Р

УДК 519.25

О. В. Рычка
Государственное образовательное учреждение высшего профессионального образования «Донецкий национальный технический университет», г. Донецк
283001,г. Донецк, ул. Артема, 58

АНАЛИЗ ЭФФЕКТИВНОСТИ УСОВЕРШЕНСТВОВАННЫХ МЕТОДОВ ПОИСКА И ОБРАБОТКИ АНОМАЛИЙ ДЛЯ НЕЛИНЕЙНЫХ МОДЕЛЕЙ С ВНУТРЕННЕЙ ЛИНЕЙНОСТЬЮ

O. V. Rychka
State Educational Institution of Higher Education “Donetsk national technical University”, Donetsk city 283001, Donetsk, Artema str., 58

EFFICIENCY ANALYSIS OF THE IMPROVED METHODS OF SEARCHING AND PROCESSING ANOMALIES FOR NONLINEAR MODELS WITH INTERNAL LINEAR

О. В. Ричка
Державна освітня установа вищої професійної освіти
«Донецький національний технічний університет», м. Донецьк
283001, м. Донецьк, вул. Артема, 58

АНАЛІЗ ЕФЕКТИВНОСТІ УДОСКОНАЛЕНИХ МЕТОДІВ ПОШУКУ І ОБРОБКИ АНОМАЛІЙ ДЛЯ НЕЛІНІЙНИХ МОДЕЛЕЙ З ВНУТРІШНЬОЇ ЛІНІЙНІСТЮ

В статье рассматривается задача обнаружения и обработки аномальных данных с целью повышения качества прогнозных регрессионных моделей. Выявлены основные недостатки существующих методов обнаружения ненадёжных данных. Предложены усовершенствованные методы для поиска и корректировки аномальных данных. Выделены основные критерии оценки эффективности разработанных методов. Выполнены численные исследования эффективности методов для нелинейных моделей с внутренней линейностью.
Ключевые слова: аномальные данные, регрессионный анализ, нелинейная регрессия, линейная модель, критерии эффективности

The article deals with the problem of detecting and processing anomalous data in order to improve the quality of predictive regression models. The main disadvantages of existing methods for detecting unreliable data are revealed. Improved methods for finding and correcting anomalous data are proposed. The main criteria for evaluating the effectiveness of the developed methods are described. Numerical studies of the effectiveness of the methods for nonlinear models with internal linearity are performed.
Keywords: anomalous data, regression analysis, nonlinear regression, linear model, performance criteria

У статті розглядається задача виявлення і обробки аномальних даних з метою підвищення якості прогнозних регресійних моделей. Виявлено основні недоліки існуючих методів виявлення ненадійних даних. Запропоновано вдосконалені методи для пошуку і коригування аномальних даних. Виділено основні критерії оцінки ефективності розроблених методів. Виконані чисельні дослідження ефективності методів для нелінійних моделей з внутрішньої лінійністю.
Ключові слова: аномальні дані, регресійний аналіз, нелінійна регресія, лінійна модель, критерії ефективності.

Общая постановка вопроса

Важным этапом анализа сложной системы является построение математической модели [1]. Построение математических моделей осуществляется на основании содержательной модели. Она состоит из входных параметров, внутренних параметров и выходных параметров. Одним из условий построения качественной адекватной модели является надёжность входных данных. Поэтому требуется первичная обработка экспериментальных данных с целью выявления ненадёжных измерений. Обнаружение аномальных значений относится к проблеме поиска образцов в данных, которые не соответствуют ожидаемому поведению. На сегодняшний день, существует большое количество работ, посвященных поиску аномальных значений [2-6]. Первые работы, касающиеся этой темы, были опубликованы ещё в XIX веке. Так, например, ещё в 1887 г. английский учёный Фрэнсис Эджуорт опубликовал статью "О противоречивых наблюдениях" [7].

В современной науке аномальные значения подразделяются на 2 класса:

- естественные;
- искусственные.

Естественные аномалии – это реальные факты или события, которые редко происходят.

Искусственные аномалии – это аномалии, вызванные ошибками ввода данных или ошибками измерений, а также некорректной работой оборудования.

