УДК 519.233

ПРОСТОЕ ДОКАЗАТЕЛЬСТВО РОБАСТНОСТИ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ С УРЕЗАНИЕМ ДЛЯ ЛИНЕЙНОЙ РЕГРЕССИОННОЙ МОДЕЛИ

А.С. Шведов

В классической линейной регрессионной модели остатки предполагаются распределенными нормально. Но реальные данные редко в точности соответствуют предположениям классической модели. При этом даже единственное резко отличающееся наблюдение может очень сильно повлиять на оценку параметров регрессии. Одним из методов робастной регрессии с высокой пороговой точкой является метод наименьших квадратов с урезанием. Дано новое доказательство теоремы о величине пороговой точки для этого метода, значительно более простое, чем оригинальное доказательство.

Ключевые слова: робастная регрессия, метод наименьших квадратов с урезанием, пороговая точка.

ВВЕДЕНИЕ

Робастные статистические методы в значительной степени созданы авторами книг [1, 2] и играют важную роль в приложениях. Современные методы робастной регрессии представлены, например, в книгах [3—5].

К работам по робастным методам примыкают работы по методам выявления резко выделяющихся наблюдений. Робастность можно понимать как уменьшение (или снятие) влияния резко выделяющихся наблюдений. При этом природа резко выделяющихся наблюдений может быть любой, от ошибок по небрежности до представления каких-то очень важных эффектов. В дальнейшем, для определенности выражений, под резко выделяющимися наблюдениями будем понимать ошибочные данные. Хотя для существа математических результатов это замечание значения не имеет.

Одним из основных показателей робастности статистического метода является пороговая точка, которая показывает, какую часть статистических данных можно испортить сколь угодно сильно, и при этом получающийся ответ будет все равно «иметь отношение к делу».

Например, чтобы одним показателем представить положение набора точек на прямой, можно воспользоваться средним арифметическим. Но тогда, испортив всего одну точку из этого набора, можно добиться того, чтобы ответ изменился сколь угодно сильно. Это пример статистического метода с низкой пороговой точкой. Но если в качестве показателя взять выборочную медиану, можно испортить до половины точек из набора, и ответ будет все равно не очень сильно отличаться от правильного (остается вопрос, что называть правильным ответом). Это пример робастного статистического метода, статистического метода с высокой пороговой точкой.

Метод наименьших квадратов (МНК), наиболее распространенный метод построения регрессионных зависимостей, имеет низкую пороговую точку. Параметр регрессии можно изменить сколь угодно сильно, изменив всего одну точку в наборе наблюдений.

Отметим, что с проблемой робастности тесно связана проблема наличия тяжелых хвостов у распределения вероятностей для ошибок регрессии. Те большие выбросы, которые маловероятны при нормальном распределении ошибок, становятся вероятными при распределениях с тяжелыми хвостами. По этим вопросам см., например, статью [6].

Одним из первых робастных методов построения регрессионных зависимостей является МНК с урезанием, подробно представленный в книге [3]. Состоит этот метод в следующем.

Пусть y_1, ..., y_n — объясняемые переменные, действительные числа; x_1, ..., x_n — регрессоры, p-мерные вектора. При каждом $\theta \in R^p$ рассматриваются ошибки

$$r_i(\theta) = y_i - \langle \theta, x_i \rangle, \quad i = 1, ..., n,$$

где через $\langle \cdot, \cdot \rangle$ обозначено скалярное произведение. Пусть эти ошибки упорядочены по возрастанию абсолютной величины

$$|r_{(1)}(\theta)| \le |r_{(2)}(\theta)| \le ... \le |r_{(n)}(\theta)|.$$

(Порядок, разумеется, разный при разных θ.) Выбирается натуральное число h, удовлетворяющее условию

$$\left[\frac{n}{2}\right] < h < n$$

(через $[\cdot]$ обозначена целая часть действительного числа), и строится функция

$$L(\theta) = \sum_{i=1}^{h} r^2_{(i)}(\theta).$$

Оценка параметров регрессии, построенная методом наименьших квадратов с урезанием, имеет вид

$$\hat{\theta}_{LTS} = \arg\min_{\theta \in R^p} L(\theta).$$

Через Z обозначим набор точек (y_1, x_1), ..., (y_n, x_n). Пусть $m < n$. Через $\Pi_m(Z)$ обозначим множество таких наборов Z', состоящих из точек (y'_1, x'_1), ..., (y'_n, x'_n), для которых, по крайней мере, $n - m$ точек содержатся в наборе Z.

