УДК 519.254

Описание и программная реализация методов обработки данных для повышения точности прогнозирования

О.В. Рычка
Донецкий национальный технический университет
olga_rychka@mail.ru

Рычка О.В. Описание и программная реализации методов обработки данных для повышения точности прогнозирования. *В работе описаны методы, позволяющие повысить точность прогнозирования при использовании линейных регрессионных моделей и их основные преимущества. Рассмотрены этапы работы программного приложения, разработанного автором на языке программирования Visual Basic for Applications для реализации предложенных методов.*

Введение

В настоящее время прогнозирование используется во всех сферах человеческой деятельности. На основании прогноза принимаются обоснованные решения. Таким образом, большое значение имеет точность измерений и прогнозов, сделанных на основе имеющихся данных.

Для прогнозирования значений переменной, часто используется регрессионный анализ. Регрессионные модели позволяют прогнозировать значения по данным объясняющих переменных. Однако в исследуемую выборку могут попасть отдельные результаты, значения которых значительно отличаются от остальных. Эти результаты представляют собой аномальные измерения. Для обнаружения таких измерений, в настоящее время, существуют различные критерии [1-5]. К данным методам относятся − метод Титьена-Мура-Бекмана, Эктона и Прескотта-Лунда. Их основным преимуществом является простота понимания и использования. Однако, в ходе исследований, был выявлен ряд существенных недостатков [6]. Основным из них является то, что при использовании данных критериев, нахождение аномальных данных происходит методом перебора. С увеличением объема выборки повышается и трудоемкость, т.к. расчеты необходимо проводить для каждого подозрительного значения в отдельности. Также в рассматриваемой выборке может оказаться большее число аномальных измерений, чем исследуется на выбросы. В связи с описанными недостатками существующих на данный момент критериев в [6] и [7] были предложены новые методы, позволяющие повысить качество регрессионных прогнозных моделей. Для упрощения использования предложенных методов и их модификаций автором было разработано программное приложение с использованием языка программирования Visual Basic for Applications (VBA).

Цель исследований

Целью данной работы является определение основных условий выбора одного из двух методов повышения точности прогнозирования или их модификаций, а также рассмотрение этапов работы программы, разработанной с помощью языка программирования VBA.

Описание методов

Задача предлагаемых методов заключается в том, что среди всех исходных статистических данных определяются измерения, которые выходят за пределы прямоугольной области со сторонами $2k\cdot\sigma_e$ и $2k\sigma'_e$. Здесь k − коэффициент (обычно $0{,}6 \le k < 3$), который соответствует вероятности попадания в заданную область. Вероятность попадания в область рассчитывается по формуле (1) [8]:

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

В таблице 1 представлены значения коэффициента k, соответствующие различным вероятностям P.

Таблица 1. Значения коэффициента k при различных вероятностях P

P	0.9	0.85	0.8	0.75	0.7	0.65	0.6	0.5
k	1.9	1.75	1.6	1.5	1.4	1.3	1.2	1.05

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

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

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

где Y_i – фактические значения;

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

$\hat{Y}'_i$ – рассчитанные значения по уравнению, график которого будет перпендикулярен исходному.

Сравнение методов

Отличие реализации методов состоит в том, что в одном методе статистические данные, выходящие за пределы прямоугольной области отбрасываются, а в другом, такие данные не исключаются, а переносятся на границы области.

Метод повышения качества прогнозной модели, основанный на отбрасывании исходных статистических данных следует применять в случае, если прогнозное значение $Y_{прогн}$ близко к среднему $\overline{Y}$, т.к. чем дальше от среднего значения $\overline{Y}$ отстоит $Y_{прогн}$, тем большей величины достигает смещение при использовании данного метода и наоборот. Метод, основанный на перемещении данных, применяется в обратной ситуации.

Схематически метод, основанный на отбрасывании "аномальных" и ненадежных измерений, а также метод, основанный на переносе данных представлены на рисунке 1.

