УДК 519.814

О.В. Рычка
Донецкий национальный технический университет

РАЗРАБОТКА И АНАЛИЗ МЕТОДА ПОВЫШЕНИЯ ТОЧНОСТИ ПРОГНОЗНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ И ЕГО МОДИФИКАЦИЙ

Запропоновано метод та його модифікації, що дозволяють підвищити якість прогнозних регресійних моделей. Метод базується на виключенні частини статистики, яка виходить за межі заданої області. Проведено аналіз ефективності методу та його модифікацій. Надано рекомендації до використання розробленого методу.

Предложен метод и его модификация, позволяющие повысить качество регрессионных моделей. Метод основан на исключении части статистики, которая выходит за пределы заданной области. Проведен анализ эффективности метода и его модификацией. Даны рекомендации по использованию предложенного метода.

The method and its modification, permitting to improve accuracy of regression models, is offered in the article. Method is based on the exception of part of statistics which are outside the scope of area. The effiency analisys of method and its modifications is conducted. Advices on the use of the offered method are given.

Ключевые слова: аномальные данные, прогнозная модель, повышение точности, метод

Введение. В настоящее время ни одна сфера деятельности человека не может обойтись без прогнозов. Прогнозирование позволяет подготавливать и принимать обоснованные решения, обеспечивающие эффективность управления, предупреждает возможные сбои и срывы в работе. Чем выше точность прогнозов, тем более весомым будет их вклад в развитие системы. Однако встречаются данные, которые представляют собой аномальные выбросы. Если не учитывать наличие таких данных, это может привести к значительному ухудшению качества прогноза.

На сегодняшний момент существует большое количество методов, позволяющих найти и устранить аномальные наблюдения. Для линейных регрессионных моделей к таким методам относятся − метод Прескотта-

© *О.В. Рычка, 2011*

Лунда, Эктона, Титьена-Мура-Бекмана[1;2]. Несмотря на свою простоту, эти методы имеют ряд недостатков:

– методы применяются последовательно к каждому из n измерений, что при большом объеме выборки увеличивает их трудоемкость;

– для использования данных методов требуется нормальное распределение случайных величин невязок;

– методы не применяются при объеме статистики менее 30;

– при отбрасывании данных, вследствие изменения положения регрессионного уравнения, возникает смещение прогнозного значения.

Последнего недостатка лишен метод Кука. Согласно ему при отбрасывании находятся уравнивающие наблюдения, которые стабилизируют параметры нового регрессионного уравнения. Т. о. величина смещения минимизируется, однако данный метод содержит сложные вычисления и большое число элементарных операций, а при отбрасывании более чем одного наблюдения при компьютерной обработке данных можно получить результаты, резко снижающие величину R^2. При исключении исходных статистических данных требуется большое число элементарных операций, т. к. количество возможных вариантов отбрасывания составляет $C_n^2 + C_n^3 + C_n^{n/2}$ [3].

Постановка проблемы. В связи с отсутствием в настоящее время достаточно простого и точного метода, устраняющего выявленные в ходе исследований недостатки, есть необходимость в разработке нового метода. Поэтому, целью исследований является разработка и оценка эффективности метода, позволяющего повысить точность регрессионных моделей. Основная особенность данного метода заключается в простоте реализации и возможности использования в современных компьютерных технологиях.

Сущность предлагаемого метода заключается в том, что среди исходных статистических данных находятся измерения, которые не попадают в заданную область, после чего они исключаются из выборки. Эти измерения представляют собой аномальные выбросы или не достаточно весомые наблюдения, которые существенно ухудшают качество прогноза. Отличие модификаций состоит в параметрах задаваемой области.

Вначале по всем исходным статистическим данным, находятся коэффициенты линейного уравнения $\hat{Y} = A \cdot x + B$. После этого определяется СКО невязок

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

Далее необходимо найти коэффициенты перпендикуляра. Уравнение будет иметь вид: $\hat{Y}_p = A_p \cdot x + B_p$. СКО невязок, в данном случае определяется по формуле 2

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

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

$$P_0 = 1-2[1-P(k)], \quad (3)$$

где $P(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{k} e^{-t^2/2} dt$.