Поиск ненадёжных измерений может иметь две различные цели – обнаружение выбросов и обнаружение новизны.

Новизна, в отличие от выбросов, не возникает в результате человеческих ошибок, а указывает на определённые изменения в системе. Основной задачей, в данном случае, является своевременное обнаружение аномалий при их появлении в выборке. После обнаружения новизны в данных, такие данные не подвергаются изменениям (отбрасыванию или корректировке). Примером задач, в которых осуществляется поиск новизны, являются: обнаружение мошенничества с кредитными картами, выявление сетевых вторжений, выявление неисправностей в функционирования оборудования.

Если целью анализа является построение наилучшей модели, то обнаруженные в экспериментальных данных выбросы подвергаются дальнейшей обработке – отбрасыванию или корректировке.

Отбрасывать аномальные данные рекомендуется, если выборка имеет достаточный объём и исключение определённого количества данных не повлияет на её репрезентативность.

В ином случае, возможно применение одного из основных методов корректировки:

- замена на наиболее вероятное значение;
- интерполяция данных;
- сглаживание данных;
- ручная замена.

Одним из распространенных инструментов анализа экспериментальных данных является регрессионный анализ. Частным случаем регрессионных моделей, имеющих большую область применения, являются парные линейные регрессионные модели. Их эффективность обусловлена тем, что они просты в понимании и применении, легко моделируются. Найденное линейное уравнение может быть начальной точкой для построения более сложных моделей. Ещё одним преимуществом парных линейных регрессионных моделей является возможность приведения большинства нелинейных моделей к линейному виду.

Анализ наиболее популярных на сегодняшний день методов обнаружения выбросов в исходных статистических данных позволил выявить ряд недостатков, основными из которых является чувствительность к объему выборки, большая трудоёмкость, существующие методы плохо формализованы и заключаются в переборе статистических данных, опираются на конкретные законы распределения вероятностей.

В связи с этим, совершенствование методов повышения качества парных линейных регрессионных моделей, основанных на поиске аномальных измерений в исходных данных, является актуальной задачей.

Целью статьи является описание усовершенствованных автором методов повышения качества линейных регрессионных моделей и доказательство их эффективности для нелинейных моделей с внутренней линейностью.

Для достижения поставленной цели были решены следующие задачи:

– усовершенствование и обоснование методов обработки статистических данных для построения более точных прогнозов;

– выбор критериев оценки эффективности предложенных методов;

– проведение математического моделирования с использованием экспериментальных данных для оценки эффективности рассматриваемых методов.

Описание используемых в работе моделей

С применением регрессионного анализа можно определить взаимосвязь между наблюдаемыми величинами, а также осуществить прогнозирование значения переменной.

Парная линейная регрессия – статистический метод, который позволяет по значениям независимой переменной X предсказывать значения зависимой переменной Y [8], [9], имеет следующий вид (1):

$$Y=\alpha X+\beta+\varepsilon, \quad (1)$$

где: α и β – параметры модели, определяемые в результате регрессионного анализа;

ε – случайные ошибки (невязки) переменной Y.

Параметры генеральной совокупности α и β нельзя определить точно, но можно найти их оценки a и b соответственно.

Таким образом, уравнение простой линейной регрессии примет вид (2):

$$\hat{Y}_i = aX_i + b, \quad (2)$$

где $\hat{Y}_i$ – предсказанное значение переменной Y для i-го наблюдения (i = 1..n);

X_i – значение переменной X для i-го наблюдения (i = 1..n);

a – коэффициент регрессии, который определяет точку пересечения прямой регрессии с осью ординат;

b – коэффициент регрессии, определяющий наклон к оси OX.

Для того чтобы найти оценки a и b, широко используется метод наименьших квадратов (МНК) [10].

В реальной жизни при моделировании различных процессов часто встречаются парные нелинейные зависимости. Например, многие экономические зависимости не являются линейными:

– эластичность спроса по цене описывается логарифмической моделью;

– производственная функция описывается степенной моделью;

– кривые Филипса представляются гиперболической функцией.

Существует два класса нелинейных регрессий:

– нелинейные по объясняющим переменным, но линейные по оцениваемым параметрам (полиномы, гиперболическая функция);

– нелинейные по оцениваемым параметрам.

