УДК 519.6

В.И. ГРИЦЮК, канд. техн. наук, доц., ХНУРЭ, Харьков

МОДИФИЦИРОВАННЫЕ РОБАСТНЫЕ ГРЕБНЕВЫЕ ОЦЕНКИ ДЛЯ ОПРЕДЕЛЕНИЯ ВЫБРОСОВ

Оценки наименьших квадратов (LSE), в множественной линейной регрессии, когда предсказатели сильно коррелированы, дают низкую точность предсказания. Гребневая регрессия, являясь регуляризованной версией регрессии на основе метода наименьших квадратов, основываясь на минимизации квадратичной функции потерь, чувствительна к выбросам. Рассмотрены сглажено сниженные ψ-функции, которые приводят к асимптотически эффективным оценкам. Для получения результирующих робастных гребневых оценок для выявления выбросов используется метод итеративно ревзвешенных наименьших квадратов (IRLS) на основе рассмотренной ψ-функции. Результаты моделирования подтверждают полученные теоретические выводы. Получена сходимость к итоговым оценкам коэффициентов с меньшим количеством итераций, чем без применения гребневой регрессии. Объединенные робастные и гребневые оценки позволяют получить стабильные коэффициенты и остатки, которые помогают определить истинные коэффициенты и выбросы.

Ключевые слова: М-оценки, робастные гребневые оценки.

Актуальность работы. В статье предлагается робастная и гребневая регрессия для одновременного решения проблемы мультиколлинеарности и определения выбросов в классической линейной регрессионной модели.

Когда предикторные переменные мультиколлинеарны, оценки наименьших квадратов могут быть слишком большими по абсолютной величине и дисперсии могут стать очень большими.

В множественной линейной регрессии, когда предсказатели тесно связаны, оценки наименьших квадратов (LSE) дают неточные прогнозы. Чтобы это исправить, Hoerl и Kennard предложили гребневую регрессию [1]. Они добавили штраф, который создаёт небольшое смещение для того, чтобы одновременно уменьшить оценку и уменьшить дисперсию, что приводит к повышению общей точности прогнозирования. Проблемы регрессии состоят как в мультиколлинеарности так и ненормальности в той или иной степени. Холланд изучал совокупную проблему, и предложил использовать взвешенную гребневую регрессию с робастным выбором весов. В данной работе представлен подход, основанный на сочетании математических формулировок программирования гребневой и

© В.И. Грицюк, 2015

робастной регрессии. Искомые коэффициенты регрессии могут быть легко вычислены путем итеративной ревзвешенной процедуры наименьших квадратов, примененной к расширенному набору данных. В результате робастные и гребневые оценки часто являются наилучшими оценками по сравнению с только либо робастными либо гребневыми оценками.

Целью настоящей работы является исследование и разработка объединённых методов робастного и гребневого оценивания, обладающих улучшенными свойствами сходимости и асимптотической эффективностью.

Методы робастного оценивания. Известно, что метод МНК ведет себя плохо, когда распределение ошибок не является нормальным, особенно, когда ошибки являются тяжелыми хвостами, то есть, если существуют отдаленные наблюдения. Эта чувствительность МНК к выбросам результатов приводит к очень обманчивым результатам. Чтобы справиться с этой проблемой была разработана методика робастной регрессии. Наиболее распространенным является метод робастной регрессии М-оценки, введенный Хубером [2, 3]. Наиболее часто используемыми робастными оценками являются Хьюбера М-оценки (Хампель и др., 1986), ММ-оценки (Йохай, 1987), GM-оценки, Сигеля оценки повторяющихся медиан (Rousseeuw и Leroy 1987), оценки наименьших квадратов медиан (LMS), LTS-оценки, (Rousseeuw 1984), S-оценки (Rousseeuw и Yohai 1984), MVE-оценки (Rousseeuw и Leroy 1987), и оценивание минимального определителя ковариационной матрицы (MCD) (Rousseeuw и Van Driessen 1998). В настоящем исследовании вводится новое семейство асимптотически эффективных, сглаженно сниженных М-оценок.