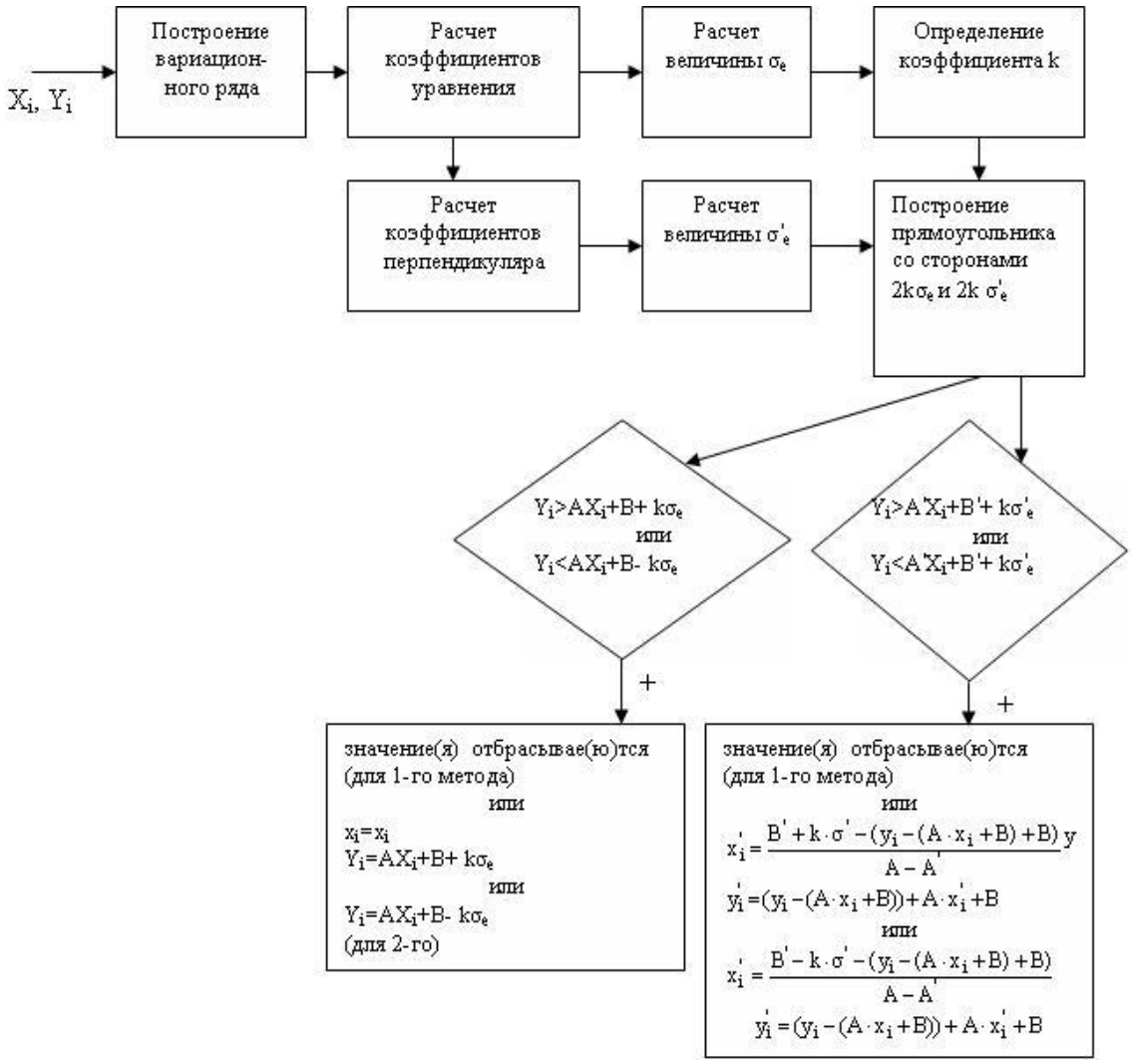

Рисунок 1 – Схематическое представление методов

Основными критериями эффективности применения предложенных методов являются:
– коэффициент детерминации R^2, который определяется соотношением (4) [9]:

$$R^2 = \frac{\sum_{i=1}^{k}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{k}(Y_i - \overline{Y})^2}; \quad (4)$$

– модуль величины смещения результата прогноза $\Delta_{прогн}$ (5):

$$\Delta_{прогн} = |(A \cdot X_{прогн} + B) - (A_н \cdot X_{прогн} + B_н)|, \quad (5)$$

