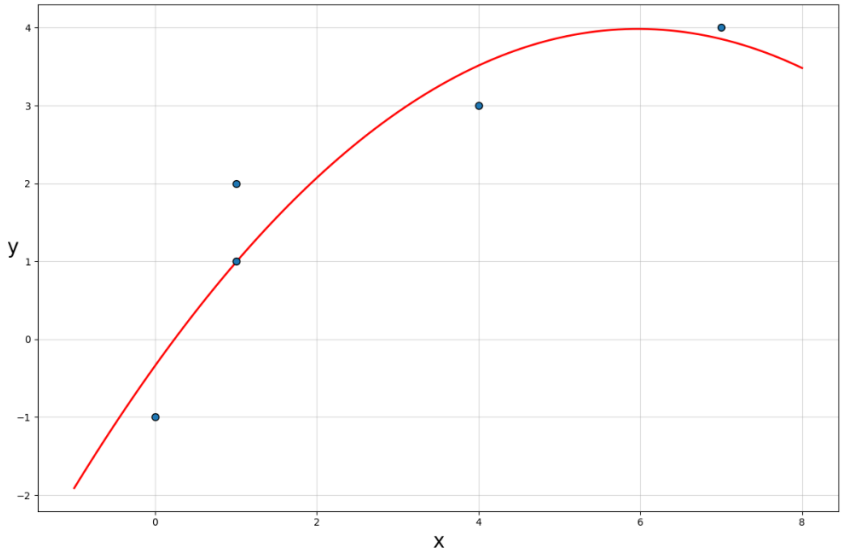

Рисунок 1.2 – аппроксимация $y = ax^2 + bx + c$

Исходная функция: $f(x_1, x_2) = k_1x_1 + k_2x_2 + k_0$

Общий вид МНК: $f = (8 - k_0 - 2k_1 - 3k_2)^2 + (12 - k_0 - 5k_1 - k_2)^2 + (16 - k_0 - 7k_1 - 4k_2)^2 + (5 - k_0 - 3k_1 - 6k_2)^2 + (20 - k_0 - 8k_1 - 2k_2)^2 + (10 - k_0 - 4k_1 - 5k_2)^2 + (11 - k_0 - 6k_1 - 7k_2)^2$

Дифференцируем по k_0, k_1, k_2:

$$\begin{cases} \dfrac{\partial f}{\partial k_0} = 14k_0 + 70k_1 + 56k_2 - 164 \\ \dfrac{\partial f}{\partial k_1} = 70k_0 + 406k_1 + 270k_2 - 938 \\ \dfrac{\partial f}{\partial k_2} = 56k_0 + 270k_1 + 280k_2 - 594 \end{cases}$$

Решаем СЛАУ методом Гаусса и получаем следующие значение параметров: $k_0 = \frac{8629}{1771}$; $k_1 = \frac{499}{252}$; $k_2 = -\frac{191}{252}$.

Полученная плоскость:

$$y = \frac{499}{252}x_1 - \frac{191}{252}x_2 + \frac{8629}{1771} \tag{1.9}$$

Покажем данную плоскость на рисунке 1.3.

