

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РОБАСТНЫХ МЕТОДОВ РЕГРЕССИИ ПРИ НАЛИЧИИ ВЫБРОСОВ В ДАННЫХ

Косяков М.М., Таргонский Д.А.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Рыкова О.В. – канд. физ.-мат. наук, доцент

В работе проведен сравнительный анализ эффективности робастных методов регрессии (Huber, RANSAC, Theil-Sen) и классической линейной регрессии (OLS) при наличии выбросов. На синтетических данных с аномалиями оценены среднеквадратичная (MSE) и средняя абсолютная (MAE) ошибки. Результаты показали, что OLS значительно уступает робастным методам из-за чувствительности к выбросам. Наименьшие ошибки продемонстрировали методы Theil-Sen (MSE=0.5982) и Huber (MSE=0.5985), тогда как RANSAC оказался менее стабильным. Исследование подтверждает преимущество робастных методов для работы с зашумленными данными.

Выбросы в данных представляют собой аномальные наблюдения, которые существенно отклоняются от общей тенденции и могут исказить результаты анализа, особенно при использовании классической линейной регрессии (OLS), основанной на методе наименьших квадратов. OLS минимизирует сумму квадратов остатков, что делает его чувствительным к экстремальным значениям из-за квадратичной функции потерь. Ниже представлена формула нахождения OLS.

$$OLS = \operatorname{argmin} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1),$$

где y_i – наблюдаемое значение, $\hat{y}_i$ – предсказанное значение; n – число наблюдений.

Робастные методы регрессии, такие как регрессия Хьюбера, RANSAC и Тейл-Сена, разработаны для минимизации влияния выбросов и повышения точности моделей в условиях зашумленных данных. Целью исследования является сравнение эффективности этих методов с OLS на синтетических данных с выбросами.

Регрессия Хьюбера использует комбинированную функцию потерь, которая переходит от квадратичной к линейной при больших остатках:

$$L_{\delta}(y, \hat{y}) = \begin{cases} \frac{1}{2} (y_i - \hat{y}_i)^2, & \text{если } |y - \hat{y}| \leq \delta \\ \delta |y - \hat{y}| - \frac{1}{2} \delta^2, & \text{иначе} \end{cases} \quad (2),$$

где δ – пороговое значение; y – наблюдаемое значение; $\hat{y}$ – предсказанное значение.

RANSAC итеративно выбирает подмножества данных, строит модель и определяет inliers, повторяя процесс для выбора модели с наибольшим числом inliers. Тейл-Сен оценивает наклон регрессии через медиану наклонов всех пар точек, что обеспечивает устойчивость к выбросам до 29,3% аномалий, но имеет вычислительную сложность $O(n^2)$.

Методика исследования. Для анализа были сгенерированы синтетические данные, состоящие из 100 точек с линейной зависимостью $y = 2x + 1$, к которым добавлен нормальный шум и 10 выбросов. Данные разделены на тренировочную (80%) и тестовую (20%) выборки с использованием функции `train_test_split` (`random_state=42`). Были применены следующие модели: классическая линейная регрессия (`LinearRegression`), регрессия Хьюбера (`HuberRegressor`), RANSAC (`RANSACRegressor`) и Тейл-Сен (`TheilSenRegressor`). Оценка моделей проводилась по метрикам среднеквадратичной ошибки (MSE) и средней абсолютной ошибки (MAE) на тестовой выборке. Для реализации использовались Python и библиотеки `scikit-learn`, `matplotlib`, `numpy`. Результаты визуализированы с помощью графика, на котором отображены данные и линии регрессии для каждой модели (рисунок 1).

График, представленный на рисунке 1, показывает исходные данные и линии регрессии для каждой модели. Линейная регрессия значительно отклоняется в сторону выбросов, тогда как робастные методы лучше соответствуют основной тенденции данных.

