УДК 519.254

ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ ПРОГНОЗИРОВАНИЯ ПРИ ИСПОЛЬЗОВАНИИ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

О.В.Рычка
Донецкий национальный технический университет

В работе рассматриваются сущность новых методов повышение эффективности прогнозирования, основные достоинства данных методов, в сравнении с уже известными. Описаны этапы работы программы, разработанной автором для реализации предложенных методов.

В настоящее время в различных сферах человеческой деятельности для прогнозирования часто используются регрессионные прогнозные модели. Для качественного прогноза необходимо, чтобы статистические данные были надежными и достоверными. Поэтому существует необходимость в обработке данных с целью выявления значений, которые значительно отличаются от остальных. Эти значения представляют собой аномальные измерения. Для обнаружения таких измерений существует ряд специальных критериев [1-3]. В ходе исследования были выявлены существенные недостатки данных методов [4]. Их основным недостатком является, то, что на "аномальность" исследуется по одному значению, что при большом количестве исходных статистических данных увеличивает трудоемкость использования данных критериев. В связи с выявленными недостатками существующих методов в [4] и [5] были предложены новые методы повышения качества регрессионных прогнозных моделей.

Суть данных методов заключается в том, что находятся данные, которые не попадают в прямоугольную область со сторонами $2k\cdot\sigma_e$ и $2k\sigma'_e$, где k − коэффициент, соответствующий вероятности попадания в заданную область. Данная вероятность рассчитывается по формуле (1):

$$P_0 = 2\cdot\frac{1}{\sqrt{2\pi}}\int_0^k e^{-t^2/2}dt - 1 \qquad (1)$$

Среднеквадратические отклонения невязок σ_e и σ'_e, определяются по формулам (2) и (3):

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}, \quad (2)$$

$$\sigma'_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}'_i)^2}{n-2}}, \quad (3)$$

где Y_i – фактические значения;

$\hat{Y}_i$ − рассчитанные значения по исходному уравнению;

$\hat{Y}'_i$ − рассчитанные значения по уравнению перпендикуляра.

Основное отличие предложенных методов между собой заключается в том, что в одном методе данные, не попадающие в прямоугольную область отбрасываются (рис.1), а в другом − корректируются (рис.2).

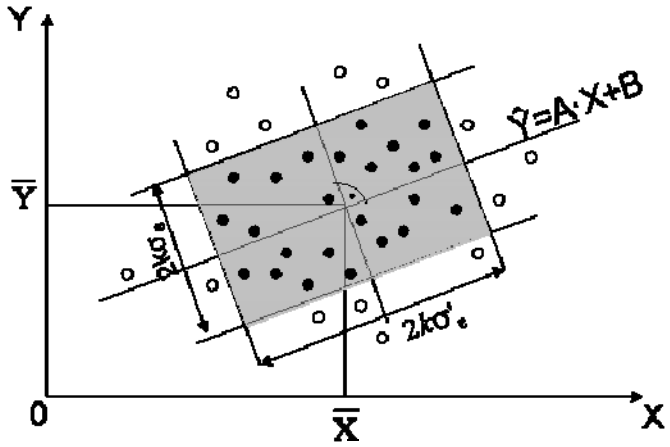

Рисунок 1 – Метод, основанный на отбрасывании данных

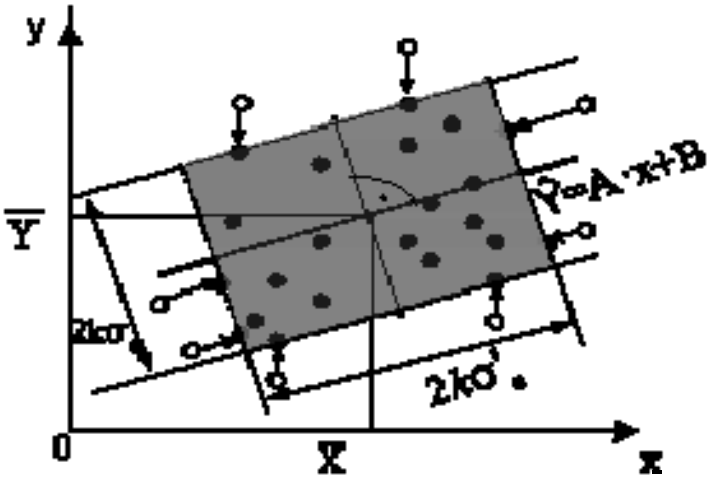

Рисунок 2 – Метод, основанный на переносе данных

Критерии оценки эффективности использования, описанных выше методов:

1. Коэффициент детерминации R^2, который демонстрирует силу взаимосвязи между Y и X (процент случайных величин x_i, который полностью объясняет поведение случайных величин y_i данным линейным регрессионным уравнением);

2. Модуль величины смещения результата прогноза Δ (является следствием изменения положения нового регрессионного уравнения при отбрасывании части исходных измерений);

3. Доверительный интервал прогнозных значений $Y_{прогн}$ (представляет собой геометрическое место расположения прогнозных значений $Y_{прогн}$ при заданном значении $X_{прогн}$ и заданной доверительной вероятности $P_{дов}$);

4. Количество элементарных операций ЭВМ, которое необходимо для реализации отбрасывания аномальных и ненадежных измерений;

5. Точность, которая рассчитывается по формуле 4:

$$T = R^2 \cdot \frac{m}{n} \qquad (4)$$

где n – исходное количество данных;
m – количество оставшихся после отбрасывания данных или количество данных, которые не подверглись изменению.

При этом наилучшим вариантом, считается вариант, при котором величина коэффициента детерминации R^2 является максимальной, при обязательном условии, что $T \geq 0.5$.

Для удобства реализации предложенных методов автором была разработана программа с использованием языка программирования Visual Basic for Applications. Основные этапы работы программы заключаются в следующем:

1. Пользователь вводит исходные данные.

2. По нажатию на кнопку, программа работает следующим образом:

а) строится вариационный ряд (по независимой переменной X);

б) находятся коэффициенты линейного регрессионного уравнения;

в) рассчитывается исходное значение коэффициента детерминации R^2;

г) определяется исходная величина доверительного интервала;

д) в программе предусмотрено определенное количество рабочих листов для каждой вероятности P попадания исходных статистических данных в заданную область (начиная с P=1 и до P=0,5

с шагом 0,05). На каждом из этих листов, отображаются результаты произведенных расчетов (выводятся оставшиеся или модифицированные данные, в зависимости от используемого метода, определяются новые коэффициенты уравнения, коэффициенты детерминации R^2, величины доверительного интервала, величины смещения и величина точности);

е) формируется итоговая таблица с рассчитанными значениями всех предложенных критериев эффективности (рис.3).

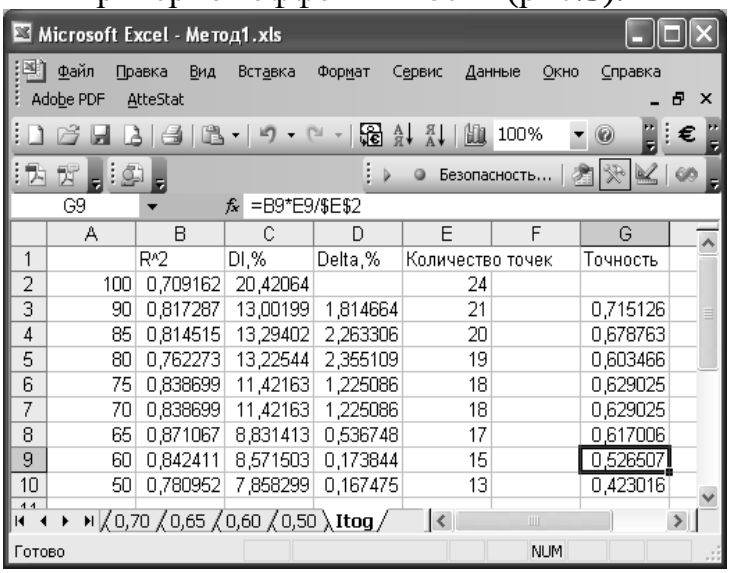

Рисунок 3 – Вид итоговой таблицы

3. В программе предусмотрено отображение графиков, изображающих соответствующие области с выделением данных, которые не попадают в определенную область (рис.4).