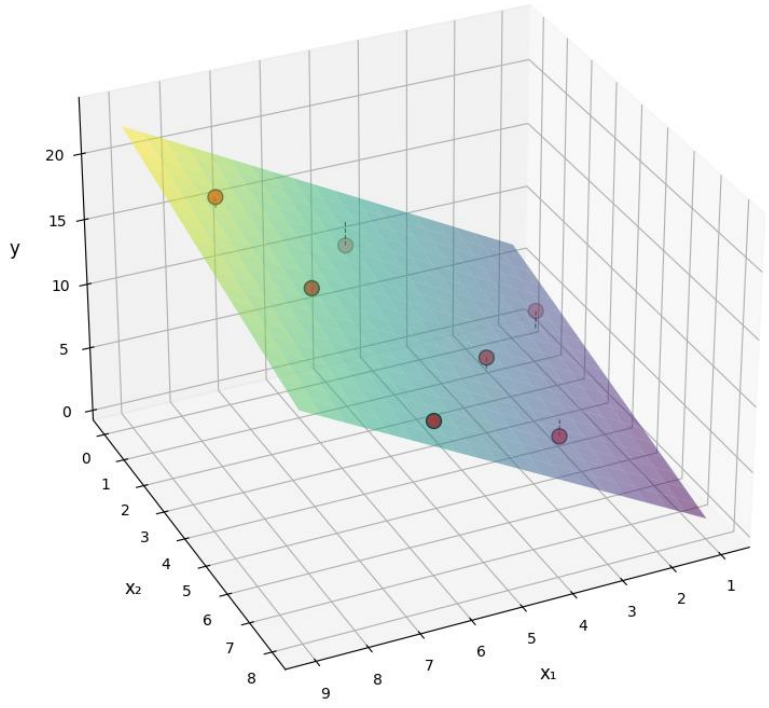

Рисунок 1.3 – аппроксимация $f(x_1, x_2) = k_1x_1 + k_2x_2 + k_0$

Для дальнейшей минимизации методов, которые будут рассмотрены, как модификации, ручной способ будет иметь больше вычислений, поэтому воспользуемся вспомогательными средствами, которые это упростят. Для этого будет использовать язык программирования Python [4] и следующие вспомогательные фреймворки, такие как:

- sklearn [5].

- sympy [6].
- matplotlib [7].
- pandas [8].
- numpy [9].

Реализацию вышеописанных полиномов ручным способом можно увидеть здесь [10]

1.2. Lasso регрессия

МНК можно применять всегда, но будет ли находится оптимальная кривая для поступающих данных – это не всегда так, потому, как данные могут быть некорректны, не всегда будут известны и т.п. Данные факторы приводят к переобучение, что приводит к неточным предсказаниям, кривая получиться не оптимальная. Существует Lasso регрессия, которая устраняет это. Lasso регрессия повышает точность прогнозирования, уменьшая некоторые коэффициенты и устанавливает другие равными нулю за счет штрафа λ. Алгебраический вид, следующий:

$$LSS = \sum_{i=1}^{n} \left(y_i - P(x_i)\right)^2 + \lambda \sum_{j=1}^{m} |w_j| \tag{1.10}$$

где w – какой-либо итеративный алгоритм нахождения.

1.3. Ridge регрессия

Так же для этого используется Ridge регрессия или же метод регуляризации Тихонова (МРТ). Метод помогает бороться с переобучением за счет λ, так называемый штраф. Чем больше значение λ, тем больше штрафуем коэффициенты и их количество, но не сводит их к нулю. Алгебраический вид нахождения коэффициентов имеет следующий вид:

$$B = (X^T X)^{-1} X^T y \tag{1.11}$$

где X – матрица независимых переменных;

y – вектор зависимых переменных.

Тем самым данный метод учитывает корреляцию предсказанных данных. Данный штрафной коэффициент имеет следующий вид:

$$L2 = \|B\|^2 = \sum_{i=1}^{n} B_i^2 \tag{1.12}$$

Так как в данных методах МНК берется за основу, остается добавить в конец (1.2) полученный коэффициент (1.12). Алгебраический вид, следующий:

$$RSS_{L2} = \sum_{i=1}^{n} \left(y_i - P(x_i)\right)^2 + \lambda \sum_{j=1}^{m} B_j^2 \tag{1.13}$$

2. ОПРЕДЕЛЕНИЕ НАИЛУЧШЕЙ МОДЕЛИ

Как и было заявлено в первом разделе – задача заключается в том, что необходимо найди наилучшею аппроксимацию, т.е. расстояние от точек до полученной кривой было минимальным. Следовательно, нужно воспользоваться такой моделью, которая максимально минимизирует это расстояние. Для определения существует несколько ключевых метрик для регрессии:

1. Коэффициент детерминации (R^2);
2. Средняя квадратичная ошибка (MSE);
3. Средняя абсолютная ошибка (MAE).

2.1 Коэффициент детерминации

1. Отображает долю дисперсии модели в общей дисперсии целевой переменной. Мера качества – нормированная квадратичная ошибка. Полученное значение находится в диапазоне от 0 до 1. Если значение близко к 1, то модели хорошо описывает данные, в противном случае, следовательно, плохо. Алгебраический вид, следующий:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - P(x_i)\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.1}$$

где P – целевая функция;

y – фактическое значение;

$\bar{y}$ – среднее арифметическое всех значений y.

