УДК 519.254

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

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ МЕТОДА ПОВЫШЕНИЯ КАЧЕСТВА ПРОГНОЗНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ И ЕГО МОДИФИКАЦИЙ

Предложен метод повышения качества прогнозных линейных регрессионных моделей. Метод основан на исключении аномальных и не достаточно надежных данных, которые не попадают в определенную область. Произведена оценка эффективности предложенного метода.

прогнозная модель, аномальные данные, метод, точность прогноза

Введение

В настоящее время в науке и технике для решения задач прогнозирования, планирования, диагностики и оперативного управления широко используются методы статистического прогнозирования. От точности прогноза зависит правильность принятых решений. Поэтому данной проблеме посвящено множество исследований [1-3]. Однако на сегодняшний день отсутствуют эффективные алгоритмы выявления и преобразования не достаточно надежных измерений, которые были бы просты в использовании и достоверны.

Основными недостатками существующих методов являются:
− привязка к конкретным законам распределения вероятностей;
− данные методы заключаются в простом переборе исходных данных, что при большом объеме исходной статистики, отрицательно влияет на продуктивность их использования.

В связи с выявленными недостатками существующих методов целью исследований является разработка нового метода повышения точности прогнозных линейных регрессионных моделей, который может быть успешно реализован в современных компьютерных технологиях, а также оценка эффективности предложенного метода и его модификаций.

Идея предлагаемого метода

Предлагаемый метод повышения качества линейной регрессионной прогнозной модели заключается в следующем. С помощью метода наименьших квадратов по всем исходным статистическим данным, находятся коэффициенты уравнения $\hat{Y} = A \cdot x + B$. Далее определяются невязки $e_i = Y_i - \hat{Y}$ и их среднеквадратическое отклонение (СКО):

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

Находятся коэффициенты перпендикуляра, т.о. уравнение имеет вид $\hat{Y}_p = A_p \cdot x + B_p$. СКО невязок, определяется по формуле 2:

$$\sigma_{p_e} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_{p_i})^2}{n-2}} \qquad (2)$$

После этого строится прямоугольник (рис.1) со сторонами $2k\sigma_e$ и $2k\sigma_{pe}$, где k − коэффициент, соответствующий вероятности попадания в заданную область (обычно $0{,}6 \leq k < 2$).

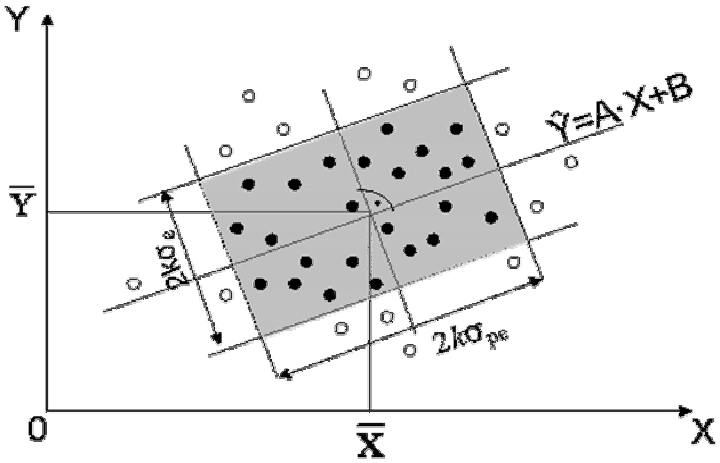

Рисунок 1 − Предложенный метод

Полученная таким образом прямоугольная область отсекает из общего числа экспериментальных данных как аномальные выбросы, так и не достаточно весомые для рассматриваемого регрессионного уравнения измерения. Вес этих отбрасываемых измерений в величине коэффициента детерминации R^2 ничтожно мал, но эти измерения существенно ухудшают качество прогнозирования.

По оставшимся после отбрасывания данным строится новое регрессионное уравнение и рассчитывается значение R^2 по формуле 3:

$$R^2 = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2}, \qquad (3)$$

где $\overline{Y} = \frac{\sum_{i=1}^{k} Y_i}{k}$ − математическое ожидание случайной величины Y_i.

Данный метод сохраняет исходную картину положения данных, а не усекает с какой-либо стороны, т.е. данные отбрасываются как по x, так и по y. В ходе исследования было выявлено, что данный метод следует использовать для выборок с исходным значением R^2, лежащим в пределах от

0.6 до 0.8. Неограниченное уменьшение прямоугольной области может привести к абсурдному научному результату, поэтому следует ограничиться отбрасыванием 30-35% исходных данных. Рост коэффициента детерминации R^2 при этом составляет не менее 20% от исходного значения R^2 (при R^2 от 0.6 до 0.75).

Первая модификация заключается в том, что вместо прямоугольника, строится коридор, состоящий из двух параллельных линий, расстояние между которыми равно $2k\sigma_e$ (рис.2).

