9. Bellahsene, Z. Forum: a flexible data integration system based on data semantics [Text] / Z. Bellahsene, S. Benbernou, H. Jaudoin, F. Pinet, O. Pivert, F. Toumani, S. Bernard, P. Colomb, R. Coletta, E. Coquery, F. De Marchi, F. Duchateau, M.-S. Hacid, A. HadjAli, M. Roche // SIGMOD Record. – 2010. – Vol. 39, Issue 2. – P. 11–18.
10. Roche, M. AcroDef: A quality measure for discriminating expansions of ambiguous acronyms [Text] / M. Roche, V. Prince // Modeling and Using Context. Springer-Verlag Berlin Heidelberg, 2007. – P. 411–424. doi: 10.1007/978-3-540-74255-5_31
11. Roche, M. Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes [Text]: PhD thesis / M. Roche. – Paris, 2004.
12. Smadja, F. Translating collocations for bilingual lexicons: A statistical approach [Text] / F. Smadja, K.R. McKeown, V. Hatzivassiloglou // Computational Linguistics. – 1996. – Vol. 22, Issue 1. – P. 1–38.
13. Dictionnaire de sigles et acronyms [Electronic resource] / G. Blandin. – Asankyeya, 2005. – Available at: http://www.sigles.net
14. Medline [Electronic resource] / R. Pike. – USA, 2004. – Available at: http://www.ncbi.nlm.nih.gov/PubMed

У множинній лінійній регресії, коли провісники сильно корельовані, оцінки найменших квадратів (LSE), як правило, дають неточні прогнози. Гребенева регресія, яка грунтується на мінімізації квадратичної функції втрат, чутлива до викидів. Розглянуто дві гладко знижені ψ-функції, засновані на принципі Вінзора, які призводять до асимптотично ефективних оцінок

Ключові слова: М-оцінки, принцип Вінзора, робастні гребеневі оцінки

В множественной линейной регрессии, когда предсказатели сильно коррелированы, оценки наименьших квадратов (LSE), как правило, дают неточные прогнозы. Гребневая регрессия, основываясь на минимизации квадратичной функции потерь, чувствительна к выбросам. Рассмотрены две сглаженно сниженные ψ-функции, основанные на принципе Винзора, которые приводят к асимптотически эффективным оценкам

Ключевые слова: М-оценки, принцип Винзора, робастные гребневые оценки

УДК 519.6

DOI: 10.15587/1729-4061.2015.37316

УЛУЧШЕННЫЕ РОБАСТНЫЕ ГРЕБНЕВЫЕ ОЦЕНКИ РЕГРЕССИИ

В. И. Грицюк

Кандидат технических наук, доцент

Кафедра проектирования и эксплуатации электронных аппаратов

Харьковский национальный университет радиоэлектроники

пр. Ленина, 14, г. Харьков, Украина, 61166

E-mail: astra_kk12@mail.ru

1. Введение

Гребневая регрессия чувствительна к выбросам. Гребневая регрессия и робастная регрессия были предложены для решения этой проблемы мультиколлинеарности и выбросов в классической линейной регрессионной модели соответственно. Эта статья предлагает робастную и гребневую регрессии 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.

2. Анализ литературных данных и постановка проблемы

В этой статье мы рассмотрим гребневые оценки. В множественной линейной регрессии, когда предсказатели тесно связаны, оценки наименьших квадратов (LSE) дают неточные прогнозы. В попытке исправить это, Hoerl и Kennard предложили гребневую регрессию [1, 2]. Они добавили штраф, который создаёт небольшое смещение для того, чтобы одновременно уменьшить оценку и уменьшить дисперсию, что приводит к повышению общей точности прогнозирования. Многие проблемы регрессии состоят как в мультиколлинеарности, так и в ненормальности в той или иной степени. Известно, что метод НК 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 Хьюбером [5–3]. Холланд изучал совокупную проблему, и предложил использовать взвешенную гребневую регрессию с робастным выбором весов. В статье представлен подход, основанный на сочетании математических формулировок программирования как гребневой регрессии, так и робастной регрессии. Искомые коэффициенты регрессии могут быть легко вычислены путем итератив-

ной ревзвешенной процедуры наименьших квадратов, примененной к расширенному набору данных. В результате робастные и гребневые оценки являются превосходными результатами по сравнению с только либо робастными, либо гребневыми оценками.