где первое и второе слагаемое, соответственно, линейное регрессионное уравнение до отбрасывания части статистики и линейное регрессионное уравнение после отбрасывания части статистики;
– доверительный интервал прогнозных значений $Y_{прогн}$ (представляет собой геометрическое место расположения прогнозных значений $Y_{прогн}$ при заданном значении $X_{прогн}$ и заданной доверительной вероятности $P_{дов}$);
– количество элементарных операций ЭВМ, необходимое для реализации методов;
– точность, которая рассчитывается по формуле 6:

$$T = R^2 \cdot \frac{m}{n} \quad (6)$$

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

Наилучший вариант при этом – вариант, при котором величина коэффициента детерминации R^2 максимальна, при обязательном условии, что $T \geq 0.5$.

Описание программной реализации методов

Для удобства и быстроты реализации предложенных методов было разработано программное приложение с использованием языка программирования Visual Basic for Applications.

Работа программы содержит следующие этапы:

1. Пользователю предлагается ввести исходные данные в определенные ячейки (рис.2).

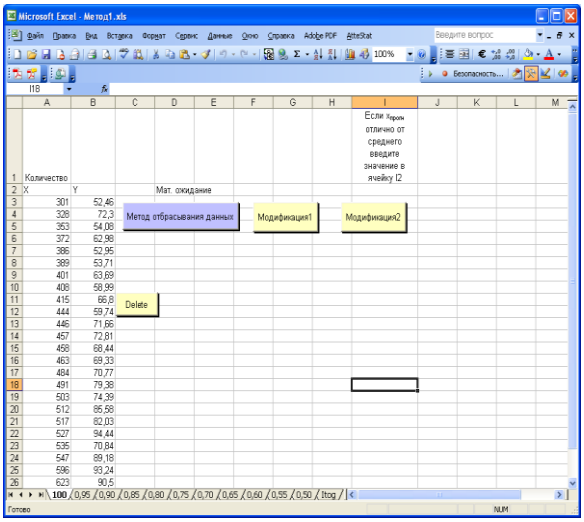

Рисунок 2 – Вид окна программы после введения данных

2. По умолчанию прогнозное значение $Y_{прогн}$ определяется при $X_{прогн}=X_{ср}$. Если же это не так, пользователь вручную может ввести в соответствующую ячейку произвольное значение $X_{прогн}$ при котором будет рассчитана величина $Y_{прогн}$.

3. Поскольку каждый из двух методов имеет по две модификации необходимо определить, следует ли применять сам метод или достаточно использовать одну из его модификаций. В случае если исследователь уверен в исходных значениях независимой переменной X (например, значения X_i даны заранее и в них не может быть случайных ошибок), следует применять первую модификацию метода. Она заключается в том, что область, в пределах которой, располагаются "надежные" данные представляет собой не прямоугольник, а коридор, границы которого равноудалены от линии исходного регрессионного уравнения. Расстояние между границами данного коридора составляет $2k\sigma_e$. В отличие от первой модификации метода повышения качества прогнозной модели во второй модификации находятся 2 линии параллельные графику, построенному перпендикулярно исходного. Расстояние между ними составляет $2k\sigma'_e$. При использовании второй модификации, отсекаются данные, которые являются аномальными и ненадежными по X. Величина коэффициента детерминации R^2 в отдельных случаях уменьшается. Однако это связано с тем, что крайние точки являются определяющими для уравнения регрессии, и если они удалены на значительное расстояние от остальных точек, то они оказывают большое

влияние на исходное уравнение. Следует отбрасывать (перемещать) не более 10-20% исходных статистических данных.

Реализация описанных выше модификаций проще, чем реализация предложенного метода, однако и первая и вторая модификации несколько снижают эффективность метода, поскольку в случае использования метода исключаются (переносятся) аномальные и ненадежные значения и по X и по Y одновременно. Т.о. картина является более точной.

После этого пользователь выбирает соответствующую кнопку. Нажатие кнопки вызывает следующие действия:

а) данные сортируются по возрастанию (по независимой переменной X);

б) рассчитывается количество введенных статистических данных;

в) определяются коэффициенты линейного регрессионного уравнения и уравнения график которого перпендикулярен исходному;

г) рассчитывается значение коэффициента детерминации R^2;

д) находится исходная величина доверительного интервала;

е) выводятся все вспомогательные значения (величины σ_e, σ'_e, $k\sigma_e$, $k\sigma'_e$ и т.д.) (рис.3).

