УДК 519.237.5

Л.А. Рябишина, кандидат технических наук, доцент

Д.А. Пак, магистр

2 курс, Уфимский государственный нефтяной технический университет, Республика Башкортостан, г. Уфа

МОДИФИКАЦИЯ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ ДЛЯ РЕГРЕССИОННОГО МОДЕЛИРОВАНИЯ В УСЛОВИЯХ «ПЛОХОЙ» СТАТИСТИКИ

Аннотация: в статье рассматривается задача повышения устойчивости регрессионных оценок в условиях малой выборки, выбросов, гетероскедастичности и неполной априорной информации о процессе. Показано, что классический метод наименьших квадратов при «плохой» статистике может давать коэффициенты, формально приемлемые по ошибке аппроксимации, но неудобные для инженерной интерпретации. Предложен модифицированный критерий оценивания, сочетающий квадратичную невязку с мягкой когнитивной коррекцией коэффициентов относительно базового вектора, заданного по физическим и технологическим соображениям. Приведены алгоритм применения метода, схема программной реализации в Python и результаты модельного эксперимента.

Abstract: the paper studies how to improve the stability of regression estimates under small samples, outliers, heteroskedastic errors, and incomplete prior information about the process. It is shown that the classical least squares method may provide coefficients with acceptable approximation error but weak engineering interpretability. A modified estimation criterion is proposed that combines the residual sum of squares with a soft cognitive correction of coefficients toward a baseline vector justified by physical and technological considerations. The paper presents the application algorithm, a Python-based implementation scheme, and the results of a model experiment.

Ключевые слова: метод наименьших квадратов, регрессионный анализ, плохая статистика, когнитивная коррекция, регуляризация, выбросы, гетероскедастичность.

Метод наименьших квадратов (МНК) остаётся одной из основных процедур оценивания параметров регрессионных моделей благодаря аналитической простоте, вычислительной эффективности и ясной интерпретации результатов. При выполнении классических предпосылок линейной модели он обеспечивает хорошие статистические свойства и потому широко применяется в эконометрике, технической диагностике и задачах идентификации объектов управления [1; 5].

Однако реальные инженерные данные редко удовлетворяют идеализированным требованиям. В архивах технологических измерений часто присутствуют малые выборки, пропуски, выбросы, различная точность датчиков и дрейф параметров режима. В таких условиях итоговая регрессия может сохранять приемлемые значения MSE или R^2, но давать коэффициенты, плохо согласованные с физикой процесса. Для систем управления это критично, поскольку модель используется не только как средство аппроксимации, но и как содержательная основа принятия решений.

Цель статьи состоит в построении компактной методики модификации МНК для условий «плохой» статистики. Для этого решаются три задачи: уточняются ограничения классического МНК, сравниваются наиболее распространённые способы его модификации и формулируется критерий оценивания с когнитивной коррекцией коэффициентов, допускающий включение экспертно-эвристической информации о процессе.

1. Ограничения классического МНК при работе с «плохой» статистикой

В линейной постановке зависимость записывается как $y = Xa + u$, где y - вектор наблюдений, X - матрица факторов, a - вектор неизвестных коэффициентов, u - ошибка. Обычный МНК минимизирует функционал $S(a) = \|y - Xa\|^2$, а при

невырожденной матрице X^TX решение имеет вид $\hat{a} = (X^TX)^{-1}X^Ty$. Теоретически этот результат корректен при отсутствии систематических нарушений структуры ошибки [1].

На практике первый источник проблем связан с ограниченным объёмом и нерепрезентативностью данных. Если выборка короткая, влияние каждой точки резко возрастает, а единичные аномальные наблюдения способны существенно изменить знак или величину коэффициента. Квадратичный характер критерия приводит к тому, что большие остатки получают непропорционально сильный вес и фактически «тянут» модель к выбросам [3; 8].

