ПРОГНОЗИРОВАНИЕ

УДК 519.6

НЕЧЕТКИЙ РОБАСТНЫЙ МЕТОД ДЛЯ ОЦЕНИВАНИЯ РЕГРЕССИОННОЙ МОДЕЛИ В СТРАХОВАНИИ

В.И. Грицюк, доц., к.т.н.,
Харьковский национальный университет радиоэлектроники

Аннотация. *В случае, если набор данных имеет выбросы, используют робастные методы оценки параметров. Когда входные данные нечеткие и набор данных имеет выбросы, нечеткая робастная регрессия исследуется вместо только метода наименьших квадратов, либо только робастных методов. В этом случае весовая матрица определяется по отношению к функции принадлежности.*

Ключевые слова: *робастная регрессия, нечеткая регрессия, функция принадлежности.*

НЕЧІТКИЙ РОБАСТНИЙ МЕТОД ДЛЯ ОЦІНЮВАННЯ РЕГРЕСІЙНОЇ МОДЕЛІ У СТРАХУВАННІ

В.І. Грицюк, доц., к.т.н.,
Харківський національний університет радіоелектроніки

Анотація. *У випадку, якщо набір даних має викиди, використовують робастні методи оцінки параметрів. Коли вхідні дані нечіткі й набір даних має викиди, нечітка робастна регресія досліджується замість тільки методу найменших квадратів, або тільки робастних методів. У цьому випадку вагова матриця визначається по відношенню до функції приналежності.*

Ключові слова: *робастна регресія, нечітка регресія, функція приналежності.*

FUZZY ROBUST METHOD FOR REGRESSION MODEL ESTIMATION IN INSURANCE

V. Gritsyuk, Assoc. Prof., Ph. D. (Eng.),
Kharkiv National University of Radio Electronics

Abstract. *If the data set has outliers, robust methods of parameters estimation are used. When the input data are fuzzy and the data set has outliers, fuzzy robust regression is investigated instead of the least squares method or only robust methods. In this case the weight matrix is defined with respect to the membership function.*

Key words: *robust regression, fuzzy regression, membership.*

Введение

Наблюдения, имеющие большие остатки, чем другие, называются выбросами. Процедура называется робастной, если она нечувствительна к появлению больших ошибок в данных. В таких случаях предпочтительнее робастные методы, чем метод наименьших квадратов (LS). Нечеткий регрессионный анализ – это нечеткий тип регрессионного анализа, который применяется для вычисления функциональных соотношений между зависимыми и независимыми переменными в

случае нечетких событий. В этом исследовании, когда входные данные – нечеткие числа

$$(X_i = (x_i, \underline{\xi}_i, \overline{\xi_i}),\ Y_i = (y_i, \underline{\eta}_i, \overline{\eta}_i))$$

и набор данных имеет выбросы, весовая матрица определяется по отношению к функции принадлежности. Поэтому актуальным является создание объединенных методов робастного и нечеткого метода наименьших квадратов для минимизации негативных воздействий выбросов на модель.

Анализ публикаций

Исследуется множественная регрессионная модель, когда зависимые и независимые переменные представлены треугольными нечеткими числами и оценки параметров – четкими числами. Tanaka и др. [1] предложил изучение линейной регрессии с нечеткой моделью. Однако приближение Tanaka может давать некорректную интерпретацию результатов нечеткой линейной регрессии, когда набор данных содержит выбросы.

Yang and Liu [2] предложили использовать алгоритм нечетких наименьших квадратов для моделей линейной регрессии. Этот алгоритм робастный против выбросов для простой регрессии. В алгоритме ортогональные условия добавлены для решения проблемы оптимизации. В Rousseeuw и др. [3] (1984) рассматривается простая регрессионная модель. Кроме этого, зависимая и независимая переменные представлены как четкие (crisp) числа и оценки параметров – четкие числа.

Цель и постановка задачи

Объект исследования – процесс разработки объединенных методов робастного и нечеткого метода наименьших квадратов.

Целью настоящей работы является исследование и разработка объединенных методов робастного и нечеткого метода наименьших квадратов, в которых возможные негативные последствия выбросов на модель сведены к минимуму.

Для достижения поставленной цели решались следующие задачи:
– анализ известных методов М-оценок, сглаженно-сниженных М-оценок;
– разработка объединенных методов робастной и нечеткой регрессии;
– получение результатов моделирования по сравнению методов Хьюбера, Хампеля, Тьюки, Андрюса, ψ-функции, МНК и с применением разработанного объединенного метода робастной и нечеткой регрессии.

Материал и результаты исследования робастной и нечеткой регрессии, результаты моделирования. Робастные методы

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

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

где ρ является симметричной функцией с минимумом в нуле

1. $\rho(0) = 0$;
2. $\rho(t) \geq 0$;
3. $\rho(t) = \rho(-t)$;
4. for $0 < t_1 < t_2 \Rightarrow \rho(t_1) \leq \rho(t_2)$;
5. ρ – непрерывная.