Нелинейные по оцениваемым параметрам, в свою очередь, делятся на:

– нелинейные с внутренней линейностью (степенная, экспоненциальная, показательная и другие функции);

– нелинейные с внутренней нелинейностью.

Внутренне линейные модели можно привести к линейному виду путём определённых преобразований (как правило, логарифмированием). После этого производится поиск ненадёжных данных и дальнейшее их изменение, а далее путем обратного преобразования возвращаются к исходному нелинейному уравнению.

Описание разработанного метода для поиска и обработки выбросов

Для обнаружения данных, которые являются ненадёжными и представляют собой аномалии, необходимо определить область, в которую не попадают такие значения. Эта область будет областью надёжности. Для регрессионных моделей вида (2) область надёжности будет представлять прямоугольник со сторонами $2k\cdot\sigma_e$ и $2k\cdot\sigma'_e$, где значение k можно определить из формулы (3), а среднеквадратические отклонения невязок σ_eи σ'_eпо формулам (4) и (5) соответственно, используя таблицу значений Лапласса [11].

$$P_0 = 2\cdot\frac{1}{\sqrt{2\pi}}\int_0^k e^{-t^2/2}dt, \tag{3}$$

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - (a\cdot X_i + b))^2}{n-2}} \tag{4}$$

$$\sigma'_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - (a'\cdot X_i + b'))^2}{n-2}} \tag{5}$$

где a' и b' − коэффициенты перпендикуляра, построенного к исходному уравнению.

После обнаружения аномальных данных, которые не попали в область надёжности, исследователь может либо отбросить их, либо откорректировать таким образом, чтобы они попали в заданную область.

Корректировка данных производится следующим образом. При непопадании части измерений в область, определенную границами, которые параллельны линии исходного регрессионного уравнения, эти данные переносятся на уровень $A\cdot X_i + B \pm k\cdot\sigma_e$. При этом значения независимой переменной X_iостаются неизменными, а величины зависимой переменной Y_i меняются на соответствующие граничные значения.

При переносе ненадёжных наблюдений на границы области, которые параллельны линии, являющейся перпендикулярной к линии регрессии, полученной по соответствующему уравнению, меняются значения не только зависимой переменной Y_i, но и независимой X_i. Значения перемещаются по траектории параллельно исходному регрессионному уравнению.

Для определения новых значений X'_i, при перемещении на границы области, полученной как $A' \cdot X_i + B' + k \cdot \sigma'_e$, используется формула (6):

$$X'_i = \frac{(A \cdot X_i + B) - k \cdot \sigma' - B'}{A'} \quad (6)$$

Новые значения X'_i при переносе на уровень $A' \cdot X_i + B' - k \cdot \sigma'_e$ находятся по формуле (7):

$$X'_i = \frac{(A \cdot X_i + B) + k \cdot \sigma' - B'}{A'} \quad (7)$$

Графически метод поиска аномальных данных можно изобразить в виде рис. 1.

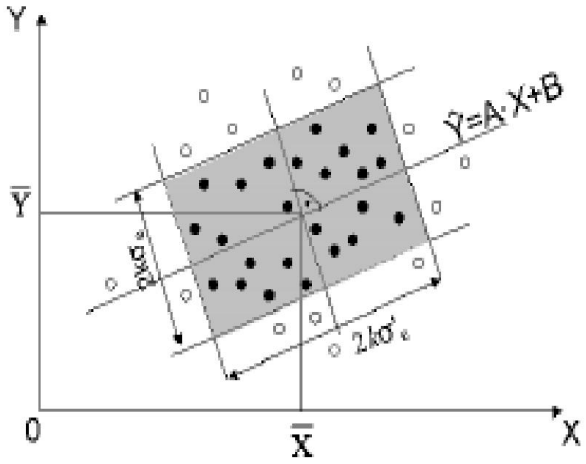

Рисунок 1 – График метода повышения качества модели

Для оценки эффективности рассматриваемых методов используются следующие критерии:

1) коэффициентом детерминации R^2 (8):

$$R^2 = \frac{\sum_{i=1}^{m} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{m} (Y_i - \bar{Y})^2}, \quad (8)$$

где $\overline{Y}$ − математическое ожидание случайной величины Y_i;

2) доверительный интервал прогнозных значений $Y_{прогн}$ – геометрическое место расположения прогнозных значений $Y_{прогн}$ при заданном значении $X_{прогн}$ и заданной доверительной вероятности $P_{дов}$.