М-оценивание основано на идее замены квадратов остатков, используемых в оценке МНК, другой функцией остатков,

$$\min_{\hat{\theta}} \sum_{i=1}^{n} \rho(r_i), \qquad (1)$$

где ρ является симметричной функцией с минимумом в нуле, при этом, ρ-функция должна обладать следующими свойствами:

1) $\rho(0) = 0$,
2) $\rho(t) \geq 0$,
3) $\rho(t) = \rho(-t)$,
4) $\rho(t_1) \leq \rho(t_2)$ для $0 < t_1 < t_2$,
5) ρ непрерывная,

Дифференцируя уравнение (1) по отношению к коэффициентам регрессии, получаем

$$\sum_{i=1}^{n} \psi(r_i) x_{ij} = 0,\ j = 1,2,\ldots,p, \tag{2}$$

$$\sum_{i=1}^{n} \psi(r_i / \hat{\sigma}) x_{ij} = 0,\ j = 1,2,\ldots,p, \tag{3}$$

где ψ является производной от ρ, x_i является вектор-строкой объясняющих переменных *i*-го наблюдения. М-оценка получается путем решения системы 'p' нелинейных уравнений. Решение не эквивариантно относительно масштабирования. Таким образом, остатки должны быть стандартизированы с помощью некоторой оценки стандартного отклонения σ, так что, они должны быть оценены одновременно. Одна возможность состоит в использовании медианы абсолютных отклонений (MAD). Шкала оценки: $\hat{\sigma} = 1.483 \mathrm{med}_i |r_i|$. Умножение на 1,483 сделано так, что для нормально распределенных данных $\hat{\sigma}$ является оценкой стандартного отклонения. Соответствующая W-функция (весовая функция) для любого ρ затем определяется как

$$\omega(t_i) = \frac{\psi(t_i)}{t_i}, \tag{4}$$

где t_i стандартизированные остатки. Используя эти W-функции в МНК, мы получаем взвешенный метод наименьших квадратов (WLS) и полученные оценки называются взвешенными оценками (Hoaglin и др., 1983). Взвешенные оценки вычисляются путем решения уравнений, где W является диагональной квадратной матрицей, имеющей диагональные элементы в качестве весов.

$$\hat{\beta} = \left(X^T W X\right)^{-1} X^T W y. \tag{5}$$

Ψ-функция Хьюбера определяется, как

$$\psi(t) = \begin{cases} -a, & t < -a, \\ t, & -a \le t \le a, \\ a, & t > a, \end{cases} \tag{6}$$

где a – так называемая константа настройки.

Сниженные М-оценки. Сниженные М-оценки были введены Hampel, который использовал три части сниженных оценок с ρ-функциями, ограниченная ψ-функция принимает значение 0 для больших (Хампель и др., 1986) $|t|$. Состоящая из трех частей сниженная ψ-функция Хампеля определяется как

$$\psi(t)=\begin{cases}\operatorname{sgn}(t)|t|, & \text{если } 0\le|t|<a,\\ a\operatorname{sgn}(t), & \text{если } a\le|t|<b,\\ \{(c-|t|)/(c-b)\}a\operatorname{sgn}(t), & \text{если } b\le|t|<c,\\ 0, & c\le|t|,\end{cases} \tag{7}$$

(Hoaglin и др.). Возникает потребность в ψ-функции сглаженно сниженной природы. Некоторые сглаженно сниженные М-оценки были предложены разными авторами. Реальные улучшения были получены Эндрюсом (Andrews, 1974) и Тьюки (Mosteller и Tukey 1977; Hoaglin и др, 1983), которые использовали волновые оценки (также называемые синус-оценки) и бивейт оценки, соответственно. И волна Эндрюса и бивейт оценки Тьюки являются сглажено сниженными ψ-функциями. Потом Кадир (1996) предложил ψ-функцию, с весовой функцией бета-функцией с α = β. Волновая функция Эндрюса

$$\psi(t)=\begin{cases}a\sin\left(\dfrac{t}{a}\right), & |t|\le\pi a\\ 0 & |t|>\pi a.\end{cases} \tag{8}$$

Бивейт функция Тьюки

$$\psi(t)=\begin{cases}t\left[1-\left(\dfrac{t}{a}\right)^2\right]^2, & |t|\le a,\\ 0, & |t|>a.\end{cases} \tag{9}$$

Результаты моделирования по методу Андрюса, Тьюки в сравнении с методом наименьших квадратов приведены ниже. В качестве примера исследован известный набор данных, взятый из Rousseeuw и Leroy (1987). Этот пример выбран, потому что этот реальный набор данных [4-6] был рассмотрен многими статистиками, такими как Danial и Wood (1971), Andrews (1974), Andrews и Pregibon (1978), Cook (1979), Draper и Smith (1981), Dempster и Gasko-Green (1981), Atkinson (1982), Rousseeuw и Leroy (1984), Carroll и Rupert (1985), Qadir (1996) и некоторыми другими с помощью различных методов. Данные описывают работу установки для окисления аммиака в азотную кислоту и состоят из 21 четырехмерных наблюдений. Stackloss (y) должен быть объяснен скоростью работы (x_1), температурой охлаждающей воды на входе (x_2), и концентрацией кислоты (x_3).