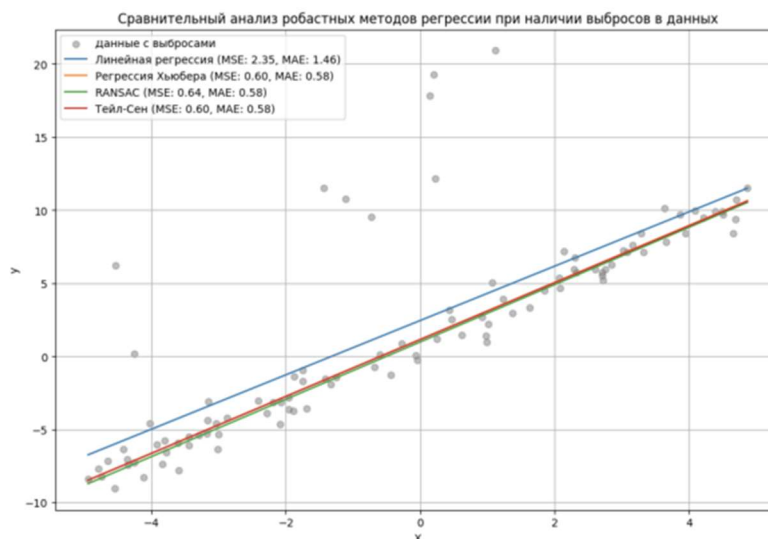


Рисунок 1 – Сравнительный анализ робастных методов регрессии при наличии выбросов в данных

Сравнительная таблица ошибок представлена в таблице 1. Линейная регрессия показала наибольшие ошибки (MSE=2.3549, MAE=1.4578), что подтверждает ее чувствительность к выбросам. Робастные методы продемонстрировали значительно меньшие ошибки: Тейл-Сен (MSE=0.5982, MAE=0.5777), Хьюбера (MSE=0.5985, MAE=0.5801) и RANSAC (MSE=0.6372, MAE=0.5833).

Таблица 1 – Сравнительная таблица ошибок моделей регрессии

Метод	MSE	MAE
Линейная регрессия	2.3549	1.4578
Регрессия Хьюбера	0.5985	0.5801
RANSAC	0.6372	0.5833
Тейл-Сен	0.5982	0.5777

Выводы. Проведенное исследование подтвердило, что классическая линейная регрессия существенно уступает робастным методам в условиях данных с выбросами. Наилучшие результаты по точности показали методы Тейл-Сена и Хьюбера, тогда как RANSAC оказался менее стабильным из-за случайного выбора подвыборок. Робастные методы регрессии предпочтительны для анализа зашумленных данных, что делает их ценным инструментом в задачах обработки данных с аномалиями.

Список использованных источников:

1. HSE ВКР: Сравнительный анализ робастных методов регрессии при наличии выбросов в данных [Электронный ресурс] : Национальный исследовательский университет «Высшая школа экономики». – Режим доступа: <https://www.hse.ru/edu/vkr/296295757>. – Дата доступа: 22.04.2025.
2. HSE Курс: Машинное обучение и анализ данных [Электронный ресурс] : Национальный исследовательский университет «Высшая школа экономики». – Режим доступа: <https://www.hse.ru/edu/courses/495283260>. – Дата доступа: 22.04.2025.
3. Robust Statistics [Электронный ресурс] : ScienceDirect. – Режим доступа: <https://www.sciencedirect.com/topics/mathematics/robust-statistic>. – Дата доступа: 22.04.2025.
4. Robust Regression [Электронный ресурс] : PennState, STAT 501: Regression Methods. – Режим доступа: <https://online.stat.psu.edu/stat501/lesson/t/t.1/t.1.1-robust-regression-methods>. – Дата доступа: 22.04.2025.
5. Robust Regression: Reduce Outlier Effects [Электронный ресурс] : Exponenta, R2021a Documentation. – Режим доступа: https://docs.exponenta.ru/R2021a_nmtnew/stats/robust-regression-reduce-outlier-effects.html. – Дата доступа: 22.04.2025.