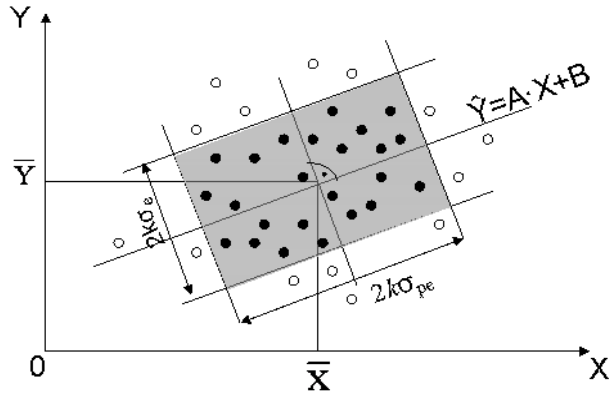


Рис. 1. Предложенный метод

Наблюдения, выходящие за рамки, построенного прямоугольника, отбрасываются, а по оставшимся данным находятся параметры нового регрессионного уравнения, после чего рассчитывается новое значение R^2 по формуле 4

$$R^2 = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2}, \qquad (4)$$

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

При использовании данного метода рекомендуется отбрасывать не более 30–35% исходной статистики. При таком отбрасывании рост коэффициента детерминации R^2, как правило, составляет не менее 20 % от исходного значения R^2 (при R^2 от 0.6 до 0.75). Предложенный метод не нарушает исходную картину положения данных, т. к. наблюдения по x и по y, отбрасываются с одинаковой вероятностью.

В некоторых практических ситуациях достаточным будет применить одну из модификаций предложенного метода. Модификации метода заключаются в том, что строится не прямоугольная область, а

определенный «коридор». В первом случае он представляет собой две параллельные линии, расстояние между которыми составляет $2k\sigma_e$ (рис.2). Здесь критерием остановки будет достижение коэффициента детерминации R^2 величины 0.9. Данная модификация применяется, когда $\sigma_y^2 > \sigma_x^2$. Во втором случае – линии параллельны построенному перпендикуляру. Расстояние между ними равно $2k\sigma_{pe}$ (рис.3). При использовании данной модификации следует отбрасывать не более 15 % наблюдений.

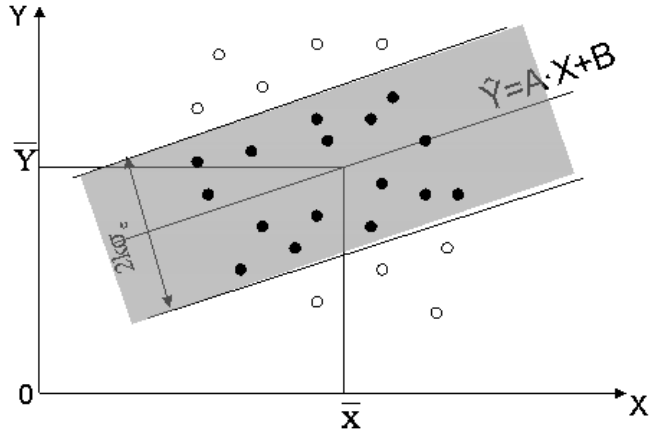


Рис. 2. Первая модификация метода

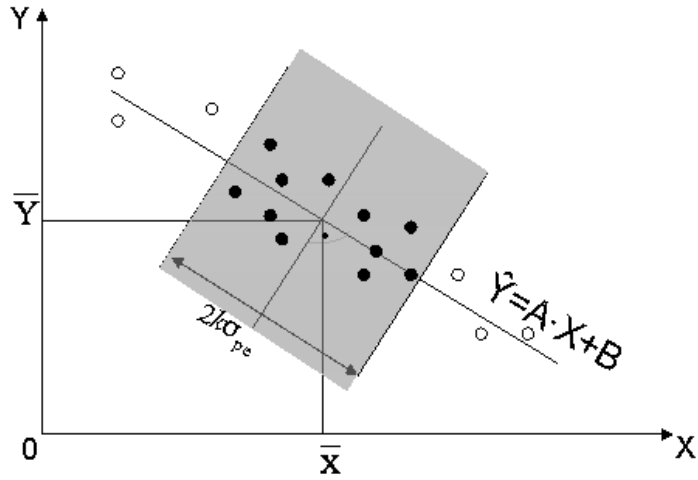


Рис. 3. Вторая модификация метода

Следует отметить, что вероятность попадания в прямоугольник равна произведению вероятности попадания в «коридор» для первой модификации на вероятность попадания для второй, при этом вероятности равны друг другу.

Анализ эффективности предложенного метода и его модификаций. Основными критериями оценки эффективности нового метода являются:

- коэффициент детерминации R^2;
- величина доверительного интервала;
- величина смещения результата прогноза;
- количество элементарных операций ЭВМ, необходимых для реализации метода.

С помощью первых двух критериев оценивается качество линейной регрессионной модели. Третий критерий является показателем трудоемкости.

При определении доверительный интервалом прогноза квантили Стьюдента $t(P_{дов}; k=n-2)$ не использовались. Это связано с тем, что закон распределения рассматриваемой статистики невязок отличен от нормального. Исключение части невязок еще более усугубляет эту

проблему. Поэтому использовались доверительные интервалы свободные от закона распределения случайных величин невязок [4].