Пусть функция T каждому набору точек Z' ставит в соответствие параметр регрессии θ. Пороговой точкой для функции T при заданном наборе Z называется величина $\varepsilon_n(T, Z) = m^*/n$, где m^* — минимальное из чисел m, обладающих тем свойством, что

$$\sup_{Z' \in \Pi_m(Z)} \|T(Z')\| = \infty.$$

Здесь и далее $\|\theta\| = \langle \theta, \theta \rangle^{1/2}$.

В частности, для обычного МНК $\varepsilon_n(T, Z) = 1/n$.

В книге [3, *с. 112—134*] при условии, что точки x_1, ..., x_n находятся в некотором «общем положении», доказывается

Теорема. *При*

$$h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right]$$

пороговая точка для МНК с урезанием

$$\varepsilon_n(T, Z) = \frac{1}{n}\left(\left[\frac{n-p}{2}\right] + 1\right). \ \blacklozenge$$

Это означает, что пороговая точка асимптотически равна 0,5 при $n \to \infty$ (т. е. является максимально высокой, можно испортить до половины всех точек из набора наблюдений).

Доказательство этой теоремы распадается на две части [3]. Отдельно доказывается, что $\frac{1}{n}\left(\left[\frac{n-p}{2}\right] + 1\right)$ является оценкой снизу для пороговой точки, и что эта же величина является оценкой сверху для пороговой точки. Утверждение, что данная величина представляет собой оценку сверху для пороговой точки, доказывается достаточно коротко [3, *с. 125*] и, скорее всего, имеет значение лишь как красивый математический результат. Значительно более трудным в работе [3] является доказательство утверждения, что данная величина представляет собой оценку снизу для пороговой точки (именно это утверждение и означает робастность МНК с урезанием).

Отметим, что МНК с урезанием может быть применен для практического построения регрессионных зависимостей лишь в сочетании с другими алгоритмами. (Либо в прикидочных расчетах, когда нет высоких требований к точности.) Это связано с тем, что если набор наблюдений (y_1, x_1), ..., (y_n, x_n) соответствует «чистой» нормальной случайной выборке, то точность МНК с урезанием существенно ниже, чем точность обычного МНК.

Одним из распространенных методов построения робастных оценок для параметров регрессии является метод *MM*-оценивания, предложенный в работе [7] (см. также книгу [4]). Построение оценки этим методом состоит из трех этапов. На первом этапе строится оценка параметров регрессии некоторым методом с высокой пороговой точкой. На втором этапе с использованием этих параметров строится *M*-оценка для масштаба. На третьем этапе с использованием найденного значения масштаба строится *M*-оценка для параметров регрессии (отсюда использование двух букв «*M*» в названии метода). Доказывается, что пороговая точка при *MM*-оценивании параметров регрессии асимптотически равна 0,5.

В настоящей работе дается значительно более простое доказательство сформулированного результата о пороговой точке для МНК с урезанием для параметров регрессии, чем доказательство из [3]; речь идет об оценке снизу для $\varepsilon_n(T, Z)$, т. е. о доказательстве робастности метода наименьших квадратов с урезанием.

1. ВСПОМОГАТЕЛЬНЫЙ РЕЗУЛЬТАТ

Во всей работе будем считать, что $n \ge p$ и что выполняется условие: любые p из векторов $x_1, ..., x_n$ образуют базис пространства R^p.

Из этого условия следует, что если для p-мерного вектора θ, $\|\theta\| = 1$, и для некоторого набора векторов $x_{i_1}, ..., x_{i_{p-1}}$ выполняется

$$\langle \theta, x_{i_1} \rangle = 0, ..., \langle \theta, x_{i_{p-1}} \rangle = 0,$$

то для любого вектора x_j, не входящего в этот набор, $\langle \theta, x_j \rangle \ne 0$.