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

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

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

где ψ является производной от ρ и 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, чтобы они могли быть оценены одновременно.

Возможность состоит в использовании медианы абсолютных отклонений (MAD). Шкала оценки: $\hat{\sigma} = 1{,}483 med_i |r_i|$. Умножение на 1,483 сделано для того, чтобы для нормально распределенных данных $\hat{\sigma}$ являлось оценкой стандартного отклонения.

W-функция (весовая функция) для любого ρ определяется как

$$\omega(t_i) = \frac{\psi(t_i)}{t_i}. \quad (4)$$

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

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

Ψ-функция Хьюбера определяется как [4, 5]

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

где a – так называемая константа настройки (a=1,5).

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) $|t|$[6]. Состоящая из трех частей сниженная ψ-функция Хампеля определяется как

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

(Hoaglin и др.), a=1,7; b=3,4; c=8,5. Возникает потребность в ψ-функции сглаженно-сниженной природы. Некоторые сглаженно-сниженные М-оценки были предложены разными авторами. Улучшения были получены Эндрюсом (Andrews, 1974) и Тьюки (Mosteller и Tukey 1977; Hoaglin и др, 1983) [6], которые использовали волновые оценки (также называемые синус-оценки) и бивейт оценки соответственно. Волна Эндрюса и бивейт оценки Тьюки являются сглаженно-сниженными ψ-функциями. Кадир (1996) [6] предложил ψ-функцию с весовой функцией бета-функцией с α = β. Волновая функция Эндрюса (a=1,5)

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

Бивейт функция Тьюки

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

a = 4,685.

Новая ψ-функция

Новая ρ-функция предложена в семействе гладко-сниженных М-оценок [6]. ψ-функция, связанная с этой ρ-функцией, достигает 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 достигает целей, ради которых он построен, потому что дает улучшенные результаты во всех ситуациях и способен выдержать значительное количество выбросов. Предлагаемая ψ-функция [6] приведена ниже.

$$\psi(t) = \begin{cases} \frac{2t}{3}\left(1 - \left(\frac{t}{a}\right)^4\right)^2, & \text{если } |t| \le a, \\ 0, & \text{если } |t| > a, \end{cases} \quad (10)$$

где a – константа настройки (a=2) и для i-го наблюдения переменная t – остатки, шкалированные MAD; ρ – функция, соответствующая ψ-функции, приведенной выше, которая удовлетворяет стандартным свойствам, как правило, связанным с обоснованной целевой функцией.

Выбросы обладают большими остатками при робастной подгонке; помимо нечувствительности к выбросам, оценки робастной регрессии легко определяют выбросы.

Остатки из LS не могут быть использованы для этих целей, так как выбросы могу обладать очень малыми LS остатками [7].

Нечеткий робастный регрессионный анализ

Оценки LS, как известно, являются наилучшими, когда данные имеют нормально распределенные ошибки. В данном исследовании рассмотрим регрессионную модель с использованием нечетких чисел, когда X и Y – треугольные нечеткие числа, оценка параметров – четкие числа. В оценивании модели эвристика не допускается.

Матрица весов определяется функцией принадлежности остатков. Метод нечеткой робастной регрессии может обнаруживать выбросы автоматически, давая каждому из них степень принадлежности, которая равна нулю или очень мала по сравнению с другими степенями.

Треугольное нечеткое число определяется как $X=\left(m,\underline{m},\overline{m}\right)$, где m – модальная величина X; $\underline{m}$ – левосторонний разброс; $\overline{m}$ – правосторонній разброс.

Когда $X_i=(x_i,\underline{\xi}_i,\overline{\xi}_i)$ и $Y_i=(y_i,\underline{\eta}_i,\overline{\eta}_i)$, $i=1,2,\ldots,n$ – треугольные нечеткие числа, нечеткая регрессионная модель определяется как

$$Y=a+bX,$$

где a, b – четкие числа.

Когда параметры четкие, проблема оптимизации нечетких наименьших квадратов определяется как

$$\begin{aligned} &Minimum \quad r(a,b)=\sum d(a+bX_i,Y_i)^2 \\ &d(a+bX_i,Y_i)^2= \\ &=\left[a+bx_i-y_i-\left(b\underline{\xi}_i-\underline{\eta}_i\right)\right]^2+ \\ &+\left[a+bx_i-y_i+\left(b\overline{\xi}_i-\overline{\eta}_i\right)\right]^2+ \\ &+\left(a+bx_i-y_i\right)^2. \end{aligned} \tag{11}$$

В этом исследовании модель нечетких наименьших квадратов представлена обобщенной многомерной моделью $Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p+\varepsilon$. В этом случае проблема оптимизации определяется как

$$\begin{aligned} &Min\ r(a,b_1,b_2,\ldots,b_p)= \\ &=\sum d(a+b_1X_{1i}+\ldots+b_pX_{pi},Y_i)^2. \end{aligned} \tag{12}$$

Параметры оцениваются минимизацией уравнения (12). Итерационную процедуру продолжают до достижения разумной степени сходимости процесса.

