Программное обеспечение многофакторного регрессионного анализа при нарушении закона о нормальном распределении наблюдений

А.И. Джангаров, Х. А. Ахметова

Чеченский государственный университет, г. Грозный

Аннотация: В данной статье рассматривается применение регрессионного анализа при нарушении предположения о нормальном законе распределения наблюдений. Проведенный анализ проиллюстрировал, что теоретическое распределение наблюдений может значительно отличаться от практических исследований, если отклоняется нормальный закон. Для решения данной проблемы, являющейся одной из центральных в статистическом анализе, были изучены и применены устойчивые методы оценивания. На основе полученных результатов был разработан программный продукт, реализующий данный статистический метод.
Ключевые слова: нормальный закон распределения наблюдений, устойчивые методы оценивания, М оценки, программная реализация.

Закон о нормальном распределении наблюдений является неотъемлемой составляющей в математической статистике. Это обусловлено существованием центральной предельной теоремы, которая гласит, что нормальное распределение наблюдений является предельным распределением суммы большого числа случайных величин, когда число проведенных опытов принимает довольно крупное значение [1]. Нормальный закон распределения наблюдений обладает целым рядом свойств; благодаря этим свойствам получена большая часть результатов математической статистики, широко используемых в различных методах регрессионного анализа. При нормальном законе распределения наблюдений, оценки параметров регрессионной модели совпадают с оценками других методов и имеют определенные оптимальные свойства [2]. Это свидетельствует о том, что данные анализа наблюдений и экспериментов в различных областях человеческой жизнедеятельности являются подлинными с довольно низкой вероятностью возникновения нарушений целостности данных и появления грубых ошибок в расчётах. Однако, если рассмотреть ситуации, в которых нормальный закон нарушается, то можно сделать вывод о том, что

вероятность возникновения грубых ошибок при экспериментальных наблюдениях резко возрастает. В связи с этим необходимо использовать оптимальные устойчивые методы оценивания распределения наблюдений.

Описание проблемы

Считается, что при нормальном законе распределения наблюдений вероятность возникновения грубых ошибок равна 0,003. График нормального распределения изображен на рисунке 1.

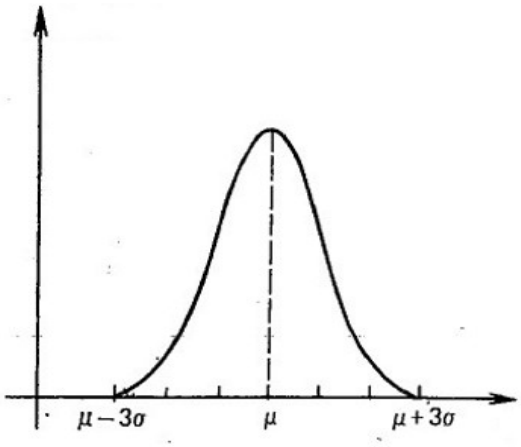

Рис. 1. – Нормальное распределение наблюдений

Характерная особенность этого распределения состоит в том, что преобладающая часть наблюдений сосредоточена в границах интервала $(\mu-3\sigma,\mu+3\sigma)$ с центром μ [3].

Исходя из этого, можно сделать вывод о том, что распределение имеет короткие хвосты (результаты наблюдений, оказавшиеся за пределами интервалов) и крайне низкую вероятность возникновения грубых ошибок.

Однако это можно отнести лишь к теоретическому распределению наблюдений. На рисунке 2 можно наблюдать, что график практического распределения наблюдений имеет гораздо более длинные хвосты, а также значительно возрастает по оси ординат. Это позволяет говорить о том, что возникновение грубых ошибок уже не является редким событием.