2.2 Средняя квадратичная ошибка

Отображает большие ошибки в данных, которые влияют на прогноз. Тем самым, полученное значение говорит о том, что чем оно больше, тем выбранная модель является менее эффективной, в противном случает модель более эффективная. Алгебраический вид, следующий:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - P(x_i)\right)^2 \tag{2.2}$$

где n – количество наблюдений.

2.3 Средняя абсолютная ошибка

Средняя абсолютная ошибка (MAE) – так же, как и MSE отображает большие ошибки, но штрафует сильнее за большие отклонения, следовательно, более чувствительна на сильные отклонения. Алгебраический вид, следующий:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - P(x_i)| \tag{2.3}$$

2.4 Расчет метрик

Посчитаем для данных из таблицы 1.1, какая целевая функция (1.1, 1.5) лучше описывает данные в стандартном МНК.

2.4.1 Целевая функция $y = kx + b$

В первом разделе разобрали пример для текущей целевой функции (1.1) и получили функцию с коэффициентами (1.4). Найдем метрики (2.1 – 2.3).

Найдем для каждого x предсказанный $\hat{y}$ и разницу между начальным и предсказанным:

$$\widehat{y_1} = \frac{49}{83} * 0 + \frac{22}{83} = \frac{22}{83} \qquad (y_1 - \widehat{y_1}) = -1 - \frac{22}{83} = -\frac{105}{83}$$

$$\widehat{y_2} = \frac{49}{83} * 1 + \frac{22}{83} = \frac{71}{83} \qquad (y_2 - \widehat{y_2}) = 1 - \frac{71}{83} = \frac{12}{83}$$

$$\widehat{y_3} = \frac{49}{83} * 1 + \frac{22}{83} = \frac{71}{83} \qquad (y_3 - \widehat{y_3}) = 2 - \frac{71}{83} = \frac{95}{83}$$

$$\widehat{y_4} = \frac{49}{83} * 4 + \frac{22}{83} = \frac{218}{83} \qquad (y_4 - \widehat{y_4}) = 3 - \frac{218}{83} = \frac{31}{83}$$

$$\widehat{y_5} = \frac{49}{83} * 7 + \frac{22}{83} = \frac{365}{83} \qquad (y_5 - \widehat{y_5}) = 4 - \frac{365}{83} = -\frac{33}{83}$$

Найдем среднее арифметическое и разницу между текущим y:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{-1+1+2+3+4}{5} = \frac{9}{5}$$

$$(y_1 - \bar{y}) = -1 - \frac{9}{5} = -\frac{14}{5}$$

$$(y_2 - \bar{y}) = 1 - \frac{9}{5} = -\frac{4}{5}$$

$$(y_3 - \bar{y}) = 2 - \frac{9}{5} = \frac{1}{5}$$

$$(y_4 - \bar{y}) = 3 - \frac{9}{5} = \frac{6}{5}$$

$$(y_5 - \bar{y}) = 4 - \frac{9}{5} = \frac{11}{5}$$

Соберем все вместе:

$$\sum_{i=1}^{n}\left(y_i - P(x_i)\right)^2 = \frac{22244}{6889} \qquad \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{370}{25}$$

$$R^2 = 1 - \frac{\frac{22244}{6889}}{\frac{370}{25}} = 1 - \frac{556100}{2548930} = 1 - 0{,}21817 = 0{,}78183$$

$$MSE = \frac{\frac{22244}{6889}}{5} = \frac{22244}{34445} = 0{,}64578 \qquad MAE = \frac{\frac{276}{83}}{5} = \frac{276}{415} = 0{,}66506$$

2.4.2 Целевая функция $y = ax^2 + bx + c$

Так же разобрали пример для (1.5) и получили функцию со следующими коэффициентами (1.8).

Найдем для каждого x предсказанный $\hat{y}$ и разницу между начальным и предсказанным:

$$\widehat{y_1} = -\frac{169}{1392} * 0 + \frac{2017}{1392} * 0 - \frac{79}{232} = -\frac{79}{232} \qquad (y_1 - \widehat{y_1}) = -1 + \frac{79}{232} = -\frac{153}{232}$$

$$\widehat{y_2} = -\frac{169}{1392} * 1 + \frac{2017}{1392} * 1 - \frac{79}{232} = \frac{229}{232} \qquad (y_2 - \widehat{y_2}) = 1 - \frac{229}{232} = \frac{3}{232}$$