Второй блок ограничений связан с гетероскедастичностью и автокорреляцией ошибок. Для технологических переменных дисперсия измерения обычно зависит от режима работы оборудования, а соседние наблюдения во временном ряду редко оказываются независимыми. В результате оценки коэффициентов становятся менее эффективными, а доверительные характеристики — менее надёжными [4; 6; 7].

Наконец, даже при приемлемой ошибке аппроксимации коэффициенты могут быть неудобны для инженерной интерпретации. При мультиколлинеарности и шуме модель нередко формирует параметры, противоречащие физическому смыслу процесса: ожидаемо положительный коэффициент оказывается отрицательным, а свободный член выходит за допустимый диапазон. Следовательно, для задач управления важна не только статистическая точность, но и физическая правдоподобность оценок.

2. Сравнение распространённых модификаций МНК

Наиболее естественной реакцией на неодинаковую точность наблюдений является взвешенный МНК. Он минимизирует сумму квадратов отклонений с весами, обратными дисперсии ошибок, и потому позволяет ослабить влияние менее надёжных измерений. Достоинство метода состоит в хорошем соответствии

метрологической логике, однако его эффективность сильно зависит от корректности выбора весов [6].

Обобщённый МНК учитывает не только гетероскедастичность, но и коррелированность ошибок, переходя к критерию $S\_GLS(a) = (y - Xa)^T\Omega^{-1}(y - Xa)$. Если ковариационная матрица Ω известна или может быть удовлетворительно оценена, метод обеспечивает более эффективные оценки. Недостаток состоит в том, что для плохо изученного технологического процесса сама структура Ω часто задаётся приблизительно [4; 5].

При мультиколлинеарности и неустойчивости широко используются методы регуляризации. Ridge-регрессия сжимает коэффициенты к нулю и улучшает обусловленность матрицы нормальных уравнений, Lasso сочетает оценивание с отбором признаков, Elastic Net объединяет обе идеи [2; 9; 10]. Робастные процедуры, напротив, уменьшают влияние выбросов за счёт изменения функции потерь [8]. Все эти методы полезны, но в явном виде не используют экспертное знание о желаемом знаке, диапазоне или номинальном уровне коэффициента.

Таблица 1. Сравнительная характеристика методов оценивания параметров регрессии

Метод	**Преимущества**	**Ограничения**	**Типовая область применения**
Обычный МНК	Простая формула, высокая вычислительная скорость, наглядная интерпретация.	Чувствителен к выбросам, мультиколлинеарности, гетероскедастичности и малой выборке.	Первичная оценка зависимости при близости данных к классическим предпосылкам.
Взвешенный МНК	Учитывает различную точность наблюдений, уменьшает влияние шумных измерений.	Требует знания или оценки весов; чувствителен к ошибкам задания дисперсий.	Данные с неодинаковой метрологической надёжностью.

Метод	Преимущества	Ограничения	Типовая область применения
ОМНК / FGLS	Корректно учитывает ковариационную структуру ошибки, повышает эффективность оценок.	Необходима правдоподобная модель матрицы Ω; возможна нестабильность на малых выборках.	Ряды с автокорреляцией и гетероскедастичностью.
Робастные методы	Ослабляют влияние выбросов и тяжёлых хвостов распределения.	Нуждаются в выборе функции потерь и порогов; возможна потеря полезных редких режимов.	Массивы с отдельными аномальными наблюдениями.
Когнитивно скорректированный МНК	Использует экспертную информацию о коэффициентах, стабилизирует решение и сохраняет физическую интерпретируемость.	Требует обоснованного выбора базовых коэффициентов и матрицы штрафов.	Инженерные модели при «плохой» статистике и наличии содержательных ограничений.

Сравнение, приведённое в табл. 1, показывает, что предлагаемая далее модификация не подменяет существующие подходы, а дополняет их. Её назначение состоит в том, чтобы совместить статистическую устойчивость решения с использованием внешнего знания о процессе — там, где одной только выборки недостаточно для формирования содержательно правдоподобной модели.