В данной работе применялись доверительные интервалы, свободные от закона распределения случайных величин невязок, поскольку использование квантилей Стьюдента не будет давать корректных результатов. Это связано с тем, что используемая статистика невязок будет иметь распределение, не соответствующее нормальному [12].

3) модуль величины смещения результата прогноза (9):

$$\Delta_{прогн} = |(a \cdot X_{прогн} + b) - (a_н \cdot X_{прогн} + b_н)|, \quad (9)$$

где: $a \cdot X_{прогн} + b$ – величина предполагаемого значения $Y_{прогн}$ при соответствующем $X_{прогн}$, полученная по линейному регрессионному уравнению, построенному по исходным данным;

$a_н \cdot X_{прогн} + b_н$ – величина предполагаемого значения $Y_{прогн}$ при соответствующем $X_{прогн}$, полученная по линейному регрессионному уравнению, построенному после отбрасывания или корректировки части исходных данных.

4) точность, определяющаяся по формуле (рассчитывается только в случае отбрасывания части статистических данных) (10):

$$T = R^2 \cdot \frac{m}{n}, \tag{10}$$

где n – количество исходных данных;

m – количество данных оставшихся после отбрасывания.

При этом наилучшим вариантом считается вариант, при котором величина коэффициента детерминации R^2 является максимальной, при обязательном условии, что $T \geq 0{,}5$.

Оценка эффективности рассматриваемых методов для нелинейных моделей с внутренней линейностью

Для анализа эффективности, предложенных автором методов повышения качества регрессионных моделей для нелинейных моделей с внутренней линейностью были использованы экспериментальные данные, которые описываются степенной моделью (рис. 2). Объём выборки составляет 50 пар значений. Первый метод, заключается в дальнейшем исключении найденных аномальных данных, а второй – в их последующей корректировке, согласно определённой методике.

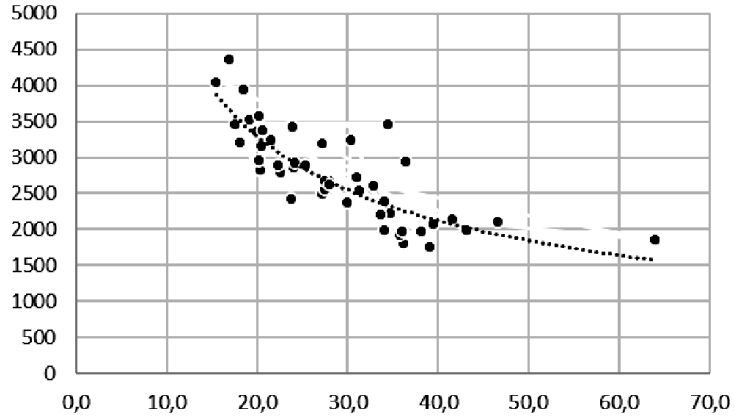

Рисунок 2 – Графическое изображение экспериментальных данных

С помощью МНК были найдены коэффициенты соответствующего уравнения регрессии. Данное уравнение имеет следующий вид: $Y=22042X^{-0{,}635}$. Коэффициент детерминации R^2 равен 0,71. Чтобы воспользоваться предложенными в работе методами повышения качества модели, воспользуемся следующими заменами: Y1=lnY, X1=lnX, B1=lnB. Тогда уравнение примет следующий вид: Y1= AX1+B1.

Воспользовавшись методом поиска аномальных данных, описанным в предыдущем разделе, были получены результаты, представленные ниже (табл. 1 и 2). В табл. 1 содержатся результаты применения метода после отбрасывания найденных аномальных измерений, а в табл. 2 – после корректировки.