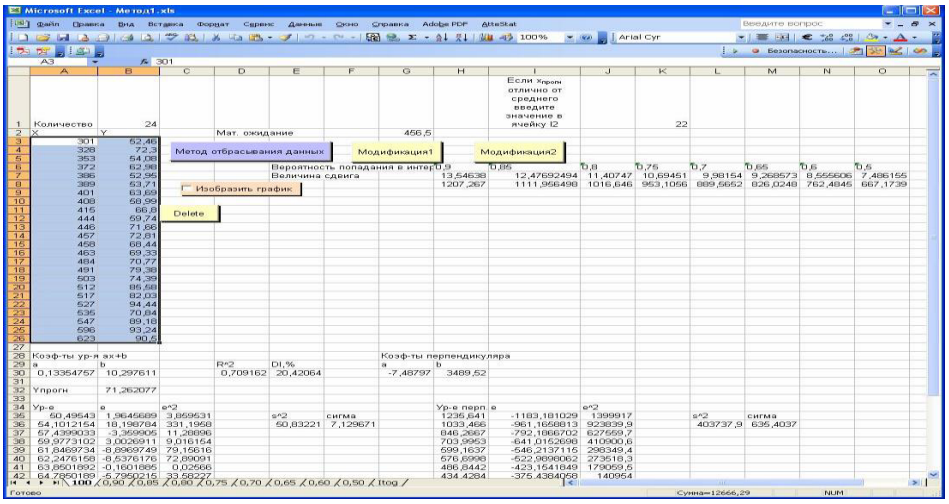

Рисунок 3 − Вид исходного листа после нажатия кнопки "Метод отбрасывания данных"

ж) Для каждой вероятности P попадания исходных статистических данных в определенную область (начиная с P=1 и до P=0.5 с шагом 0.05) предусмотрен отдельный рабочий лист, на котором отображаются результаты произведенных расчетов (выводятся оставшиеся или скорректированные данные, в зависимости от применяемого метода, определяются новые коэффициенты уравнения, коэффициенты детерминации R^2, величина доверительного интервала, величины смещения, количество элементарных операций и величина точности). Пользователь может переключаться между этими рабочими листами и смотреть всю необходимую ему информацию;

з) формируется итоговый рабочий лист, на котором в виде таблицы представлены данные по всем основным критериям эффективности метода для каждого значения вероятности попадания в заданную область. На основании данной таблицы пользователь может сделать вывод о том, до каких пределов следует уменьшать вероятность попадания в "коридор", а также увидеть полученный от применения метода выигрыш (рис.4).

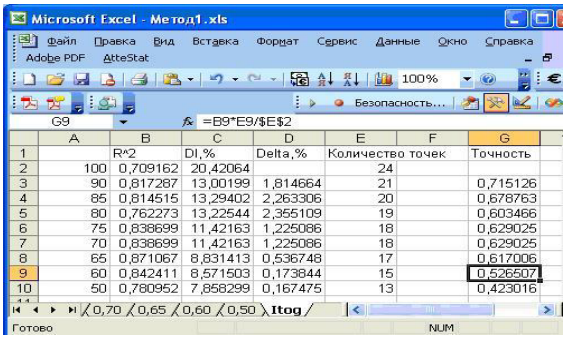

Рисунок 4 − Вид итоговой таблицы

4. Для наглядности пользователю доступна возможность отображения графиков, изображающих соответствующие области с выделением наблюдений, которые не попадают в

заданную область (см. рис.5). Одной из особенностей разработанной программы является возможность построения графика перпендикуляра, т.к. стандартные инструменты Microsoft Excel этого сделать не позволяют.

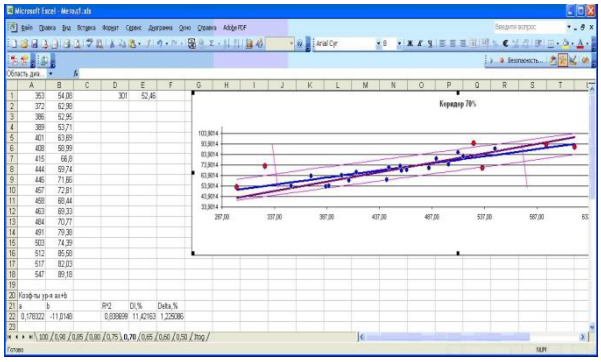

Рисунок 5 – Вид листа «0.70» после нажатия кнопки «Изобразить график»

5. По окончанию работы программы, все полученные данные можно сохранить с помощью стандартных инструментов программы Excel, а также очистить все рабочие листы нажатием кнопки "Delete".