3. Цель и задачи исследования

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

Для достижения поставленной цели решались следующие задачи:

– анализ известных методов М-оценок, анализ сглаженно сниженных М-оценок, выбор ψ-функций с улучшенной эффективностью;

– разработка объединённых методов робастной и гребневой регрессии с использованием в качестве помехоустойчивой меры разброса $\hat{\sigma}$ медианы абсолютных отклонений, основанных для оценивания коэффициентов на итеративно ревзвешенном методе наименьших квадратов (IRLS) и учитывающих наличие в модели свободного члена (intercept);

– получение результатов моделирования по сравнению методов Андрюса, Тьюки, МНК и с применением разработанных объединённых методов робастной и гребневой регрессии на основе выбранных ψ-функций.

4. Материал и результаты исследований робастной и гребневой регрессии, результаты моделирования

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). Введём новое семейство асимптотически более эффективных, сглаженно сниженных М-оценок. Этот новый подход основан на хорошо известном принципе Винзора (Winsor в Тьюки), в котором говорится, что все распределения являются нормальным явлением по происхождению.

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

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

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

1. $\rho(0) = 0$;

2. $\rho(t) \geq 0$;

3. $\rho(t) = \rho(-t)$;

4. for $0 \prec t_1 \prec t_2 \Rightarrow \rho(t_1) \leq \rho(t_2)$;

5. $\rho -$ непрерывная,

где ρ является симметричной функцией. Дифференцируя уравнение (1) по отношению к коэффициентам регрессии, получаем

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

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

где ψ является производной от ρ и x_i является вектор-строкой объясняющих переменных i-го наблюдения. М-оценка получается путем решения этой 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 для того, чтобы для нормально распределенных данных $\hat{\sigma}$ было 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 ω-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.

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

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

$$\psi(t) = \begin{cases} \mathrm{sign}(t)|t|, & \text{если} \quad 0 \leq |t| \prec a, \\ a\,\mathrm{sgn}(t), & \text{если} \quad a \leq |t| \prec b, \\ \{(c - |t|)/(c-b)\}\, a\,\mathrm{sgn}(t), & \text{если} \quad b \leq |t| \prec c, \\ 0, & \text{если} \quad c \leq |t|, \end{cases} \tag{6}$$

(Hoaglin и др.). Возникает потребность в ψ-функции сглаженно сниженной природы. Некоторые сглаженно

сниженные М-оценки были предложены время от времени. Реальные улучшения пришли от Андрюс (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 Андрюса, и бивейт-оценки Тьюки являются сглаженно-сниженными ψ-функциями. Потом Кадир (1996) предложил ψ-функцию, с весовой функцией бета-функцией с α=β. В последнее время Асад (2004) предложил другую ψ-функцию, которая имеет большую линейность в её центральной части. Волновая функция Андрюса

$$\psi(t)=\begin{cases} a\sin\left(\frac{t}{a}\right), & |t|\le \pi a, \\ 0 & \text{в друг. случ.} \end{cases} \qquad (7)$$

Бивейт-функция Тьюки

$$\psi(t)=\begin{cases} t\left[1-\left(\frac{t}{a}\right)^2\right]^2, & |t|\le a, \\ 0 & \text{в друг. случ.} \end{cases} \qquad (8)$$

Результаты моделирования по методу Андрюса, Тьюки в сравнении с методом МНК приведены ниже. В качестве примера исследован известный набор данных, взятый из Rousseeuw и Leroy (1987). Этот пример выбран, потому что этот реальный набор данных [6–8] 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). Были получены оценки коэффициентов, которые включены в уравнения:

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

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

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

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

Уравнение 1) включает коэффициенты, полученные МНК. Уравнение 2) содержит коэффициенты, полученные МНК с удалёнными точками 1, 3, 4 и 21. Уравнение 3) содержит соответственно коэффициенты, полученные методом Андрюса (a=1,5), уравнение 4) содержит коэффициенты, полученные с функцией бивейт Тьюки (a=4,685).

Из табл. 1 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.

Таблица 1

Остатки, полученные по методу Андрюса, методу Тьюки в сравнении с методом МНК

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

Асимптотическая вариация и эффективность М – оценок.