Таблица 1 – Исходные данные, описывающие работу установки для окисления аммиака в азотную кислоту

№	Y	x_1	x_2	x_3
1	42	80	27	89
2	37	80	27	88
3	37	75	25	90
4	28	62	24	87
5	18	62	22	87
6	18	62	23	87
7	19	62	24	93
8	20	62	24	93
9	15	58	23	87
10	14	58	18	80
11	14	58	18	89
12	13	58	17	88
13	11	58	18	82
14	12	58	19	93
15	8	50	18	89
16	7	50	18	86
17	8	50	19	72
18	8	50	19	79
19	9	50	20	80
20	15	56	20	82
21	15	70	20	91

В результате моделирования были получены следующие оценки.

$$E(y) = -39.919 + 0.716\,x_1 + 1.295\,x_2 - 0.152\,x_3, \quad (10)$$

$$E(y) = -37.652 + 0.798\,x_1 + 0.577\,x_2 - 0.067\,x_3, \quad (11)$$

$$E(y) = -37.061 + 0.821\,x_1 + 0.513\,x_2 - 0.074\,x_3, \quad (12)$$

$$E(y) = -36.908 + 0.827\,x_1 + 0.495\,x_2 - 0.075\,x_3. \quad (13)$$

Уравнение (10) включает оценки коэффициентов, полученные МНК. Уравнение (11) содержит оценки коэффициентов, полученные МНК с удалёнными точками 1, 3, 4 и 21. Уравнение (12) содержит соответственно оценки коэффициентов, полученные методом Андрюса (a = 1,5), уравнение (13) содержит оценки коэффициентов, полученные с функцией бивейт Тьюки (a = 4,685).

Асимптотическая вариация и эффективность *M*-оценок. Для больших n можно выразить $\hat{\beta}$ как примерно нормально распределенное

$$D(\hat{\beta}) \approx N_p\left(\beta, \hat{\upsilon}(X^T X)^{-1}\right), \quad (14)$$

где $$\hat{\upsilon} = \hat{\sigma}^2 \frac{\text{ave}_i\left\{\psi\left(r_i / \hat{\sigma}\right)^2\right\}}{\left[\text{ave}_i\left\{\psi'\left(r_i / \hat{\sigma}\right)\right\}\right]^2} \frac{n}{n-p}, \quad (15)$$

где $\text{ave}_i(z_i)$ – среднее набора данных z.

На практике можно оценить

$$\left[E\left(\psi^2\right)\right] \quad \text{как} \quad \frac{1}{n}\sum_{i=1}^{n}\psi^2$$

$$\text{и} \quad \left[E\left(\psi'\right)\right]^2 \quad \text{как} \quad \left(\frac{1}{n}\sum_{i=1}^{n}\psi'\right)^2.$$

Новая ψ-функция. Новая ρ-функция, предложена в семействе гладко сниженных М-оценок [7]. ψ-функция, связанная с этой ρ-функцией, достигает гораздо большей линейности в центральной части прежде, чем она спадает, по сравнению с другими ψ-функциями, такими, как синус Эндрюса, бивейт Тьюки и Кадира бета-функция, в результате ее повышенной эффективности. Многократно ревзвешенный метод наименьших квадратов (IRLS) на основе предложенной ρ-функции явно обнаруживает выбросы и игнорирует выбросы, которые уточняются при последующем анализе. Метод действительно достигает целей, ради которых он построен, потому что дает улучшенные результаты во всех ситуациях и способен выдержать значительное количество выбросов. Предлагаемая ψ-функция [7] приведена ниже.

$$\psi(t) = \begin{cases} \frac{2t}{3}\left(1 - \left(\frac{t}{a}\right)^4\right)^2, & \text{если } |t| \le a, \\ 0, & \text{если } |t| > a, \end{cases} \quad (16)$$

где a – так называемая константа настройки и для i-ого наблюдения переменная t – остатки, шкалированные MAD.

ρ – функция, соответствующая ψ-функции, приведенной выше, удовлетворяет стандартным свойствам, как правило связанным с обоснованной целевой функцией.

Объединенный метод робастного и гребневого оценивания. Можно заметить, что сглаженно сниженные М-оценки ведут себя очень плохо, если ошибки действительно нормально распределены.

Из [7] видно, что асимптотическая вариация и эффективность предложенной ψ функции, намного улучшены по сравнению с другими версиями.