Таблица 1 – Результаты метода, основанного на отбрасывании данных

Процент данных	R^2	DI,%	Delta,%	Количество данных	Точность
100	0,71	4,13		50	
90	0,84	2,52	1,00	46	0,77
85	0,82	2,47	0,70	43	0,70
80	0,82	2,45	0,51	42	0,69
75	0,84	2,17	0,75	40	0,67
70	0,83	2,15	0,71	38	0,63
65	0,81	2,09	0,40	34	0,55
60	0,84	2,00	0,65	32	0,54
50	0,81	1,10	0,14	23	0,37

Проанализировав данные (табл. 1), можно увидеть, что значение коэффициента детерминации увеличивается и достигает величины 0,84 при отбрасывании 4 ненадёжных измерений (что составляет менее 10% от общего количества данных). Коэффициенты линейного уравнения регрессии при этом равны: A=-0,729, B1=10,288. Теперь необходимо осуществить обратные преобразования, чтобы вернутся к степенной функции. Для этого необходимо найти значение коэффициента B из выражения: $B=e^{B1}$. Таким образом, новое уравнение степенной функции, описывающее наилучшую модель для рассматриваемых данных, будет иметь вид: $Y=29378X^{-0,729}$.

Таблица 2 – Результаты метода, основанного на переносе данных

Процент данных	R^2	DI,%	Delta,%
100	0,7091	4,13	
90	0,7570	3,18	0,35
85	0,7695	2,97	0,47
80	0,7706	2,88	0,56
75	0,7706	2,86	0,61
70	0,7700	2,97	0,69
65	0,7658	3,09	0,80
60	0,7589	3,27	0,92
50	0,7368	3,62	1,27

В данном случае значение коэффициента детерминации увеличивается на 6% и достигает значения 0,77 при корректировке 6 исходных данных.

На рис. 3 представлено наглядное изображение данных, которые изменяются.

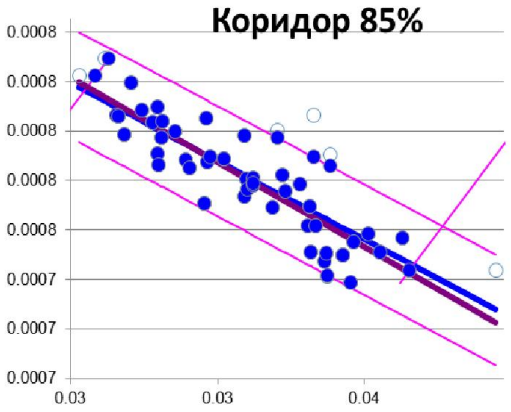

Рисунок 3 – Представление изменившихся данных эксперимента

При этом коэффициенты линейного уравнения равны A=-0,6875 и B1=10,1685. Тогда нелинейное уравнение примет вид: $Y=26069X^{-0,6875}$.

Как видно из полученных результатов, оба метода дают положительные результаты, что даёт основания рекомендовать их применение с целью повышения качества нелинейных моделей.

Выводы

Предлагаемые в статье методы, основанные на поиске и дальнейшей обработке аномальных данных, позволяют повысить точность прогнозных линейных регресссионных моделей. С помощью выбранных критериев эффективности, проведён анализ качества полученных регрессионных моделей на экспериментальных данных. Основные преимущества предложенных методов являются:

1. Использование описанных методов одинаково эффективно как для линейных, так и для нелинейных регрессионных прогнозных уравнений с внутренней нелинейностью.

2. Применение подхода, состоящего в обнаружении и дальнейшем изменении значений аномальных измерений, позволяет обеспечить его реализацию для выборок малого объёма, поскольку в отличие от метода, где аномальные данные отбрасываются, в этом подходе сохраняется исходное количество данных, что не уменьшает репрезентативность выборки.

3. Предложенные методы хорошо формализованы и позволяют наиболее быстро и точно проводить процедуру анализа данных на наличие грубых выбросов.

Список литературы

1. Александрова О. В. Групповой анализ и эргодичность стохастических процессов [Текст] / О. В. Александрова // Проблемы искусственного интеллекта. – Донецк: ГУ ИПИИ. – 2018. – № 4 (11). – С. 40–51.
2. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников [Текст] / Кобзарь А. И. − М. : ФИЗМАТЛИТ, 2006. − 816 с.
3. Тарасик В. П. Математическое моделирование технических систем [Текст] / Тарасик В. П. – Минск : Новое знание, 2013. – 584 с.
4. Rawlings John O. Applied regression analysis: a research tool. [Текст] / John O. Rawlings, Sastry G. Pentula, David A. Dickey – 2nd ed.− USA.: Springer, 1998.