3. Модифицированный критерий с когнитивной коррекцией коэффициентов

В инженерных приложениях априорная информация о коэффициентах обычно имеет не вероятностный, а экспертно-эвристический характер. Технолог или исследователь может не знать точное распределение параметра, но способен указать его ожидаемый знак, порядок величины или номинальное значение. Такое знание удобно включать в процедуру оценивания через мягкое предпочтение, а не через жёсткое ограничение.

Предлагается минимизировать функционал $J(a) = ||y - Xa||^2 + (a - a\_b)^T\Lambda(a - a\_b)$, где a_b - базовый вектор коэффициентов, сформированный на основе физики процесса, а Λ - диагональная матрица штрафов. Второе слагаемое интерпретируется как цена ухода от технологически обоснованных параметров. Если исследователь уверен в конкретном коэффициенте, соответствующий элемент λ_j выбирается больше; при слабом знании λ_j может быть малым или равным нулю.

Из условия минимума следует система $(X^TX + \Lambda)\hat{a}\_c = X^Ty + \Lambda a\_b$, откуда получаем $\hat{a}\_c = (X^TX + \Lambda)^{-1}(X^Ty + \Lambda a\_b)$. Формально эта оценка близка к гребневой регрессии, но принципиально отличается смыслом: коэффициенты сжимаются не к нулю, а к вектору, имеющему предметную интерпретацию. Поэтому когнитивная коррекция может рассматриваться как содержательно ориентированная форма регуляризации.

Практический эффект такого критерия двоякий. С одной стороны, диагональная матрица Λ повышает численную устойчивость задачи и уменьшает влияние случайных возмущений в выборке. С другой стороны, итоговые коэффициенты оказываются более согласованными с физической логикой объекта. При большой и качественной выборке вклад штрафного члена естественным образом уменьшается, тогда как при «плохой» статистике именно он удерживает решение от ухода в неинтерпретируемую область.

4. Практическая реализация метода

Применение метода целесообразно организовать как последовательность из нескольких этапов. Сначала выполняется предварительный анализ данных: проверка диапазонов, поиск пропусков, визуальная оценка выбросов и коллинеарности факторов. Затем строится базовая модель обычного МНК, по которой рассчитываются остатки и определяется характер возможных статистических искажений.

Следующий этап - формирование базового вектора a_b. Он может задаваться по упрощённой физической модели, по режимным расчётам, по предыдущим периодам эксплуатации объекта или по экспертной оценке специалистов. Важно, чтобы эти значения имели внешнее по отношению к текущей шумной выборке обоснование.

После этого выбирается матрица штрафов Λ. На практике наиболее удобна диагональная структура, поскольку степень уверенности в разных коэффициентах обычно неодинакова. Для ответственных факторов штраф может быть выше, для свободного члена и второстепенных параметров — ниже. Настройка возможна по перекрёстной проверке, анализу чувствительности или экспертной интервальной оценке.

Для прикладной проверки метода был реализован программный модуль на Python с использованием библиотек NumPy, Pandas, Matplotlib и tkinter. Программа позволяет загружать выборку из Excel, рассчитывать обычную и модифицированную регрессию, отображать линии на одном графике, вычислять MSE и R^2, а также проводить предварительную фильтрацию выбросов по расстоянию от линии регрессии. Такой формат делает метод пригодным не только для теоретического анализа, но и для повседневной инженерной работы с архивными данными.

5. Модельный эксперимент и обсуждение результатов

Для иллюстрации поведения метода рассмотрим модельный пример, имитирующий технологические измерения. Истинная зависимость задаётся выражением $y = 2{,}0 + 0{,}8x$. В выборку включаются наблюдения с умеренной гетероскедастичностью и двумя аномальными точками. Такая схема позволяет одновременно проследить влияние выбросов, неодинаковой точности измерений и ограниченного объёма выборки.