Более эффективным представляется объединение робастного и гребневого оценивания. В этом случае используем следующие соотношения, применяя робастную и гребневую оценку. Для набора данных регрессии (X, y) с $X \in R^{n\times p}$ и $y \in R^n$

$$\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min\left\{L(X, y, \beta : \beta_0 \in R, \beta_1 \in R^P\right\}, \tag{17}$$

$$L(X, y, \beta) = \sum_{i=1}^{n} \rho\left(\frac{r_i(\beta)}{\hat{\sigma}_{ini}}\right) + \frac{\lambda}{\hat{\sigma}_{ini}^2}\|\beta_1\|^2, \tag{18}$$

$$r = (r_1, \ldots, r_n)^T = \left(y - \hat{\beta}_0 1_n - X\hat{\beta}_1\right).$$

Как известно, классическая оценка гребневой регрессии (RR) соответствует нормальным уравнениям

$$\hat{\beta}_0 = \bar{y} - \bar{x}^T\hat{\beta}_1,$$
$$\left(X^T X + \lambda I_p\right)\hat{\beta}_1 = X^T\left(y - \hat{\beta}_0 1_n\right), \tag{19}$$

I_p – единичная матрица, $\bar{x}$ и $\bar{y}$ – средние X и y соответственно.

Система уравнений, соответствующая робастной гребневой оценке (RRR)

$$\psi(t) = \rho'(t),\ W(t) = \frac{\psi(t)}{t}. \tag{20}$$

Пусть

$$\sigma = \hat{\sigma}\left(r\left(\hat{\beta}\right)\right),\ t_i = \frac{r_i}{\sigma},\ \omega_i = \frac{\psi(t_i)}{2t_i},$$
$$w = (\omega_1, \ldots, \omega_n)^T, \tag{21}$$
$$W = \mathrm{diag}(w).$$

Приравняем производную по β в (18) нулю для RRR.

$$w^T\left(y - \hat{\beta}_0 1_n - X\hat{\beta}_1\right) = 0,$$
$$\left(X^T W X + \lambda I_p\right)\hat{\beta}_1 = X^T W\left(y - \hat{\beta}_0 1_n\right). \tag{22}$$

В результате исследований было доказано, что оценивание на основе обоих смещенных и робастных методов может быть полезной процедурой в тех случаях, когда наборы данных подвержены одновременно неортогональности и ненормальным ошибкам. Процедура оценки состоит из увеличения исходного набора данных так, что обычный метод наименьших квадратов даёт смещенную оценку данных. Затем многократно ревзвешенный метод наименьших квадратов, (IRLS) предполагающий итеративную процедуру, может быть использован для получения результирующих робастных и гребневых оценок.

Таблица 2 – Остатки, полученные по методу Андрюса, методу Тьюки, робастной гребневой регрессии с ψ-функцией в сравнении с методом МНК

№	У	Остатки МНК	Остатки МНК без выброс	Остатки Андрюс	Остатки Тьюки	Остатки RRR с ψ-функцией
1	42	3,24	6,22	6,02	6,04	5,87
2	37	–1,92	1,15	0,95	0,96	0,84
3	37	4,56	6,43	6,23	6,24	5,96
4	28	5,70	8,17	8,25	8,26	8,31
5	18	–1,71	–0,67	–0,74	–0,74	-0,88
6	18	–3,01	–1,25	–1,24	–1,24	-1,28
7	19	–2.39	–0,42	–0,30	–0,28	-0,49
8	20	–1,39	0,58	0,71	0,72	0,50
9	15	–3,14	–1,06	–0,94	–0,93	-0,89
10	14	1,27	0,36	0,04	0,02	-0,09
11	14	2,64	0,96	0,72	0,69	0,20
12	13	2,78	0,47	0.15	0,11	-0,43
13	11	–1,43	–2,51	–2,81	–2,83	-3,03
14	12	–0,05	–1,35	–1,48	–1,5	-2,08
15	8	2,36	1,34	1,33	1,33	1,00
16	7	0,91	0,14	0,10	0,09	-0,10
17	8	–1,52	–0,37	–0,45	–0,46	0,05
18	8	–0,46	0,1	0,07	0,07	0,27
19	9	–0,60	0,59	0,65	0,65	0,89
20	15	1,41	1,93	1,84	1,83	1,86
21	15	–7,24	–8,63	–9,05	–9,07	-9,74

При моделировании были получены следующие оценки коэффициентов.

$$E(y) = -39.883 + 0.849\,x_1 + 0.404\,x_2 - 0.032\,x_3. \qquad (23)$$

Уравнение (23) содержит оценки коэффициентов, полученные с применением робастных и гребневых оценок с функцией $\psi(t)$ (a = 2). Параметр λ определяется согласно методу, приведенному в [8].