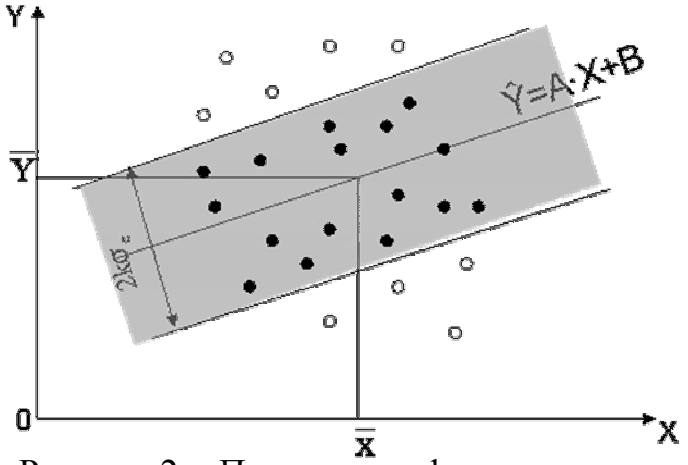

Рисунок 2 − Первая модификация метода

Критерием остановки, в данном случае является, достижение коэффициента детерминации R^2 величины 0.9.

В отличие от первой модификации метода повышения качества прогнозной модели, во второй модификации находятся 2 линии параллельные перпендикуляру (рис. 3). Расстояние между ними составляет $2k\sigma_{pe}$. В данном случае отбрасывать следует не более 15% данных.

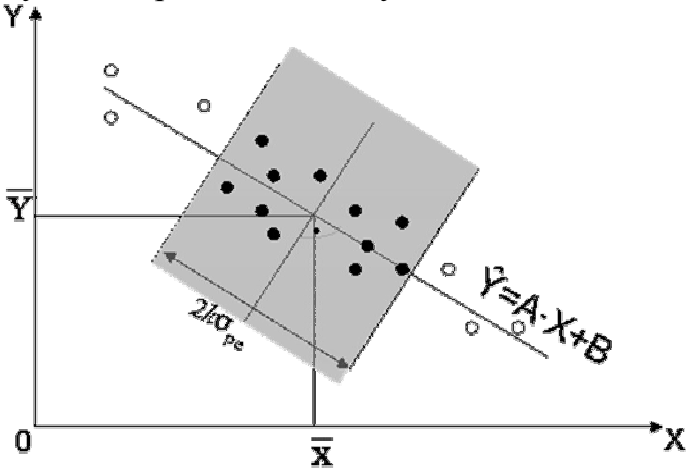

Рисунок 3 − Вторая модификация метода

Определение эффективности использование предлагаемого метода

Оценка эффективности предлагаемого метода и его модификаций, осуществлялась с помощью следующих критериев:

− коэффициент детерминации R^2;
− доверительный интервал прогнозных значений $Y_{прогн}$;
− количество элементарных операций ЭВМ, которое необходимо для реализации отбрасывания.

Для исследования эффекта полученного от применения метода были рассмотрены различные выборки. Исходное значение коэффициента детерминации R^2 для первой выборки составляет 0.709, для второй 0.89.

Первая выборка представляет собой зависимость стоимости перевозок (y) от объема продаж (x).

Таблица 1. − Данные первой выборки

x	301	328	353	372	386	389	401	408	415	444
y	52,46	72,3	54	62,98	52,95	53,7	63,7	58,99	66,8	59,7
x	446	457	458	463	484	491	503	512	517	527
y	71,66	72,81	68,44	69,33	70,77	79,38	74,39	85,58	82,03	94,44
x	535	547	596	623						
y	70,84	89,18	93,24	90,5						

Вторая выборка − зависимость количества подписчиков от количества часов на телефонный маркетинг.

Таблица 2. − Данные второй выборки

x	162	185	191	193	212	216	222	226	227
y	585	567	808	666	769	840	851	1005	848
x	232	251	257	261	263	266	269	282	286
y	1103	995	1033	975	1228	923	1118	1256	1133
x	311	321	325	331	336	345			
y	1180	1333	1289	1366	1405	1437			

После применения предложенного метода, а также его модификаций были получены результаты, представленные в таблицах 3-5. В таблицах 3 и 4, содержатся зависимости величин R^2 и доверительного интервала от количества отбрасываемых точек в процентах для первой и второй выборки соответственно. В таблице 5 находится информация о количестве элементарных операций, необходимых при использовании метода.

Таблица 3 − Значения R^2 и величина доверительного интервала для первой выборки

Количество отброшенных точек, %	1-я модиф. R^2	2-я модиф. R^2	Метод, R^2	1-я модиф. ДИ,%	2-я модиф. ДИ, %	Метод. ДИ, %
0	71	71	71	20,42064	20,42064	20,42064
10	83,684	58	78,9	12,94504	21,02419	17,61326
15	87	75,72	81,45	11,07202	17,78211	13,29402
20	89	75,72	76,22	9,681284	17,78211	13,22544
30	91,15	73,1	83,87	8,334503	17,83178	11,42163
35	92,26	71,76	87,1	7,714467	18,07439	8,831413
40	92,26	82,73	84,24	7,714467	9,209786	8,571503
50	94	68,5	78,1	5,500294	9,806398	7,858299