Для больших n можно выразить $\hat{\beta}$ как примерно нормально распределенное

$$D\left(\hat{\beta}\right)\approx N_p\left(\beta,\hat{\upsilon}\left(X^TX\right)^{-1}\right), \qquad (9)$$

где

$$\hat{\upsilon}=\hat{\sigma}^2\frac{ave_i\left\{\psi\left(r_i/\hat{\sigma}\right)^2\right\}}{\left[ave_i\left\{\psi'\left(r_i/\hat{\sigma}\right)\right\}\right]^2}\frac{n}{n-p}, \qquad (10)$$

где $ave_i(z_i)$ – среднее набора данных z.

На практике можно оценить

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

Принцип Винзора. Принцип Винзора гласит о том, что все распределения нормальные в середине. Таким образом, ψ-функция М-оценки должна быть похожа на ту, которая оптимальна для гауссовских данных в середине.

Новые ψ–функции. Предлагаются несколько новых ψ-функций, и обсудим их свойства по сравнению с другими ψ-функциями: функцией Андрюса и бивейт-функцией Тьюки. Предлагаемые ψ-функции [9] приведены ниже.

$$\psi_1(t)=\begin{cases} \frac{t}{2}\left(1-\left(\frac{t}{a}\right)^6\right)^2, & \text{если } |t|\le a, \\ 0, & \text{если } |t|\succ a. \end{cases} \qquad (11)$$

$$\psi_2(t)=\begin{cases}\frac{t}{2}\left(1-\left(\frac{t}{a}\right)^8\right)^2, & \text{если } |t|\le a,\\ 0, & \text{если } |t|\succ a,\end{cases} \quad (12)$$

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

ρ – функции, соответствующие ψ-функциям, приведенным выше, удовлетворяют стандартным свойствам, как правило связанным с обоснованной целевой функцией.

Асимптотическая эффективность предложенных M-оценок. Можно заметить, что сглаженно сниженные M-оценки ведут себя очень плохо, если ошибки действительно нормально распределены.

Видно [9], что асимптотическая вариация и эффективность обеих предложенных ψ функций, то есть ψ_1 и ψ_2, намного улучшены по сравнению с другими версиями.

Таким образом, используем следующие соотношения, применяя робастную и гребневую оценку.

Для набора данных регрессии (X, y) с $X\in R^{n\times p}$ и $y\in R^n$

$$\hat\beta=\left(\hat\beta_0,\hat\beta_1\right)=$$
$$=\arg\min\left\{L(X,y,\beta):\beta_0\in R,\beta_1\in R^p\right\}, \quad (13)$$

$$L(X,y,\beta)=\sum_{i=1}^{n}\rho\left(\frac{r_i(\beta)}{\hat\sigma_{ini}}\right)+\frac{\lambda}{\hat\sigma_{ini}^2}\|\beta_1\|^2, \quad (14)$$

$$r=(r_1,\ldots,r_n)^T=y-\hat\beta_0 1_n-X\hat\beta_1.$$

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

$$\hat\beta_0=\bar y-\bar x^T\hat\beta_1,$$
$$\left(X^TX+\lambda I_p\right)\hat\beta_1=X^T\left(y-\hat\beta_0 1_n\right), \quad (15)$$

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

Система уравнений, соответствующая робастной гребневой оценке (RRR)

$$\psi(t)=\rho'(t),\ W(t)=\frac{\psi(t)}{t}. \quad (16)$$

Пусть

$$\sigma=\hat\sigma(r(\hat\beta)),\ t_i=\frac{r_i}{\sigma},\ \omega_i=\frac{W(t_i)}{2},$$
$$w=(\omega_1,\ldots,\omega_n)^T,$$
$$W=\mathrm{diag}(w). \quad (17)$$

Приравняем производную по β в (14) нулю для RRR.

$$w^T\left(y-\hat\beta_0 1_n-X\hat\beta_1\right)=0,$$

$$\left(X^TWX+\lambda I_p\right)\hat\beta_1=X^TW\left(y-\hat\beta_0 1_n\right). \quad (18)$$

На основе исследований было обнаружено, что оценивание на основе обоих смещенных и робастных методов может быть полезным инструментом в тех случаях, когда наборы данных ухудшены одновременно от неортогональности и ненормальных ошибок. Процедура оценки состоит из увеличения исходного набора так, что обычный метод наименьших квадратов даст желаемую смещенную оценку данных. Затем многократно ревзвешенный метод наименьших квадратов, (IRLS) предполагающий итеративную процедуру, может быть использован для получения результирующих робастных и гребневых оценок. В результате моделирования получены оценки коэффициентов, включённые в уравнения:

5) $E(y)=-39.68+0.846\,x_1+0.421\,x_2-0.038\,x_3$,

6) $E(y)=-39.666+0.846\,x_1+0.426\,x_2-0.038\,x_3$.