Пусть e_1, e_2, e_3, …, e_i порядковые статистики и $e_1<e_2<e_3 \ldots e_{i-1}< e_i$. При этом доверительный интервал (e_1, e_i) симметричен относительно своего медианного значения e_k и определяется доверительной вероятностью:

$$P_{дов} = I_{1/2}(1,i) - I_{1/2}(i,1),$$

$$I_x(a,b) = 1 - I_{1-x}(b,a),\ (0 \le x \le 1),$$

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1} \cdot (1-t)^{b-1} dt,$$

$$B(z,\omega) = \frac{\Gamma(z) \cdot \Gamma(\omega)}{\Gamma(z+\omega)}\ [5],$$

$\Gamma(n+1) = 1 \cdot 2 \cdot 3 \ldots (n-1) \cdot n = n!$ для целых значений аргумента,

где: $I_x(a,b)$ − неполная Бета-функция; $B(z, \omega)$ – Бета-функция; $\Gamma(n+1)$ – Гамма-функция.

Для вычисления доверительного интервала подбирались пары 1, i; 2, i-1; 3, i-2 и т. д. симметрично расположенные относительно медианного значения e_k, обеспечивающие заданное значение $P_{дов}$. Найденная пара представляет собой размах доверительного интервала. Естественно, что из-за дискретного характера порядковой статистики точное значение величины доверительного интервала таким способом найти не удается, однако с этим приходится мириться.

Анализ метода и его модификаций по предложенным критериям, проводился на различных наборах наблюдений. Для наглядного отражения результатов использования предложенного метода и его модификаций в данной статье, рассматриваются два варианта выборок с различными исходными значениями коэффициента детерминации R^2.

Первая выборка представляет зависимость стоимости перевозок y от объема продаж x (табл.1). Исходное значение коэффициента детерминации R^2 равно 0.71.

Таблица 1

Исходные данные первой выборки

x	301	328	353	372	386	389	401	408
y	52,46	72,3	54	62,98	52,95	53,7	63,7	58,99
x	415	444	446	457	458	463	484	491

y	66,8	59,7	71,66	72,81	68,44	69,33	70,77	79,38
x	503	512	517	527	535	547	596	623
y	74,39	85,58	82,03	94,44	70,84	89,18	93,24	90,5

Вторая выборка отражает зависимость процентной доли учеников, успешно прошедших тестирование от средней посещаемости уроков. Исходная величина R^2 составляет 0.6

Таблица 2

Исходные данные третьей выборки

x	88,67	91,93	92,08	92,21	92,22	92,63	92,8	92,87
y	26	38	60	42	49	58	53	40
x	93,05	93,06	93,22	93,39	93,67	93,71	93,8	93,93
y	43	48	22	59	45	58	70	60
x	93,95	93,96	94,03	94,04	94,16	94,2	94,3	94,36
y	59	59	67	44	71	49	46	78
x	94,4	94,4	94,42	94,5	94,54	94,66	94,7	94,72
y	65	69	68	60	66	69	66	62
x	94,76	94,83	94,93	94,97	95,56	95,66	95,7	95,76
y	67	55	38	58	84	88	79	84
x	95,76	95,77	95,93	96,07	96,13	96,33	96,8	
y	74	72	78	82	92	83	93	

Результаты применения предложенного метода, а также его модификаций представлены в таблицах 3–6. 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. Количество отбрасываемых точек соответствует вероятности непопадания в заданную область. Величины смещения прогнозного значения $Y_{прогн}$, рассчитанные в процентах, полученные в результате использовании предложенного метода, представлены в таблице 5. В таблице 6 приведена информация о числе элементарных операций, необходимых при использовании метода.

Таблица 3

Значения R^2 и величина доверительного интервала для первой выборки

Процент отброшен-ных 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	0.71	0.71	0.71	20,42	20,42064	20,42064
10	0.837	0.58	0.799	12,945	21,02419	17,61326
15	0.87	0.7572	0.8145	11,07	17,78211	13,29402
20	0.89	0.7572	0.7622	9,68	17,78211	13,22544
30	0.9115	0.731	0.8387	8,33	17,83178	11,42163
35	0.9226	0.7176	0.871	7,71	18,07439	8,831413
40	0.9226	0.8273	0.8424	7,71	9,209786	8,571503
50	0.94	0.685	0.781	5,5	9,806398	7,858299

Т. о. из таблицы видно, что коэффициент детерминации R^2, при использовании предложенного метода после отбрасывания 35 % исходных данных вырос с 0.71 до 0.87, величина доверительного интервала уменьшилась в 2.3 раза. Использование первой модификации также дало хорошие результаты − рост R^2 до рекомендуемого значения и уменьшение доверительного интервала в 2.5 раза. В данном примере вторая модификация также дает результаты, однако не такие значительные, как метод и первая модификация.