Некоторым обобщением этого утверждения является

Лемма. *Существует число* $\delta > 0$ *такое, что если для p-мерного вектора* θ, $\|\theta\| = 1$, *и для некоторого набора векторов* $x_{i_1}, ..., x_{i_{p-1}}$ *выполняется*

$$|\langle \theta, x_{i_1} \rangle| < \delta, ..., |\langle \theta, x_{i_{p-1}} \rangle| < \delta,$$

то для любого вектора x_j, *не входящего в этот набор,* $|\langle \theta, x_j \rangle| \geqslant \delta$.

Доказательство. Рассмотрим в качестве набора векторов $x_{i_1}, ..., x_{i_{p-1}}$ набор $x_1, ..., x_{p-1}$. Проведя ортогонализацию Грама — Шмидта, построим ортонормированную систему $e_1, ..., e_{p-1}$,

$$e_i = \sum_{k=1}^{i} \gamma_{ik} x_k, \quad i = 1, ..., p-1.$$

Проекция вектора θ на гиперплоскость, порожденную векторами $e_1, ..., e_{p-1}$, имеет вид

$$\langle \theta, e_1 \rangle e_1 + ... + \langle \theta, e_{p-1} \rangle e_{p-1} = \sum_{i=1}^{p-1} \left(\sum_{k=1}^{i} \gamma_{ik} \langle \theta, x_k \rangle \right) e_i.$$

Будем считать, что все скалярные произведения $\langle \theta, x_k \rangle$ настолько малы по абсолютной величине, что длина этой проекции не превосходит $1/\sqrt{2}$.

Выберем вектор x_j, $j > p-1$. Поскольку этот вектор не является линейной комбинацией векторов $x_1, ..., x_{p-1}$, можно записать

$$x_j = \sum_{i=1}^{p-1} \alpha_i x_i + \beta t,$$

где t — единичный вектор ортогональный гиперплоскости, порожденной векторами $e_1, ..., e_{p-1}$; $\beta \ne 0$.

Длина проекции вектора θ на гиперплоскость, порожденную векторами $e_1, ..., e_{p-1}$, не превосходит $1/\sqrt{2}$. Поэтому длина проекции вектора θ на вектор t должна быть не меньше, чем $1/\sqrt{2}$. Имеем

$$\langle \theta, t \rangle = \frac{1}{\beta} \langle \theta, x_j \rangle - \frac{1}{\beta} \sum_{i=1}^{p-1} \alpha_i \langle \theta, x_i \rangle.$$

Будем считать, что все скалярные произведения $\langle \theta, x_i \rangle$ настолько малы по абсолютной величине, что

$$\left| \frac{1}{\beta} \sum_{i=1}^{p-1} \alpha_i \langle \theta, x_i \rangle \right| \le \frac{1}{\sqrt{2}} - \frac{1}{2}.$$

Тогда

$$|\langle \theta, x_j \rangle| \ge |\beta|/2.$$

Дополнительно наложим условие $\delta < |\beta|/2$.

Пока утверждение леммы доказано для одного j. Такое же рассмотрение можно провести для всех $j = p, ..., n$ и выбрать минимальное требуемое δ. Затем нужно рассмотреть все существующие наборы $x_{i_1}, ..., x_{i_{p-1}}$ (всего таких наборов C_n^{p-1}) и вновь выбрать минимальное требуемое δ. Лемма доказана.

2. РЕЗУЛЬТАТ О ПОРОГОВОЙ ТОЧКЕ

Выберем числа m и h так, что $m + p \le h \le n - m$, и покажем, что тогда для МНК с урезанием

$$\sup_{Z' \in \Pi_m(Z)} \|T(Z')\| < \infty.$$

Для набора $Z' \in \Pi_m(Z)$ остатки обозначим

$$\rho_i(\theta) = y'_i - \langle \theta, x'_i \rangle, \quad i = 1, ..., n.$$

Положим $\theta = 0$, и пусть наименьшими по абсолютной величине h остатками являются

$$\rho_{k_1}(0), ..., \rho_{k_h}(0),$$

где $k_1, ..., k_h$ — различные числа из набора $1, ..., n$. В силу условия $m \le n - h$ в наборе Z' содержатся точки $(y_{j_1}, x_{j_1}), ..., (y_{j_h}, x_{j_h})$. Тогда

$$L(0) = \sum_{i=1}^{h} \rho_{k_i}^2(0) \le \sum_{i=1}^{h} r_{j_i}^2(0) \le \sum_{i=1}^{n} r_i^2(0) = \sum_{i=1}^{n} y_i^2.$$

