УДК 519.254

А.В. Смирнов, О.В. Рычка
Донецкий национальный технический университет
smirnov_dntu@ukr.net

Новый метод улучшения качества прогнозных регрессионных моделей

Предложен оригинальный метод повышения качества регрессионных прогнозных моделей, основанный на нелинейном преобразовании невязок. При этом аномальные и ненадежные исходные измерения не удаляются, а изменяются их координаты.
***Ключевые слова**: преобразование невязок; изменение координат; улучшение качества; регрессионная прогнозная модель.*

Введение

Методы статистического прогнозирования в настоящее время широко используются в науке и технике. Несмотря на то, что регрессионные прогнозные модели известны давно, их усовершенствование, модернизация продолжается и в настоящее время [1-3 и др.]. В [3] предложен новый оригинальный метод повышения качества прогнозных регрессионных моделей, основанный на исключении из регрессионного анализа как "аномальных ошибок" так и ненадежных исходных данных. Отбрасывание таких данных в разумных пределах способствует повышению качества исходных прогнозных моделей по ряду критериев: увеличивается коэффициент детерминации R^2, уменьшается величина доверительного интервала прогноза при заданной величине доверительной вероятности $P_{дов}$. Кроме того, как показали исследования в [3], существенно уменьшается количество элементарных операций, необходимых для реализации метода по сравнению с известным методом-прототипом Кука. В рамках дальнейших исследований у авторов [3] возник вопрос, что лучше: исключать из регрессионного анализа аномальные и ненадежные исходные данные или их сохранить, но изменить их координаты? Так возник новый метод, который описан в данной статье.

Цель исследований

Целью настоящих исследований является разработка нового оригинального метода повышения качества регрессионных прогнозных моделей и сравнение его на многокритериальной основе с известным методом Кука и ранее предложенным методом [3]. Для реализации поставленной цели необходимо решить следующие задачи:

- разработать механизм трансформации координат аномальных и ненадежных исходных данных;
- рассмотреть теоретические предпосылки улучшения качества регрессионных моделей при использовании предложенного метода.

Используемая модель статистических данных

В данных исследованиях используется модель Тьюки:

$$\omega(y) = (1-\beta)\cdot\varphi(y;a;\sigma_0^2) + \beta\cdot\varphi(y;a;\sigma_1^2), \quad (1)$$

где: $\varphi(y;a;\sigma^2)$ – нормальный закон распределения со средним значением a и дисперсией σ^2; β – доля "засоряющих" аномальных наблюдений.

В (1) обычно $\sigma_0^2 < \sigma_1^2$. В данной работе, как и в [3], модель (1) может быть существенно расширена, путем отхода от нормального закона $\varphi(y;a;\sigma^2)$. В качестве $\varphi(y;a;\sigma^2)$ может быть использована любая плотность распределения вероятностей случайных величин y_i, обладающая свойствами симметрии относительно a и убывающая при росте величины $|y_i - a|$. Это может быть любой из известных законов: двустороннее экспоненциальное распределение (распределение Лапласа); Симпсона и др.

Следует отметить особенности подхода авторов к решению поставленной задачи. Они заключаются в том, что "вредными" для исследователя, считаются не только случайные величины y_i, описываемые вторым слагаемым (1), но и другие y_i, которые расположены на умеренном расстоянии от a, но не могут достойно поддерживать используемый регрессор (они относятся к первому слагаемому).

Сущность предлагаемого метода и его модификации

Сущность предлагаемого метода повышения качества линейной регрессионной прогнозной модели заключается в следующем. На первом этапе исследователь, используя все исходные статистические данные, находит вид линейного уравнения с использованием традиционного метода наименьших квадратов $\hat{Y} = A \cdot x + B$. Далее определяются невязки $e_i = Y_i - \hat{Y}$ и их СКО:

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

Находим СКО невязок e'_i относительно уравнения $\hat{Y}'_i$:

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

В формуле (3) значения $\hat{Y}'_i$ находятся исходя из следующего уравнения:

$$\hat{Y}' = -\frac{1}{A}x + \bar{y} + \frac{\bar{x}}{A}, \qquad (4)$$

где: $\bar{x}_i = \frac{\sum_{i=1}^{n} x_i}{n}$ – математическое ожидание случайных величин x_i;

$\bar{y}_i = \frac{\sum_{i=1}^{n} y_i}{n}$ – математическое ожидание случайных величин y_i.