5. Мудров В. И. Методы обработки измерений: Квазиправдоподобные оценки [Текст] / Мудров В. И., Кушко В. Л. − Изд. 2-е, перераб. и доп. − М.: Радио и связь, 1983. − 304 с.
6. Лемешко Б. Ю. Области применения критериев типа Граббса, используемых при отбраковке аномальных измерений [Текст] / Б. Ю. Лемешко, С. Б. Лемешко // Измерительная техника. – 2005. – № 6 – С. 13–20.
7. Edgeworth F. Y. On discordant observations [Текст] / Edgeworth F. Y. // The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. – 1887. – Vol. 23, no.5. – Pp.364–375.
8. Левин Дэвид М. и др. Статистика для менеджеров с использованием Microsoft Excel [Текст]. − 4-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2004. – 1312 с.
9. Левин В. И. Интервальная математика и построение прямых и обратных характеристик преобразователей информации [Текст] / В. И. Левин, Е. А. Немкова // Проблемы искусственного интеллекта. – Донецк : ГУ ИПИИ. – 2018. – № 1 (8). – С. 4–12.
10. Норман Р. Дрейпер. Прикладной регрессионный анализ [Текст] / Норман Р. Дрейпер, Гарри Смит. – 3-е изд.: Пер. с англ. – М. : Вильямс, 2007. – 912 с.
11. Ллойд Э. Справочник по прикладной статистике : в 2 т. [Текст] / Т. 1: Пер. с англ. / Под ред. Э. Ллойда, У. Ледермана, Ю. Н. Тюрина. − М. : Финансы и статистика, 1989. − 510 с.
12. Рычка О. В. Разработка и анализ метода повышения точности прогнозных регрессионных моделей и его модификаций [Текст] / О. В. Рычка // Питання прикладної математики і математичного моделювання : зб. наук. пр. / ред. кол....О. М. Кісельова (голов. ред.) та ін. –Д. : Вид-во Дніпропетр. нац. ун-ту, 2011. – С. 200–212.

References

1. O. V. Aleksandrova. Gruppovoy analiz i ergodichnost' stokhasticheskikh protsessov [Group analysis and ergodicity of stochastic processes]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], Donetsk, State Institution IPAI, 2018, No. 4 (11), P. 40–51.
2. Kobzar A.I. *Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov* [Applied Mathematical Statistics. For engineers and scientists], M., FIZMATLIT, 2006, 816 p.
3. Tarasik V.P. *Matematicheskoye modelirovaniye tekhnicheskikh sistem* [Mathematical modeling of technical systems], Minsk, New knowledge, 2013, 584 p.
4. John O. Rawlings, Sastry G. Pentula, David A. Dickey. *Applied regression analysis: a research tool*. 2nd ed. USA, Springer, 1998.
5. Mudrov V.I., Kushko V.L. *Metody obrabotki izmereniy: Kvazipravdopodobnyye otsenki* [Measurement processing methods: Quasi-likelihood estimates]. Ed. 2nd, rev. and add., M ., Radio and communication, 1983, 304 p.
6. Lemeshko B. Yu, Lemeshko S.B. Oblasti primeneniya kriteriyev tipa Grabbsa, ispol'zuyemykh pri otbrakovke anomal'nykh izmereniy [Expansion of the application area of Grubbs-type criteria used in the rejection of anomalous measurements]. *Izmeritelnaya tekhnika* [Measuring equipment], 2005, No. 6, P.13–20.
7. Edgeworth F. Y. On discordant observations. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1887, Vol. 23, no. 5, Pp. 364–375.
8. Levin, David M. et al. *Statistics for managers using Microsoft Excel* [Statistika dlya menedzherov s ispol'zovaniyem Microsoft Excel], 4th ed.: Per. from English, M ., Publishing house "Williams", 2004, 1312 p.
9. Levin V. I., Nemkova E. A. Interval'naya matematika i postroyeniye pryamykh i obratnykh kharakteristik preobrazovateley informatsii [Interval Mathematics And Construction Of Direct And Inverse Characteristics Of Information Converters]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], Donetsk, State Institution IPAI, 2018, no 1 (8), P. 4–12.
10. Norman R.Draper, Harry Smith. *Prikladnoy regressionnyy analiz* [Applied Regression Analysis], 3rd ed: Transl. from English, M., Williams, 2007, 912 p.
11. Lloyd E. *Spravochnik po prikladnoy statistike* [Handbook of Applied Statistics]. In 2 volumes.Vol. 1: Transl. from English / Ed. E. Lloyd, W. Lederman, Yu.N. Tyurin, M., Finance and statistics, 1989, 510 p.
12. Rychka O.V. Razrabotka i analiz metoda povysheniya tochnosti prognoznykh regressionnykh modeley i yego modifikatsiy [Development and analysis of a method for increasing the accuracy of predictive regression models and its modifications]. Pytannya prykladnoyi matematyky i matematychnoho modelyuvannya : zb. nauk. pr. [Questions of applied mathematics and mathematical modeling: coll] / ed. count . Oh. M. Kiselyova (head editor) and another, D ., Publ. of Dnipropetr. nat. un-ty, 2011, P. 200-212.