Расчёт обычного МНК по полной выборке даёт коэффициенты $a_0 = 2{,}276$ и $a_1 = 0{,}744$ при $R^2 = 0{,}897$. После простой фильтрации выбросов по правилу 2σ коэффициенты изменяются до $a_0 = 1{,}970$ и $a_1 = 0{,}803$. Уже этот шаг показывает, что классическая оценка заметно чувствительна к отдельным аномалиям.

Далее к очищенной выборке применяется когнитивная коррекция с базовым вектором a_b = (2,0; 0,8) и матрицей штрафов $\Lambda = \mathrm{diag}(1; 12)$. В результате получаем $a_0 = 1{,}975$ и $a_1 = 0{,}802$. По критериям подгонки метод практически не проигрывает классическому МНК на отфильтрованных данных, но сохраняет коэффициенты ближе к физически ожидаемым значениям.

С инженерной точки зрения это означает, что предлагаемый подход полезен прежде всего тогда, когда задача состоит не в формальном минимуме ошибки на текущей выборке, а в построении устойчивой и интерпретируемой модели. Метод особенно эффективен в сочетании с предварительной диагностикой качества данных: фильтрация выбросов, взвешивание наблюдений и когнитивная коррекция не конкурируют между собой, а образуют последовательную схему повышения надёжности регрессионного оценивания.

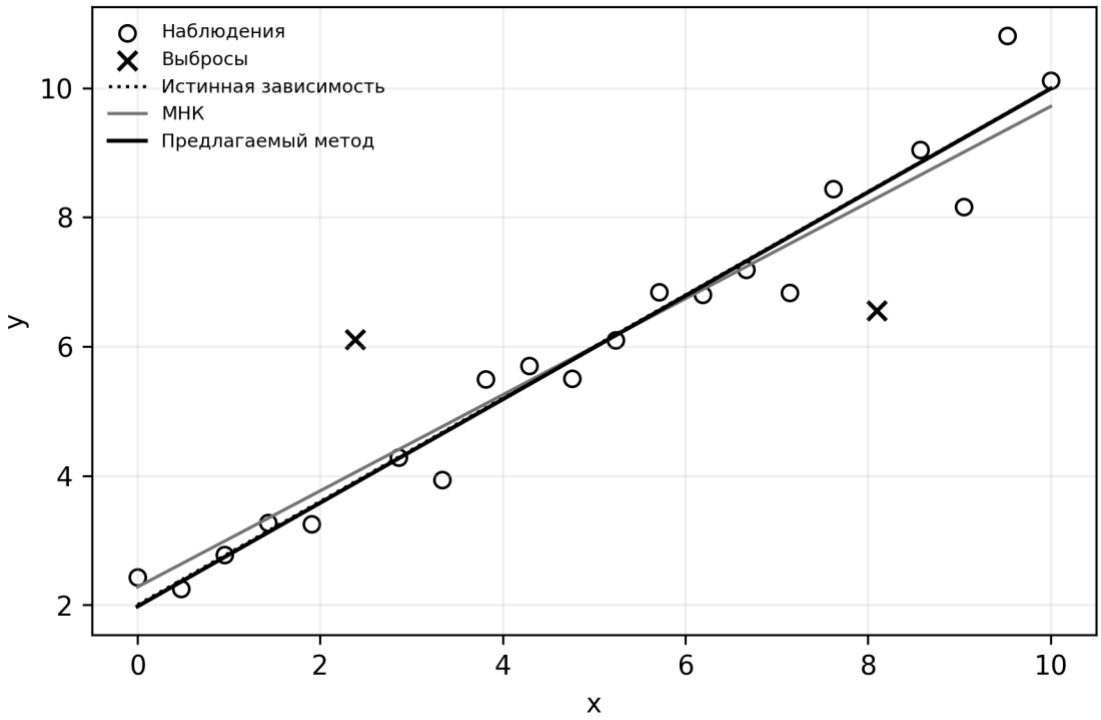

Рисунок 1 - Сопоставление обычного МНК и метода с когнитивной коррекцией на модельной выборке