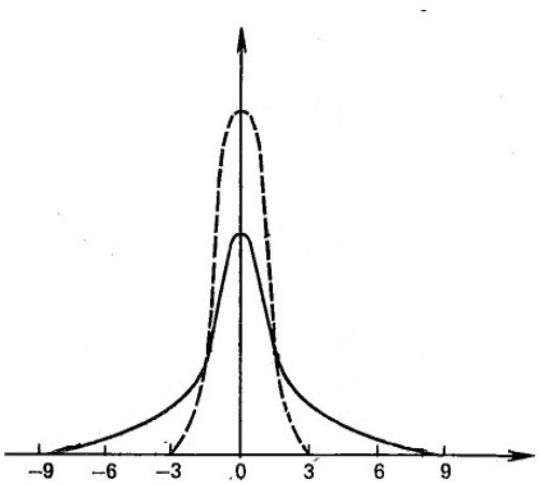

Рис. 2. – Практическое распределение наблюдений

Решение проблемы

Описанная выше проблема является одной из центральных в современном регрессионном анализе. Для её решения был разработан целый математический аппарат и введены устойчивые методы оценивания. Существует три основные группы оценок и в данном исследовании были рассмотрены так называемые М-оценки, предложенные швейцарским учёным Хубером.

Эти оценки, обозначаемые b_M, получают, решая задачу оптимизации, которая выглядит следующим образом:

$$\sum_{u=1}^{N}\rho\left(y_u-\sum_{i=1}^{k}\beta_i f_{iu}\right)=\sum_{u=1}^{N}\rho\left(e_u\right)-\min(\beta_1...\beta_k),$$

где $e_u=\sum\beta_i f_{iu}$ — остатки результатов наблюдений, а $\rho(z)$ — подходящая выбранная функция. Свойства оценок b_M, полученных на основе данной формулы, зависят от вида функции $\rho(z)$ [4]. Для обеспечения несмещенности, состоятельности и высокой эффективности М-оценок эта функция должна удовлетворять определенным условиям. Самое общее

требование — она должна расти медленнее, чем квадратичная функция, использующаяся в методе наименьших квадратов (классический метод исследования распределения наблюдений при нормальном законе). При таком выборе $\rho(z)$ функция и полученные М-оценки оказываются значительно менее чувствительными к грубым ошибкам, нежели оценки того же метода наименьших квадратов [5].

Также на функцию $\rho(z)$ накладывается следующее условие – она должна выбираться так, чтобы обеспечить минимальную дисперсионную матрицу оценок (диагональная матрица, элементы которой позволяют судить об интенсивности возникновения грубых ошибок) при самом неблагоприятном распределении грубых ошибок. Решение минимизационной задачи имеет вид:

$$\rho(z)=\begin{cases}(1/2)z^2 & \text{при } |z|\le c,\\ c|z|-(1/2)c^2 & \text{при } |z|>\text{c}.\end{cases}$$

Из этого можно сделать вывод, что оптимальная функция $\rho(z)$ не зависит от конкретного вида распределения грубых ошибок, а только от их интенсивности, которую принято обозначать – δ. Интенсивность грубых ошибок можно рассматривать, как относительную долю ошибок во всей совокупности [6].

Программная реализация

Для более детального изучения этой темы было разработано программное обеспечение, реализующее метод регрессионного анализа (при отклонении нормального закона распределения наблюдений).

Алгоритм программы использует математическую составляющую, описанную в М-оценках. Дополнительно были интегрированы такие важные критерии, как классический критерий Фишера (F-критерий) и критерий Стьюдента [7, 8, 9].

После ввода экспериментальных данных, программа позволяет оценить влияние каждого фактора на итоговый результат исследования, а также на основе критериев Фишера и Стьюдента сделать вывод об устойчивости модели, что в рамках отклонения нормального закона распределения первостепенно [10].

Диалоговое окно для ввода данных наблюдений выглядит следующим образом:

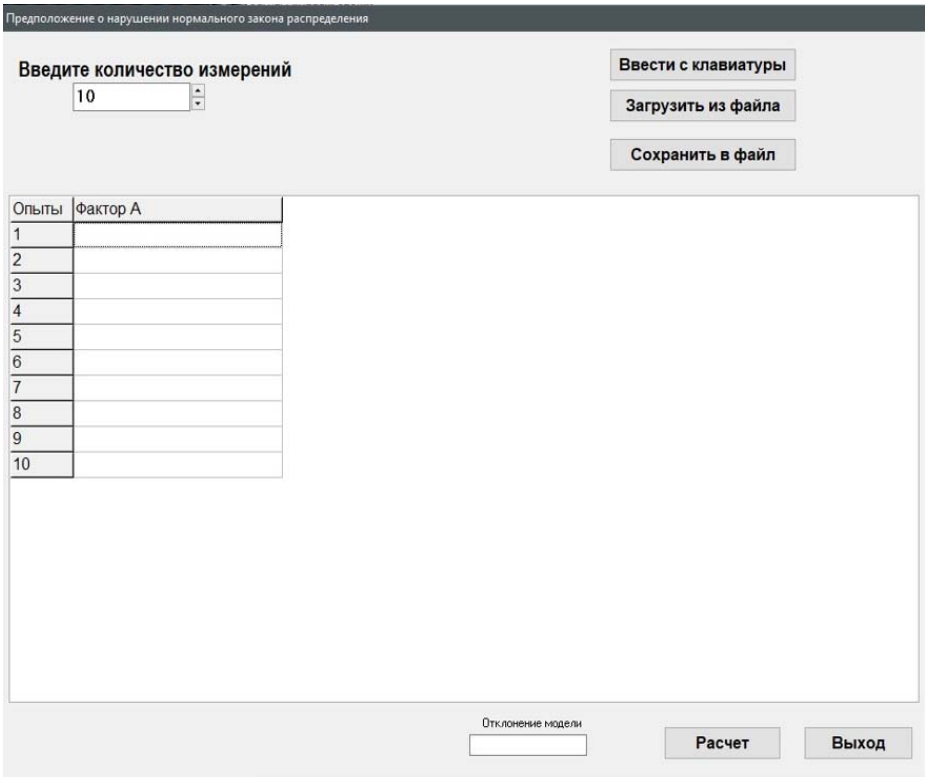

Рис. 3. – Окно ввода данных

Указав параметр δ (отклонении модели) и нажав на кнопку «Расчет», получаем следующую информацию:

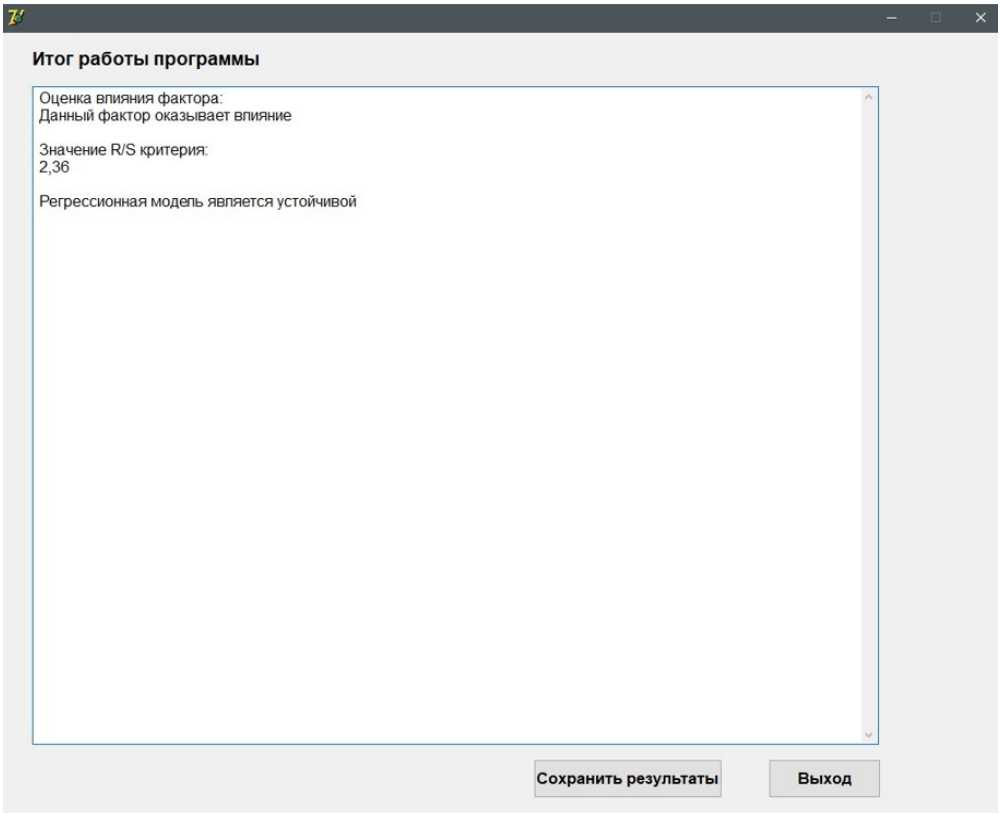

Рис. 4. – Результаты работы программы

Заключение

Регрессионный анализ является мощным статистическим инструментом, основная задача которого – оценка влияния факторов на выходной параметр. В ходе данной научно-исследовательской работы было установлено, что подобные оценки нуждаются в дополнительной проверке и интерпретации. И как оказалось, этого можно достичь при помощи различных математических моделей и критериев.

Результаты исследований продемонстрировали, что устойчивые методы оценивания распределения наблюдений способны серьезно облегчить и ускорить анализ данных. Использование М-оценок в устойчивых моделях распределения наблюдений привело к получению достоверных результатов с небольшой вероятностью возникновения грубых ошибок. С их внедрением и автоматизацией процесса – исследование практических

результатов – уже не является трудной задачей, требующей большого количества времени.

Литература

1. Norman R. Draper, Harry Smith. Applied Regression Analysis, third edition. 1998. 473 p.

2. Jacob Cohen. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. 2002. 691 p.

3. Воскобойников, Ю.Е. Регрессионный анализ данных в пакете Mathcad. СПб.: Лань, 2011. 224 с.

4. Дрейпер, Н. Прикладной регрессионный анализ. М.: Вильямс. 2007. 706 с.

5. Яковлев М. Я. Янгирова А. В. Метод и результаты численной оценки эффективных механических свойств резинокордных композитов для случая двухслойного материала // Инженерный вестник Дона, 2013, №2 URL: ivdon.ru/ru/magazine/archive/n2y2013/1639

6. Соколов, Г.А. Введение в регрессионный анализ и планирование регрессионных экспериментов в экономике. М.: ИНФРА-М, 2012. 202 с.

7. Magomedov I. A. Mezhieva A.I. Suleymanova M.A. Analysis and methods of squeal reduction of disk brake system. Инженерный вестник Дона, 2018, №4. URL: ivdon.ru/ru/magazine/archive/n4y2018/5334

8. Носко В.П. Регрессионный анализ временных рядов. М.: ИД РАНХиГС, 2011. 672 с.

9. Julian J.Faraway. Extending the Linear Model with R. 2006. 324 p.

10. Frank E. Harrell, Jr. Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression and Survival Analysis. 2015. p. 3

References

1. Norman R. Draper, Harry Smith. Applied Regression Analysis, third edition. 1998. 473 p.

2. Jacob Cohen. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. 2002. 691 p.

3. Voskobojnikov, Yu.E. Regressionny`j analiz danny`x v pakete Mathcad. [Regression analysis of data in the Mathcad package]. SPb.: Lan`, 2011. 224 p.

4. Drejper, N. Prikladnoj regressionny`j analiz. [Applied Regression Analysis]. M.: Vil`yams. 2007. 706 p.

5. Yakovlev M. Ya. Yangirova A. V. Inženernyj vestnik Dona (Rus), 2013, №2. URL: ivdon.ru/ru/magazine/archive/n2y2013/1639/

6. Sokolov, G.A. Vvedenie v regressionny`j analiz i planirovanie regressionny`x e`ksperimentov v e`konomike. [Introduction to regression analysis and planning of regression experiments in economics]. M.: INFRA-M, 2012. 202 p.

7. Magomedov I. A. Mezhieva A.I. Suleymanova M.A. Inženernyj vestnik Dona (Rus). 2018. №4 URL: ivdon.ru/ru/magazine/archive/n4y2018/5334.

8. Nosko V.P. Regressionny`j analiz vremenny`x ryadov. M.: ID RANXiGS, 2011. 672 p.

9. Julian J.Faraway. Extending the Linear Model with R. 2006. 324 p.

10. Frank E. Harrell, Jr. Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression and Survival Analysis. 2015. p. 3