$$\widehat{y_3} = -\frac{169}{1392} * 1 + \frac{2017}{1392} * 1 - \frac{79}{232} = \frac{229}{232} \qquad (y_3 - \widehat{y_3}) = 2 - \frac{229}{232} = \frac{235}{232}$$

$$\widehat{y_4} = -\frac{169}{1392} * 16 + \frac{2017}{1392} * 4 - \frac{79}{232} = \frac{815}{232} \qquad (y_4 - \widehat{y_4}) = 3 - \frac{815}{232} = -\frac{119}{232}$$

$$\widehat{y_5} = -\frac{169}{1392} * 49 + \frac{2017}{1392} * 7 - \frac{79}{232} = \frac{447}{116} \qquad (y_5 - \widehat{y_5}) = 4 - \frac{447}{116} = \frac{34}{232}$$

Среднее арифметическое будет такое же, следовательно, можно не вычислять. Значит, соберем все вместе:

$$\sum_{i=1}^{n}\left(y_i - P(x_i)\right)^2 = \frac{93960}{53824} \qquad \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{370}{25}$$

$$R^2 = 1 - \frac{\frac{93960}{53824}}{\frac{370}{25}} = 1 - \frac{2349000}{19914880} = 1 - 0{,}11795 = 0{,}88205$$

$$MSE = \frac{\frac{93960}{53824}}{5} = \frac{93960}{269120} = 0{,}34914 \qquad MAE = \frac{\frac{544}{232}}{5} = \frac{276}{415} = 0{,}46897$$

Сравним результаты, записав их в таблицу 1.2.

Таблица 1.2 – Сравнение метрик целевых функций

Метрика / Целевая функция	R^2	MSE	MAE
$y = kx + b$	0,78183	0,64578	0,66506
$y = ax^2 + bx + c$	0,88205	0,34914	0,46897

Исходя из полученных метрик можно сделать вывод о том, что для данных из таблицы 1.1 подходит МНК с целевой функций (1.5) лучше, чем с (1.1).

Заключение

Метод наименьших квадратов лежит в основе задач регрессии. Он отлично показывает оценку и точность предсказанных данных по отношению к исходным данным. Но также перед тем, как предсказывать данные, необходимо всегда проанализировать исходные данные, т.к. могут быть аномалии, данные выглядят нелинейно. Тогда, чтобы процесс минимизации был более эффективен, необходимо использовать модификации метода или же задавать целевую функцию в иной форме, а может и все вместе нужно использовать.

Литература

1. ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ В СТАТИСТИЧЕСКИХ ЗАДАЧАХ: [Электронный ресурс] // Белорусский государственный университет информатики и радиоэлектроники, 2023. URL: https://libeldoc.bsuir.by/bitstream/123456789/51960/1/Legeida_Osobennosti.pdf. (Дата обращения: 01.12.2025)

2. Least Absolute Shrinkage and Selection Operator [Электронный ресурс] // ScienceDirect, 2015. URL: https://www.sciencedirect.com/topics/engineering/least-absolute-shrinkage-and-selection-operator. (Дата обращения: 01.12.2025)

3. Multifactorial evolutionary algorithm enhanced by symmetry transformation and ridge regression Operator [Электронный ресурс] // ScienceDirect, 2025. URL: https://www.sciencedirect.com/science/article/abs/pii/S095741742503859X. (Дата обращения: 01.12.2025)

4. Python [Электронный ресурс] // Python, 1991. URL: https://www.python.org/. (Дата обращения: 01.12.2025)

5. Sklearn [Электронный ресурс] // Sklearn, 2007. URL: https://scikit-learn.org/stable/. (Дата обращения: 01.12.2025)

6. Sympy [Электронный ресурс] // Sympy, 2007. URL: https://www.sympy.org/en/index.html. (Дата обращения: 01.12.2025)

7. Matplotlib [Электронный ресурс] // Matplotlib, 2003. URL: https://matplotlib.org/. (Дата обращения: 01.12.2025)

8. Pandas [Электронный ресурс] // Pandas, 2008. URL: https://pandas.pydata.org/. (Дата обращения: 01.12.2025)

9. Numpy [Электронный ресурс] // Numpy, 2006. URL: https://numpy.org/. (Дата обращения: 01.12.2025)

10. Описание модификаций МНК [Электронный ресурс] // GitHub, 2025. URL: https://github.com/egorbeckish. (Дата обращения: 01.12.2025)