Используя (2) и (3), строим прямоугольник со сторонами $2k\cdot\sigma_e$ и $2l\sigma'_e$, где k и l числа (обычно $0{,}6 \le k < 3$ и $0{,}6 \le l < 3$). Для реализации метода необходимо выполнить условие пропорциональности k и l. При несоблюдении этого условия нарушаются статистические связи исходных данных. Построенный прямоугольник изображен на рис.1. При $k \ge 3$ и $l \ge 3$ этот прямоугольник описывает геометрическое место точек, характеризующих исходные данные регрессионной модели. Если исходные данные подчиняются двумерному нормальному распределению, то их проекция на плоскость XOY представляет собой эллипс. Построенный прямоугольник описывает в этом случае внутренний эллипс и его параметры совпадают с параметрами эллипса.

Рассматриваемый прямоугольник отсекает из общего числа экспериментальных данных как аномальные выбросы, так и не достаточно весомые для рассматриваемого регрессионного уравнения измерения. Вес этих трансформируемых измерений в величине коэффициента детерминации R^2 достаточно мал, но эти измерения существенно ухудшают качество прогнозирования. Аналогично с методом, который описан в [3], существуют два варианта упрощения предложенного метода в зависимости от соотношения величин σ_e и σ'_e (рис. 2).

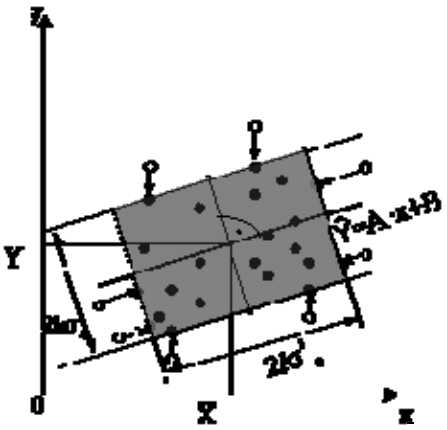
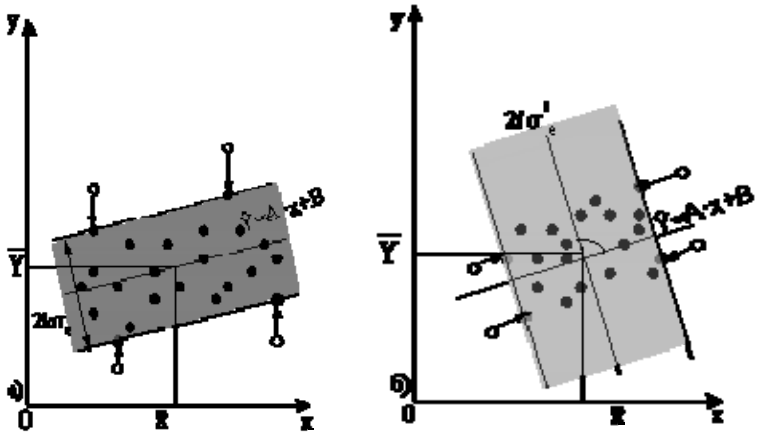


Рисунок 1 − Реализация предложенного метода



а) б)

Рисунок 2− Упрощенные модификации метода:
а) $\sigma'^2_e < \sigma^2_e$ и б) $\sigma'^2_e > \sigma^2_e$

Кратко рассмотрим статистические особенности предложенного метода по сравнению с ранее опубликованным в [3]. Предложенный в данной работе метод является нелинейным по отношению к исходным данным по которым построено регрессионное уравнение

$\hat{Y} = A \cdot x + B$. Рассмотрим его особенности (рис. 3).