Уравнения 5) и 6) содержат коэффициенты, полученные с применением робастных и гребневых оценок с функциями $\psi_1(t)$ (a=2,7) и $\psi_2(t)$ (a=2,6) соответственно. Параметр λ определяется согласно методу, приведенному в [10].

Таблица 2

Остатки для робастной гребневой регрессии с функциями $\psi_1(t)$ и $\psi_2(t)$

№	Остатки RRR с ψ_1	Остатки RRR с ψ_2
1	5,833	5,829
2	0,795	0,790
3	5,956	5,954
4	8,273	8,268
5	–0,875	–0,876
6	–1,301	–1,304
7	–0,500	–0,503
8	0,499	0,497
9	–0,915	–0,920
10	–0,050	–0,048
11	0,290	0,297
12	–0,321	–0,313
13	–2,974	–2,971
14	–1,984	–1,978
15	1,063	1,066
16	–0,051	–0,048
17	–0,007	–0,012
18	0,258	0,255
19	0,870	0,865
20	1,870	1,865
21	–9,644	–9,637

Из табл. 2 видно, что объединённые методы робастного и гребневого оценивания, основанные на ψ_1- и ψ_2-функциях, подтверждают факт, что наблюдения 1, 3, 4 и 21 являются выбросами, так как предложенные методы дают высокие величины остатков для этих наблюдений.

5. Выводы

Метод итеративно ревзвешенных наименьших квадратов (IRLS) на основе предложенных ψ-функций может быть использован для получения результирующих робастных гребневых оценок для выявления выбросов и игнорирования выбросов с нулевыми весами. 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_1 и ψ_2 – функциями, автоматически находят эти коэффициенты и определяют выбросы.

Литература

1. Owen, A. A robust hybrid of lasso and ridge regression [Text] / A. Owen. – Technical report, Stanford University, CA, 2006. – P. 1–14.
2. Cortez, P. Modeling wine preferences by data mining from physicochemical properties [Text] / P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis// Decision Support Systems. – 2009. – Vol 47, Issue 4. – P. 547–553. doi: 10.1016/j.dss.2009.05.016
3. Alma, Ö. G. Comparison of Robust Regression Methods in Linear Regression [Text] / Ö. G. Alma // Int. J. Contemp. Math. Sciences. –2011. – Vol. 6, Issue 9.–P. 409–421.
4. Asad, A. A Modified M-Estimator for the Detection of Outliers [Text] / A. Asad, M. F. Qadir // Pakistan Journal of Statistics and Operation Research. – 2005. – Vol. 1. – P. 49–64.
5. Qadir, M. F. Robust Method for Detection of Single and Multiple Outliers [Text] / M. F. Qadir // Scientific Khyber. –1996. – Vol. 9. – P. 135–144.
6. Deniel, C. Fitting Equations to Data [Text] / C. Deniel, F. S. Wood. – John Wiley and Sons, New York,1999. – 459 p.
7. Rousseeuw, P. J. and Leroy A.M. Robust Regression and Outlier Detection [Text] / P. J. Rousseeuw, A. M. Leroy. – John Wiley and Sons. New York, 1987. – 334 p. doi: 10.1002/0471725382
8. Rousseeuw, P. J. Recent Development in PROGRESS [Text] / P. J. Rousseeuw, M. Hubert //Computational Statistics and Data Analysis. – 1996. – Vol. 21. –P. 67–85.
9. Asad, A. Regression Outliers: New M-Class ψ-Functions Based on Winsor's Principle With Improved Asymptotic Efficiency [Text] / A. Asad, M. F. Qadir, Salahuddin // Jornal of Statistics. – 2006. –Vol 13, Issue 1. – P. 67–83.
10. Грицюк, В. И. Модифицированный алгоритм наименьших квадратов и выбор модели [Текст] / В. И. Грицюк // Вестник национального технического университета "ХПИ". Серия Автоматика и приборостроение. – 2004. – № 17. – С. 47–50.