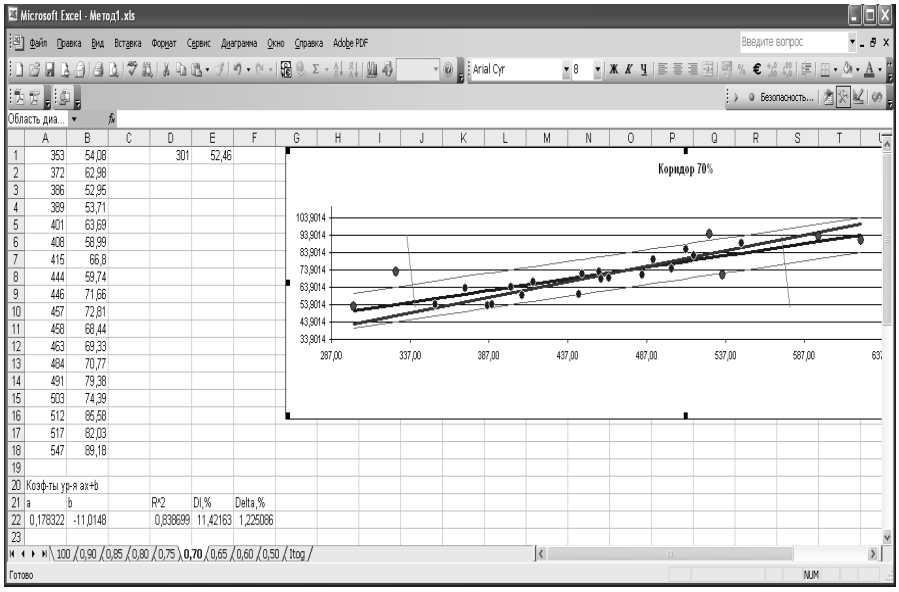

Рисунок 4 – Вид рабочего листа "0,70" после нажатия кнопки "Изобразить график"

4. По окончанию работы программы рабочие листы очищаются, нажатием кнопки "Delete".

Таким образом, в данной работе были описаны методы, позволяющие повысить точность прогнозирования при использовании линейных регрессионных моделей. Основными достоинствами данных методов является:

- простота понимания и применения;
- возможность найти и обработать несколько подозрительных значений одновременно, что позволяет избежать простого перебора данных;
- хорошая формализуемость, что позволило реализовать данные методы в компьютерных технологиях.

Разработанная программа позволяет:

- сократить время реализации методов;
- определить выигрыш от использования методов;
- наглядно изобразить результаты работы методов.

Список литературі

1. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. − М.: ФИЗМАТЛИТ, 2006. − 816 с.

2. Дрейпер Н.Р., Смит Г. Прикладной регрессионный анализ. 3-е изд.: Пер. с англ. − М.: Вильямс, 2007. − 912 с.

3. Rawlings, John O. Applied regression analysis: a research tool. — 2nd ed. / John O. Rawlings, Sastry G. Pentula, David A. Dickey − USA.: Springer, 1998.

4. Смирнов А.В., Рычка О.В. Метод повышения качества прогнозных регрессионных моделей. // Наукові праці Донецького національного технічного університету. Серія "Інформатика, кібернетика та обчислювальна техніка". Випуск 12(165) – Донецьк: ДВНЗ "ДонНТУ". – 2010. – С.141-147.

5. Смирнов А.В., Рычка О.В. Новый метод улучшения качества прогнозных регрессионных моделей. // Наукові праці ДонНТУ Серія "Інформатика, кібернетика та обчислювальна техніка". Випуск 13(185) – Донецьк: ДВНЗ "ДонНТУ". – 2011. – С.168-172.

Отримано 10.07.2011