Выводы

В данной статье были представлены методы, повышающие точность линейных регрессионных прогнозных моделей. Основными достоинствами данных методов является:

- простота понимания и применения;
- возможность нахождения и обработки нескольких аномальных и ненадежных измерений одновременно, что позволяет сократить время исследования, за счет отказа от простого перебора данных;
- хорошая формализуемость, что позволило реализовать данные методы в компьютерных технологиях.

Разработанное программное приложение имеет следующие преимущества:

1. Доступность понимания и использования.
2. Сокращение времени расчетов, по сравнению с ручной обработкой данных.
3. Возможность вводить произвольное количество исходных статистических данных.
4. Расчет прогнозного значения $Y_{прогн}$ для любого $X_{прогн}$.
5. Вывод основных результатов на отдельном рабочем листе для соответствующей вероятности попадания в определенную область, что позволяет проследить за основными изменениями, полученными в результате использования методов или их модификаций.
6. Отображение итоговой таблицы, которая позволяет определить выигрыш от применения метода и сделать вывод о величине, до которой следует уменьшать вероятность попадания в область.
7. Наглядное изображение областей, соответствующих выбранным вероятностям и данных, которые при этом являются "аномальными" и ненадежными.
8. Возможность изображения на графике перпендикулярных прямых, путем расширения стандартных инструментов Excel.
9. Не требуется предварительная установка на компьютере, котором уже есть Microsoft Office.
10. Маленький размер программного приложения (около 1 МБ).

Литература

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. Мудров В.И., Кушко В.Л. Методы обработки измерений: Квазиправдоподобные оценки. − Изд. 2-е, перераб. и доп. − М.: Радио и связь, 1983. − 304 с.

5. Лемешко Б. Ю, Лемешко С. Б. Расширение области применения критериев типа Граббса, используемых при отбраковке аномальных измерений// Измерительная техника. – 2005. – № 6 – С.13-20

6. 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.

7. 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.

8. Ллойд Э. Справочник по прикладной статистике. В 2-х т. Т.1: Пер. с англ./ Под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. − М.: Финансы и статистика, 1989. − 510 с.

9. *Абрамовиц М.* Справочник по специальным функциям с формулами, графиками и математическими таблицами./ Под. ред. М.Абрамовица и Н. Стигана: Пер. с англ. под ред. В.А. Диткина и Л.И. Кармазиной. − М.: Наука, 1979. − 830 с.

УДК 519.254

Рычка О.В. Описание и программная реализация методов обработки данных для повышения точности прогнозирования. В работе описаны методы, позволяющие повысить точность прогнозирования при использовании линейных регрессионных моделей и их основные преимущества. Рассмотрены этапы работы программного приложения, разработанного автором на языке программирования Visual Basic for Applications для реализации предложенных методов.
Ключевые слова: точность прогнозирования, линейная регрессионная модель, программное приложение, аномальные измерения.

УДК 519.254

Ричка О.В. Опис та програмна реалізація методів обробки даних для підвищення точності прогнозування. У роботі описано методи, що дозволяють підвищити точність прогнозування при використанні лінійних регресійних моделей і їх основні переваги. Розглянуто етапи роботи програмного додатка, розробленого автором мовою програмування Visual Basic for Applications для реалізації запропонованих методів.

Ключові слова: точність прогнозування, лінійна регресійна модель, програмний додаток, аномальні виміри.

UDC 519.254

Rychka O.V. Description and software implementation of data processing methods to improve the accuracy of forecasting. Methods to improve forecasting accuracy using linear regression models and their main advantages are described. The stages of work of software, developed by the author in the programming language Visual Basic for Applications to implementation the proposed methods are considered.
Key words: forecasting accuracy, linear regression model, software, the anomalous measurements

Статья поступила в редакцию 21.05.2016
Рекомендована к публикации д-ром техн. наук А.С. Миненко