RESUME

O. V. Rychka

Efficiency Analysis of the Improved Methods of Searching and Processing Anomalies for Nonlinear Models with Internal Linear

The theme of detecting anomalous measurements in statistical data has been of interest to scientists since the 19th century. To date, dozens of different outlier detection methods have been described. However, they all have several disadvantages. Therefore, the task of improving methods for detecting and processing anomalous data is actual.

The paper proposes a method for searching for unreliable measurements and their further processing to improve the quality of forecasts made on the basis of linear regression models. The search method is to build a rectangular area. Its size depends on the specified probability and the values of the standard deviations. Data that falls into this area is considered reliable, and data outside the area is considered anomalous. After the discovery of such data, they are subject to either discarding or correcting, in accordance with certain rules.

Using the described performance criteria, it was found that the proposed methods can improve the quality of linear regression models, thereby increasing the forecast accuracy. In addition, the effectiveness of the proposed methods has been experimentally proved for nonlinear models with internal linearity.

The proposed method for detecting unreliable measurements shows high efficiency for samples of various sizes, is well formalized, which will allow it to be successfully implemented in software packages, and can be used both for linear regression models and nonlinear models with internal linearity.

РЕЗЮМЕ

О. В. Рычка

Анализ эффективности усовершенствованных методов поиска и обработки аномалий для нелинейных моделей с внутренней линейностью

Тема обнаружения аномальных измерений в статистических данных интересовала учёных ещё с XIX века. На сегодняшний день описаны десятки различных методов обнаружения выбросов. Однако все они имеют ряд недостатков. Поэтому задача усовершенствования методов обнаружения и обработки аномальных данных является актуальной.

В работе предложен метод поиска ненадёжных измерений и дальнейшей их обработки для повышения качества прогнозов, сделанных на основе линейных регрессионных моделей. Метод поиска заключается в построении прямоугольной области. Её размер зависит от заданной вероятности и значений среднеквадратических отклонений. Те данные, которые попадают в эту область, считаются надёжными, а данные, выходящие за границы области – аномальными. После обнаружения таких данных, они подлежат либо отбрасыванию, либо корректировке, в соответствии с определёнными правилами.

С использованием описанных критериев эффективности было получено, что предлагаемые методы позволяет повысить качество линейных регрессионных моделей, тем самым увеличив точность прогноза. Помимо этого, экспериментально доказана эффективность предложенных методов и для нелинейных моделей с внутренней линейностью.

Предложенный метод обнаружения ненадёжных измерений показывает высокую эффективность для выборок различного объема, хорошо формализуем, что позволит его успешно реализовать в пакетах прикладных программ, может использоваться как для линейных регрессионных моделей, так и для нелинейных моделей с внутренней линейностью.

Статья поступила в редакцию 04.09.2020.