Положим

$$C_0 = \max_{1 \le j \le n} |y_j|$$

и выберем число C_1 такое, что

$$C_1 > \frac{1}{\delta} \left(\left(\sum_{i=1}^{n} y_i^2 \right)^{1/2} + C_0 \right),$$

где $\delta > 0$ определяется леммой.

Возьмем $\theta \in R^p$ такое, что $\|\theta\| \ge C_1$. Пусть наименьшими по абсолютной величине h остатками являются $\rho_{k_1}(\theta), ..., \rho_{k_h}(\theta)$. В силу условия

$m + p \le h$ среди точек (y'_{k_1}, x'_{k_1}), ..., (y'_{k_h}, x'_{k_h}) присутствуют точки (y_{i_1}, x_{i_1}), ..., (y_{i_p}, x_{i_p}).

Воспользовавшись тем, что для любых действительных чисел α и β выполняется неравенство

$$|\alpha - \beta| \ge \|\alpha| - \beta\|,$$

имеем

$$L(\theta) = \sum_{i=1}^{h} \rho^2_{k_i}(\theta) \ge \sum_{l=1}^{p} r^2_{i_l}(\theta) \ge \sum_{l=1}^{p} (|\langle \theta, x_{i_l}\rangle| - |y_{i_l}|)^2.$$

Пусть $\theta_0 = \theta/|\theta|$. На основании леммы хотя бы для одного из индексов, например, для i_p выполняется $|\langle \theta_0, x_{i_p}\rangle| \ge \delta$. Тогда

$$L(\theta) \ge (\|\theta\| \cdot |\langle \theta_0, x_{i_p}\rangle| - |y_{i_p}|)^2 \ge (C_1\delta - C_0)^2 > \sum_{i=1}^{n} y_i^2.$$

Тем самым установлено, что для любого набора $Z' \in \Pi_m(Z)$ выполняется $\|T(Z')\| \leqslant C_1$.

Нетрудно проверить, что при

$$m = \left[\frac{n-p}{2}\right], \quad h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right]$$

неравенства $m + p \le h \le n - m$ имеют место. Следовательно,

$$m^* \ge \left[\frac{n-p}{2}\right] + 1.$$

ЗАКЛЮЧЕНИЕ

Результат работы заключается в новом доказательстве известной теоремы о робастности метода наименьших квадратов с урезанием для линейной регрессионной модели; под робастностью понимается высокая пороговая точка, асимптотически равная 0,5 при больших размерах выборки. Это доказательство существенно проще, чем оригинальное доказательство, приводимое в книге [3].

ЛИТЕРАТУРА

1. *Хьюбер П.Дж.* Робастность в статистике. — М.: Мир, 1984. — 303 с.
2. *Робастность* в статистике. Подход на основе функций влияния / Ф. Хампель и др. — М.: Мир, 1989. — 512 с.
3. *Rousseeuw P.J., Leroy A.M.* Robust regression and outliers detection. — N.-Y.: Wiley, 1987. — 329 p.
4. *Maronna R., Martin D., Yohai V.* Robust statistics: Theory and methods. — Chichester: Wiley, 2006. — 403 p.
5. *Andersen R.* Modern methods for robust regression. — Los Angeles: Sage Publications, 2008. — 107 p.
6. *Шведов А.С.* Робастная регрессия с применением *t*-распределения и *EM*-алгоритма // Экономический журнал ВШЭ. — 2011. — Т. 15. — С. 68—87.
7. *Yohai V.J.* High breakdown-point and high efficiency robust estimates for regression // Annals of Statistics. — 1987. — Vol. 15. — P. 642—656.

Статья представлена к публикации членом редколлегии А.Г. Кушнером.

Шведов Алексей Сергеевич — д-р физ.-мат. наук, профессор, Национальный исследовательский университет «Высшая школа экономики», г. Москва,
✉ ashvedov@hse.