Из табл. 2 видно, что робастные процедуры по методу Андрюса (a = 1,5) и методу Тьюки (a = 4,685) ведут к идентификации четырёх выбросов и дают те же оценки, что и метод наименьших квадратов, когда из данных удалены четыре выброса.

Из таблицы также видно, что предложенный объединённый метод робастного и гребневого оценивания, основанный на ψ-функции (a = 2), подтверждает факт, что наблюдения 1, 3, 4 и 21 являются выбросами, так как предложенный метод даёт высокие величины остатков для этих наблюдений.

Выводы. Применение объединённых робастных и гребневых оценок позволяет получить сходимость к итоговым оценкам коэффициентов с меньшим количеством итераций, чем без использования гребневых оценок. Использование разработанной процедуры приводит к получению устойчивых коэффициентов и остатков, которые позволяют определить истинные коэффициенты и выбросы. Оценки, полученные с ψ-функцией, автоматически находят эти коэффициенты и определяют выбросы. Главное преимущество заключается в обнаружении наблюдений для дальнейшего изучения.

Список литературы: 1. Owen A. A robust hybrid of lasso and ridge regression // Technical report, Stanford University, CA. – 2006. – P. 1-14. **2.** Alma Ö. Gürünlü. Comparison of Robust Regression Methods in Linear Regression // Int. J. Contemp. Math. Sciences. – 2011. – Vol. 6, no. 9. – P. 409-421. **3.** Qadir, M.F. Robust Method for Detection of Single and Multiple Outliers / M. F. Qadir // Scientific Khyber. –1996. – Vol. 9. – P. 135-144. **4**. Deniel C. and Wood F.S. Fitting Equations to Data, John Wiley and Sons. – New York, 1999. – 459 p. **5.** Rousseeuw P.J. and Leroy A.M. Robust Regression and Outlier Detection, John Wiley and Sons. – New York, 1987. – 334p. **6.** Rousseeuw P.J. and Hubert M. Recent Development in PROGRESS //Computational Statistics and Data Analysis. – 1996. – Vol. 21. – P. 67-85. **7.** Asad, A.A. Modified M-Estimator for the Detection of Outliers / A. Asad, M.F. Qadir // Pakistan Journal of Statistics and Operation Research. – 2005. – Vol. 1. – P. 49-64. **8.** Грицюк В.И. Модифицированный алгоритм наименьших квадратов и выбор модели. // Вестник НТУ "ХПИ". Серия Автоматика и приборостроение. – 2004. – № 17. – С. 47-50.

Bibliorafy (transliterated): 1. Owen, A. "A robust hybrid of lasso and ridge regression". *Technical report, Stanford University, CA.* (2006): 1-14. Print. **2**. Alma, Ö. Gürünlü. "Comparison of Robust Regression Methods in Linear Regression", *Int. J. Contemp. Math. Sciences*. Vol 6, No 9 (2011): 409-421. Print. **3**. Qadir, M.F. "Robust Method for Detection of Single and Multiple Outliers", *Scientific Khyber*. 9 (1996): 135-144. Print. **4**. Deniel, C. and Wood, F.S. *Fitting Equations to Data*. New York: John Wiley and Sons, (1999): 459. Print. **5**. Rousseeuw, P.J. and Leroy, A.M. *Robust Regression and Outlier Detection.* New York: John Wiley and Sons, (1987): 334. Print. **6**. Rousseeuw, P.J. and Hubert, M. "Recent Development in PROGRESS". *Computational Statistics and Data Analysis.* 21 (1996): 67-85. Print. **7.** Asad, A. and Qadir, M.F. "A Modified M-Estimator for the Detection of Outliers." *Pakistan Journal of Statistics and Operation Research*. 1 (2005): 49-64. Print. **8**. Gritsyuk, V.I. "The modified least squares algorithm and model choice". *Vestnik of National Technical University "KPI", Series Automation and priborostroenie*. 17 (2004): 47-50. Print.

Поступила (received) 28.02.2015

Грицюк Вера Ильинична, кандидат технических наук, доцент кафедры "Проектирование и эксплуатация электронных аппаратов" Харьковского национального университета радиоэлектроники. Защитила диплом инженера-электрика по специальности "Автоматика и телемеханіка" в 1971г. в Харьковском институте радиоэлектроники, диссертацию кандидата технических наук в Харьковском государственном техническом университете радиоэлектроники по специальности "Управление в технических системах" в 1995г.
Научные интересы: стохастические системы управления.