Численный пример

В исследуемом примере данные собраны от хорошо известной страховой компании [8, 9]. X_1, X_2 и Y представляют номер месяца, количество требований в соответствующем месяце, платежи в соответствующем месяце. Эта структура – часть изучения Dalkilis и др. (2009). В Xu и Li (2001) величины разброса предполагались авторами [10]. В рассмотренном примере независимые переменные – нечеткие, в анализе нечеткой робастной регрессии (FRR) метода взяты величины независимых переменных центр, левый и правый разброс как x_i, $\underline{\xi}_i=x_i/8$ и $\overline{\xi}_i=x_i/7$ соответственно. Зависимая переменная величина – нечеткая, в FRR методе взяты величины зависимых переменных: центр, левый и правый разброс определены как y_i, $\underline{\eta}_i=y_i/8$ и $\overline{\eta}_i=y_i/7$ соответственно.

Таблица 1 Набор данных

X_1	X_2	$Y*10^4$	X_1	X_2	$Y*10^4$
1	1270	125	7	3169	631
2	2630	387	8	3448	545
3	3653	589	9	3163	583
4	3045	591	10	3096	606
5	3232	609	11	3765	753
6	3681	654	12	4481	898

Был выполнен анализ для LS, M и нечеткого робастного метода (FRR). Результаты получены используя набор данных, который приведен в табл. 1. Результаты анализа остатков показывают, что восьмое наблюдение является выбросом.

Оценки параметров регрессионной модели даны в табл. 2. Оценки параметров для FRR – такие же по знаку и почти такие же по величине, как и полученные робастными методами, хотя весовая матрица получена используя функцию принадлежности. Можно заметить, что на FRR не влияют выбросы.

Сумма квадратов ошибок дана в табл. 2 для LS, M и FRR методов. Можно отметить, что сумма квадратов ошибок больше для M методов, чем для других методов, благодаря величинам остатков с выбросами в робастных методах. Также можно заметить, что сумма квадратов ошибок для FRR метода ближе к данным M методов. Веса, полученные LS, M и FRR методами, показаны в табл. 3.

Таблица 2 Оценки параметров регрессии

Метод	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	Sum of Square Residual
LS	–118,4505	12,2627	0,1924	22503,91
Huber	–123,0438	14,1149	0,1908	22843,58
Hampel	–121,3489	13,8046	0,1908	22783,41
Tukey	–130,7773	16,5404	0,1847	25221,06
Andrews	–120,6023	13,238	0,1915	22636,21
ψ функция	–125,8294	17,1891	0,1824	25232,49
FRR	–103,12	15,5868	0,1808	23388,85

Таблица 3 Веса, полученные LS, M и FRR методами

Method	LS	Huber	Hampel	Tukey	Andrews	ψ функция	FRR
1	1,0000	1,0000	1,0000	0,9808	0,473	0,6642	0,9555
2	1,0000	1,0000	1,0000	0,9991	0,4625	0,6341	0,9611
3	1,0000	1,0000	1,0000	0,9828	0,4526	0,5803	0,9765
4	1,0000	0,4566	0,5497	0,0000	0,3407	0, 0000	0,2812
5	1,0000	0,7607	0,9362	0,0000	0,4267	0,3975	0,611
6	1,0000	1,0000	1,0000	0,9713	0,4734	0,6659	1,0000
7	1,0000	0,6787	0,8192	0,0000	0,4089	0,1658	0,5877
8	1,0000	0,3619	0,3526	0,0000	0,249	0,0000	0,0852
9	1,0000	1,0000	1,0000	0,7043	0,4639	0,6587	0,8617
10	1,0000	1,0000	1,0000	0,9999	0,4762	0,6666	1,0000
11	1,0000	1,0000	1,0000	0,9632	0,4753	0,6651	1,0000
12	1,0000	1,0000	1,0000	0,9933	0,4763	0,6665	1,0000

Весовая матрица получена через функцию принадлежности. Так минимизировано негативное влияние выбросов на модель. Получены оценки параметров регрессионной модели, где *X* и *Y,* треугольные нечеткие числа. В этом случае влияние выбросов меньше, чем в методе LS. Видно, что восьмое наблюдение является выбросом, так как имеет большой остаток (–99, 9409).