Таблица 4 − Значения R^2 и величина доверительного интервала для второй выборки

Количество отброшенных точек, %	1-я модиф. R^2	2-я модиф. R^2	Метод, R^2	1-я модиф. ДИ,%	2-я модиф. ДИ, %	Метод. ДИ, %
0	88,83	88,83	88,83	16,32158	16,32158	16,32158
10	94,6	87,37	93,37	12,0806	16,31454	13,23732
20	96,145	74,2	93,8145	10,27117	15,51852	12,10691
25	97,35	74,2	93,217	8,618845		10,00939
30	97,886	63,8	89,85	7,729944	15,24757	9,885149
35	97,886	63,8	87,655	7,729944		10,49687
40	98,79	66,175	90,036	5,185637	14,51236	10,2069
50	99,14		81,99	4,537033		8,877101

Таблица 5 − Количество элементарных операций

Количество отброшенных точек, %	1-я модиф.	2-я модиф.	Метод
0	219	219	219
5	639	646	843
15	613	620	809
20	600	607	792
30	574	581	775
40	561	542	724
50	535	503	690

Как видно из таблиц, предложенный метод и его первая модификация дают хорошие результаты. Значение R^2 растет, а величина доверительного интервала падает. Для первой выборки при использовании метода наибольшее значение R^2 достигается при вероятности попадания в

прямоугольник 0.65. Для того, чтобы получить область, соответствующую данной вероятности, необходимо умножить вероятность попадания в коридор для 1-й модификации на вероятность попадания для второй. Т.о. данные вероятности составляют 0.8. Это следует учитывать при сравнении первой модификации и самого метода.

В данных примерах применение второй модификации метода показало результат хуже, чем применение первой. Однако существуют ситуации, когда вторая модификация дает хорошие результаты.

Пусть дана следующая выборка (табл.7).

Таблица 7. − Исходные данные

x	2	4	6	8	10	30
y	7	11	15	19	23	45

Значение R^2, полученное по исходной выборке равно 0,977, а рассчитанное после отбрасывания 10% данных по второй модификации составляет 1.

Отсюда можно сделать вывод, что данная модификация применима для ситуаций, когда есть ярко выраженные аномальные значения переменной x.

Выводы

На основании проделанных исследований можно сделать следующие выводы:

1. Предложенный метод повышения качества линейных регрессионных моделей и его модификации дают хорошие результаты, т.к. коэффициент детерминации R^2 растет и достигает достаточного для прогнозирования значения.

2. Разработанный метод целесообразно применять при исходном значении R^2 от 0.6 до 0.8.

3. При применении предложенного метода, следует исключать не более 35% исходной статистики. Критерием остановки при использовании первой модификации является достижение коэффициента детерминации R^2 значения 0.9. Во второй модификации рекомендуется отбрасывать не более 15% измерений, т.к. в противном случае искажается картина исходных данных

4. Предложенный метод, является более точным, чем его модификации, т.к. сохраняет исходную картину представления данных.

5. Метод легко программируется и не содержат большого количества операций, поэтому может быть успешно реализован в современных компьютерных технологиях.

Список литературы

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

2. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. − М.: ФИЗМАТЛИТ, 2006. − 816 с.
3. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочн. изд. / С.А. Айвазян, И.С. Енюков, Л.Д. Мещалкин. − М.: Финансы и статистика, 1983. − 471 с.
4. Ханк Дж. Э. Бизнес-прогнозирование, 7-е изд.: пер с англ. / Ханк Дж. Э, Райтс А. Дж., Уичерн Д.У. − М.: Вильямс, 2003. − 656 с.
5. Смирнов А.В. Метод повышения качества прогнозных регрессионных моделей / А.В. Смирнов, О.В. Рычка // Наукові праці Донецького національного технічного університету. Серія: Інформатика, кібернетика та обчислювальна техніка. – 2010. - Вип. 12(165). – С.141-147.
6. 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.
7. Справочник по прикладной статистике: в 2-х т. : пер. с англ./ под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. − М.: Финансы и статистика, 1989. - Т.1.− 510 с.

Надійшла до редакції 02.10.2010 *Рецензент: канд.техн.наук, доц. Смірнов О.В.*

О.В. Ричка
Донецький національний технічний університет

Дослідження ефективності застосування метода підвищення якості прогнозних регресійних моделей та його модифікацій. Запропоновано метод підвищення якості прогнозних лінійних регресійних моделей. Метод заснований на виключенні аномальних і не достатньо надійних даних, які не попадають у певну область. Зроблено оцінку ефективності запропонованого методу.
прогнозна модель, аномальні дані, метод, точність прогнозу

O.V. Rychka
Donetsk National Technical University

Efficiency Analysis of a Method for Improving the Accuracy of Forecasting Regression Models and Its Modifications. A method for improving the accuracy of forecasting linear regression models is offered. The method is based on finding and excepting anomalous and unreliable data which miss a destination area. The estimation of the efficiency of the offered method is made.
forecasting model, anomalous data, method, forecasting accuracy