## Медианные методы

Еще в середине прошлого века Тайл и Сен предложили метод оценки параметров, заключающийся в нахождении медианы среди параметров, полученных с помощью всевозможных минимально необходимых поднаборов исходных данных [1], [2]. Такой метод, используемый, например, для оценки параметров прямой, остается устойчивым при наличии не более чем  выбросов 29,3% выбросов. Однако при увеличении размера минимального необходимого набора данных это значение будет уменьшаться [3].

Развитием этого метода является предложенный в 1982 году метод повторной медианы [4], устойчивый к 50% выбросов. Обозначим q-ый параметр модели, восстановленный по минимально необходимому набору данных с индексами $i_1,\ i_2 \dots\ i_m$, как $\theta_q(i_1,\ i_2 \dots\ i_m)$. Возьмем медиану среди значений параметра $\theta_q$, полученного по наборам данных, где $m-1$ первых индексов фиксировано, и варьируется только последний индекс:

$$\widehat{\theta_q}(i_1,\ i_2 \dots\ i_{m-1}) = \operatorname*{med}_{j^1} \theta_q(i_1,\ i_2 \dots\ i_{m-1}, j^1)$$

Тогда каждому набору индексов $i_1,\ i_2 \dots\ i_{m-1}$ поставлено в соответствие вычисленное значение параметра $\widehat{\theta_q}$. Теперь аналогично зафиксируем $m-2$ первых индекса и получим медиану среди значений параметра $\widehat{\theta_q}$ для различных индексов $i_{m-1}$:

$$\widehat{\theta_q}(i_1,\ i_2 \dots\ i_{m-2}) = \operatorname*{med}_{j^2}\ \widehat{\theta_q}(i_1,\ i_2 \dots\ i_{m-2}, j^2)$$

На каждом шаге количество фиксируемых индексов уменьшается на один. Тогда на последнем шаге получим:

$$\widehat{\theta_q} = \operatorname*{med}_{j^m}\ \widehat{\theta_q}(j^m)$$

Объединив все операции нахождения медианы, запишем:

$$\widehat{\theta_q} = \operatorname*{med}_{j^m} \operatorname*{med}_{j^{m-1}} \dots \operatorname*{med}_{j^1} \theta_q(j^m, j^{m-1} \dots\ j^1)$$

Отметим, что описанный процесс необходимо повторить для каждого параметра модели.

Также существует метод наименьших медиан квадратов (Least Median of Squares) [5]. Для каждого минимально необходимого набора данных $s = \{i_1,\ i_2 \dots\ i_m\}$ из набора всех данных $S$, вычисляются параметры модели $\theta(s)$, а также отклонения $r(i,s)$ каждого элемента исходных данных $x_i \in S, i = 1..|S|$ от этой модели $\theta(s)$. Тогда результатом работы алгоритма будут параметры $\hat{\theta}$, полученные по формуле:

$$\hat{\theta} = \min_s \operatorname*{med}_i r(i,s)^2$$

Этот метод также устойчив при количестве выбросов не более 50% [6].

1. Theil H. A rank-invariant method of linear and polynomial regression analysis // Proceedings of the Royal Netherlands Academy of Sciences 53, 1950. P. 386–392, 521–525, 1397–1412.
2. Sen P. K., Kumar P. "Estimates of the regression coefficient based on Kendall's tau", Journal of the American Statistical Association 63, 1968. P. 1379–1389.
3. Wilcox R. Theil–Sen Estimator // Introduction to Robust Estimation and Hypothesis Testing, 2005 // P. 423–427.
4. Siegel A. F. Robust Regression Using Repeated Medians // Biometrika, 1982. Vol. 69, No. P. 242-244.
5. Rousseeuw P. J. Least median of squares regression // Journal of the American Statistical Association, 1984. Vol. 79, No. 388. P. 871–880.
6. Rosin P. L. “Further Five Point Fit Ellipse Fitting” // Graphical Models and Image Processing 09, 1999. Vol. 61, No 5. P. 245-259.