В расчетах зависимые и независимые переменные – четкие числа в LS и M оценках, в то время как в FRR – нечеткие числа. Оценки параметров регрессии – четкие числа во всех методах. Веса восьмого наблюдения «0,3619», «0,3525», «0», «0,249», «0» и «0,0852» для методов Huber, Hampel, Tukey, Andrew, ψ -функции и FRR соответственно. Веса, которые получены в FRR методе, – степень принадлежности каждого наблюдения. Эти принадлежности показывают воздействие наблюдений на модель. В табл. 3 показано, что выбросы влияют на модель очень малой степенью принадлежности, в то время как степень принадлежности других наблюдаемых величин 1, или близкая к 1.

Обсуждение результатов исследования робастной и нечеткой регрессии, результаты моделирования

Достоинством метода является то, что в данном случае нечеткая робастная множествен-

ная регрессия робастна для оценивания нечеткой регрессионной модели, особенно, когда существуют выбросы. Данный метод позволяет автоматически обнаруживать выбросы. Весовая матрица получена через функцию принадлежности, каждое наблюдение включается в оценку регрессионной модели в зависимости от степени принадлежности. Поэтому воздействие выброса на модель минимизировано. Главное преимущество заключается в обнаружении наблюдений для дальнейшего изучения.

Выводы

В результате проведенных исследований проведен анализ известных методов М-оценок, сглаженно-сниженных М-оценок (волны Эндрюса и бивейт оценки Тьюки, ψ-функции);

– разработан объединенный метод робастной и нечеткой регрессии (FRR);

– получены результаты моделирования по сравнению методов МНК, Хьюбера, Хампеля, Тьюки, Андрюса, ψ-функции и с применением разработанного объединенного метода робастной и нечеткой регрессии (FRR).

В проведенном исследовании метод нечеткой робастной регрессии предложен для построения модели для описания соотношения между зависимыми и независимыми переменными вместо метода наименьших квадратов и классического метода робастной регрессии. Исследована множественная регрессионная модель с использованием нечетких чисел, когда зависимая и независимая переменные являются треугольными нечеткими числами и оценки параметров – четкие числа. При рассмотрении табл. 2 видно, что оценки параметров регрессии, полученные методом нечеткой робастной регрессии, имеют тот же самый знак и почти ту же величину, что и оценки, полученные робастными методами. Также видно, что сумма квадратов ошибок FRR метода ближе к данным М методов.

Взвешенный метод нечетких наименьших квадратов применяется используя весовую матрицу, которая определяется из функции принадлежности остатков. Нечеткий робастный регрессионный метод может автоматически определять выбросы. Таким образом, возможное влияние выброса на модель минимизировано.

Литература

1. Tanaka H. Linear regression analysis with fuzzy model / H. Tanaka, S. Uegima and K. Asai // IEEE Trans. Systems Man Cybernet 12: . – 1982. – P. 903–907.
2. Yang M.S. and Liu H.H. Fuzzy Least Squares Algorithms for Interactive Fuzzy Linear Regression Models // FuzzySets and Systems. – 2003. – Vol. 135. – P. 305–316.
3. Rousseeuw P. Applying robust regression to insurance / P. Rousseeuw B. Daniels, A. Leroy // Insurance: Mathematics and Economics. – 1984. – Vol. 3. – P. 67–72.
4. Alma Ö.G. Comparison of Robust Regression Methods in Linear Regression / Ö.G. Alma // Int. J. Contemp. Math. Sciences. – 2011. –Vol. 6, Issue 9. – P. 409–421.
5. Qadir M.F. Robust Method for Detection of Single and Multiple Outliers / M.F. Qadir // Scientific Khyber. – 1996. – Vol. 9. – P. 135–144.
6. Asad A.A. Modified M-Estimator for the Detection of Outliers / A.A. Asad, M.F. Qadir // Pakistan Journal of Statistics and Operation Research. – 2005. – Vol. 1. – P. 49–64.
7. Rousseuw P.J. and Leroy A.M. Robust regression and outlier detection / P.J. Rousseuw and A.M. Leroy. – JohnWiley & Sons, New York, 1997. –334 p.
8. Kula K.S. Fuzzy Robust Regression analaysis Based on the Ranking of Fuzzy Sets / K.S. Kula, A. Apaydin // Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS). – 2008. – Vol. 16. – P. 663–681.
9. Kula Kamile Şanlı A study on fuzzy robust regression and its application to insurance / Kamile Şanlı Kula, Fatih Tank, Turkan Erbay dalk // Mathematical and Computational Applications. – 2012. – Vol. 17, No. 3. – P. 223–234.
10. Xu R. and Li C. Multidimensional least-squares fitting with a fuzzy model / R. Xu and C. Li // Fuzzy Sets and Systems. – 2001. – Vol. 119. – P. 215–223.

Рецензент: А.В. Бажинов, профессор, д.т.н., ХНАДУ.

Статья поступила в редакцию 01 апреля 2016 г.