Таблица 4

Значения R^2 и величина доверительного интервала для второй выборки

Процент отброшенных точек	1-я модиф. R^2	Метод, R^2	1-я модиф. ДИ,%	Метод. ДИ, %
0	0,602	0,6	39,02	39,02
10	0,729	0,68	24,83	28,1
20	0,786	0,67	21,5	25,03
25	0,79	0,685	19,73	25,01
30	0,85	0,74	18,24	19,71
35	0,897	0,687	14,33	19,61
40	0,917	0,594	11,43	19,32
50	0,919	0,716	8,984	14,77

В данном случае при использовании нового метода коэффициент детерминации вырос на 23 % от исходного значения равного 0,6 и составил 0,74, величина доверительного интервала при этом уменьшилась

в 2,1 раза. При использовании первой модификации, здесь также как и в предыдущем примере достигается рекомендуемое значение R^2 равное 0.9. Однако, в данном случае, использование второй модификации значительно снизило величину коэффициента детерминации (в таблице это не отражено). Это связано со спецификой исходных данных.

Таблица 5

Значения величины смещения Δ

Процент отброшенных точек	1-я выборка Δ,%	2-я выборка Δ, %
10	0,906846	1,562204
15	2,263314	0,997816
20	2,355117	0,21496
30	1,225094	0,50746
35	0,536757	0,125088
40	0,173853	0,048557
50	0,167467	2,575949

Таблица 6

Количество элементарных операций

Количество отброшенных точек, %	1-я выборка			2-я выборка		
	1-я модиф.	2-я модиф.	Метод	1-я модиф.	2-я модиф.	Метод
0	219	219	219	426	426	426
5	639	646	843	1247	1254	1654
15	613	620	809	1208	1215	1620
20	600	607	792	1195	1150	1603
30	574	581	775	1104	1098	1535
40	561	542	724	1039	1020	1433
50	535	503	690	1000	968	1331

Применение предложенного метода дало следующие результаты:

− коэффициент детерминации R^2 растет и достигает рекомендуемого значения при вероятности попадания в заданную область равную 65 %;

− величина доверительного интервала уменьшается до 2.5 раза;

− величина смещения прогнозного значения $Y_{прогн}$ не превышает 2.5 %;

– число элементарных операций увеличилось в 3.5 раза и составило 750 и 1535 при исходном объеме выборки 24 и 47 значений соответственно, но оно является меньшим, чем при применении известных методов.

Т. о. можно сказать, что применение предложенного метода и его первой модификации дало хорошие результаты.

В рассматриваемых примерах использование второй модификации метода показало результат хуже, чем при применении метода или его первой модификации. 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. В результате проведенных исследований можно сделать следующие выводы:

1. Применение разработанного метода повышения качества прогнозных регрессионных моделей приводит к росту коэффициента детерминации R^2.
2. Величины доверительного интервала снижается.
3. Смещение прогнозной величины $Y_{прогн}$ является незначительным.
4. При использовании предложенного метода рекомендуется исключать не более 30–35% наблюдений. Рост коэффициента детерминации R^2 при этом, как правило, составляет не менее 20 % от исходного значения R^2 (при R^2 от 0.6 до 0.75).
5. Для первой модификации критерием остановки является увеличение R^2 до 0.9.
6. Применяя вторую модификацию, следует отбрасывать не более 15 % данных для того, чтобы не произошло искажения исходного вида представления независимой переменной x.
7. Исходное значение коэффициента детерминации R^2 не должно выходить за пределы интервала от 0.6 до 0.8.

8. Предложенный метод и его модификации легко реализуются в современных компьютерных технологиях, т. к. не содержат большого числа элементарных операций и не требуют постоянного контроля со стороны пользователя.

Библиографические ссылки

1. **Кобзарь А.И.** Прикладная математическая статистика. Для инженеров и научных работников. / А.И. Кобзарь − М., 2006. − 816 с.
2. **Дрейпер Н.Р., Смит Г.** Прикладной регрессионный анализ. 3-е изд.: Пер. с англ. / Н.Р. Дрейпер, Г. Смит − М., 2007. − 912 с.
3. **Смирнов А.В., Рычка О.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». Вип. 12(165), 2010. – С.141-147.
4. **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., 1989. − 510 с.
5. **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., 1979. − 830 с
6. **Айвазян С.А.** Прикладная статистика: Основы моделирования и первичная обработка данных. Справочн. изд./ С.А. Айвазян, И.С. Енюков, Л.Д. Мещалкин. − М., 1983. − 471 с.

Надійшла до редколегії 10.11.10