Вариант расчёта	**a_0**	**a_1**	**MSE по наблюдениям**	**R^2 по наблюдениям**
Обычный МНК по полной выборке	2,276	0,744	0,579	0,897
МНК после фильтрации 2σ	1,970	0,803	0,611	0,892
Предлагаемый метод	1,975	0,802	0,610	0,892

Таблица 2 - Результаты модельного эксперимента

Как видно из табл. 2, когнитивная коррекция практически не ухудшает MSE и R^2 по сравнению с очищенной выборкой, но делает итоговые параметры более согласованными с базовой зависимостью. Именно эта комбинация устойчивости и интерпретируемости определяет прикладную ценность предложенного подхода.

Заключение

В работе показано, что классический метод наименьших квадратов в условиях «плохой» статистики сталкивается не только с ростом численной неустойчивости, но и с потерей содержательной интерпретируемости коэффициентов. Рассмотренные модификации МНК - взвешенный, обобщённый, робастный и регуляризованный варианты - снимают отдельные ограничения, однако не решают задачу прямого включения экспертно-технологической информации о параметрах модели.

Предложенный критерий с когнитивной коррекцией коэффициентов позволяет мягко смещать оценки в сторону физически обоснованного базового вектора без отказа от данных как основного источника информации. Модельный эксперимент и программная реализация показывают, что такой подход может использоваться как практический инструмент стабилизации регрессионных оценок в задачах идентификации и анализа технологических процессов при малых и неоднородных выборках.

СПИСОК ЛИТЕРАТУРЫ

1. Колмогоров А.Н. К обоснованию метода наименьших квадратов // Успехи математических наук. 1946. Т. 1. Вып. 1(11). С. 57–70. URL: https://www.mathnet.ru/rus/rm7015 (дата обращения: 23.03.2026).

2. Жданов А.И. Оптимальная регуляризация решений приближенных стохастических систем линейных алгебраических уравнений // Журнал

вычислительной математики и математической физики. 1990. Т. 30. № 10. С. 1588–1593. URL: https://www.mathnet.ru/rus/zvmmf3194 (дата обращения: 23.03.2026).
3. Шведов А.С. Простое доказательство робастности метода наименьших квадратов с урезанием для линейной регрессионной модели // Проблемы управления. 2016. № 5. С. 10–13. URL: https://www.mathnet.ru/pu987 (дата обращения: 23.03.2026).
4. Болотин Ю.В. Обобщенный метод наименьших квадратов в задаче оценивания по угловым измерениям // Автоматика и телемеханика. 1997. № 2. С. 65–74. URL: https://www.mathnet.ru/at2493 (дата обращения: 23.03.2026).
5. Aitken A.C. On least squares and linear combination of observations // Proceedings of the Royal Society of Edinburgh. 1935. Vol. 55. P. 42–48. URL: https://doi.org/10.1017/S0370164600014346.
6. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity // Econometrica. 1980. Vol. 48. No. 4. P. 817–838. URL: https://doi.org/10.2307/1912934.
7. Newey W.K., West K.D. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix // Econometrica. 1987. Vol. 55. No. 3. P. 703–708. URL: https://doi.org/10.2307/1913610.
8. Huber P.J. Robust estimation of a location parameter // The Annals of Mathematical Statistics. 1964. Vol. 35. No. 1. P. 73–101. URL: https://doi.org/10.1214/aoms/1177703732.
9. Tibshirani R. Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society. Series B. 1996. Vol. 58. No. 1. P. 267–288. URL: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
10. Zou H., Hastie T. Regularization and variable selection via the elastic net // Journal of the Royal Statistical Society. Series B. 2005. Vol. 67. No. 2. P. 301–320. URL: https://doi.org/10.1111/j.1467-9868.2005.00503.x.