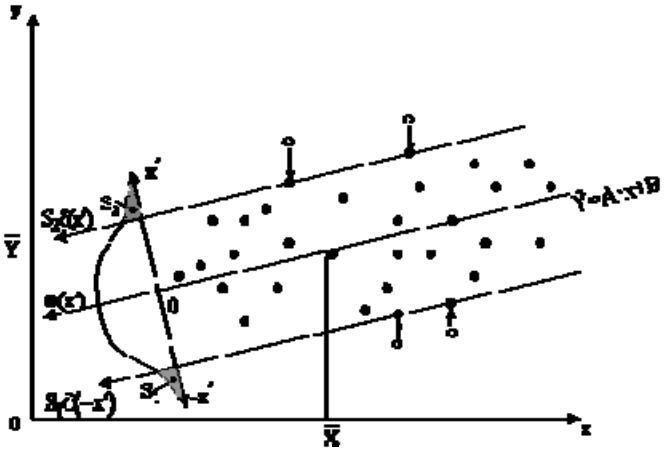


Рисунок 3 − Статистические особенности метода

Исходные невязки e_i имеют нормальное распределение. После осуществления операции переноса аномальных и ненадежных измерений на уровни $\hat{Y} \pm u$, где u − новые положения исходного уравнения $\hat{Y}$ (которые симметричны относительно исходного), происходит трансформация исходной нормальной плотности вероятностей ω(x') в новую плотность вероятностей невязок e_i:

$$\omega'(x') = \begin{cases} \omega(x'), \text{если } \hat{Y} - u < x' < \hat{Y} + u, \\ S_1 \cdot \delta(-x'), \text{если } x' = \hat{Y} + u, \\ S_2 \cdot \delta(x'), \text{если } x' = \hat{Y} - u \end{cases} \quad (5)$$

где: δ () − дельта-функция Дирака [4].

В (5) $S_1 = S_2 = \int_{u+\hat{Y}}^{\infty} \omega(x')dx'$, поскольку ± u симметричны относительно исходного уравнения $\hat{Y}$. Преобразование исходного закона ω(x') в ω'(x') может быть осуществлено, если невязки e_i исходного регрессионного уравнения $\hat{Y}$ пропустить через двухсторонний безинерционный ограничитель с линейным участком [4]:

$$y = f(x) = \begin{cases} -u, & -\infty < x < -u, \\ x, & -u < x < u, \\ u, & u < x < \infty. \end{cases} \quad (6)$$

При осуществлении нелинейного преобразования (6) исходные данные, которые находятся между уровнями ±u, не претерпевают изменений, а другие − привязываются к уровню +u или –u. Такое нелинейное преобразование изменяет спектральные (корреляционные) характеристики исходных данных, а данные, которые не претерпели изменений, характеризуются прежними вероятностными характеристиками.

Совершенно другими статистическими характеристиками описывается метод улучшения качества прогнозных регрессионных моделей [3]. Здесь осуществляется линейная операция отбрасывания аномальных и ненадежных исходных данных (рис.4).

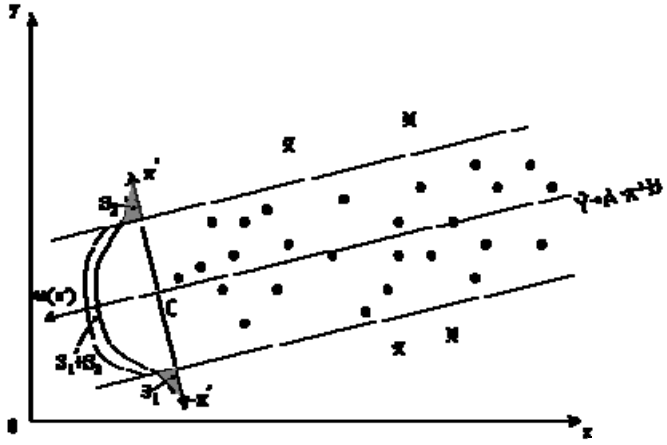


Рисунок 4 − Статистические характеристики метода [3]

Как видно из рисунка 4, при реализации отбрасывания аномальных и ненадежных данных происходит увеличение вероятности качественных исходных данных. При этом спектральные (корреляционные) характеристики в результате такого преобразования меняются в меньшей степени.

Сопоставление методов

Сравним между собой метод-прототип Кука [1], известный метод [3] и предложенный оригинальный метод улучшения качества прогнозных регрессионных моделей при одинаковых условиях. Для их взаимного тестирования будем снова использовать известный пример [1] о взаимосвязи возраста ребенка, в котором он произнес свое первое слово (X), и результаты адаптивного теста Гесселя (Y). Рисунок с исходной статистикой и исходным регрессионным уравнением не приводим и отсылаем читателя в [3], где он приведен. При использовании 100% исходных данных линейное регрессионное уравнение имеет вид: $\hat{Y}$=109,87-1,127·X (R^2= 0,41). При $X_{прогн}$=14,381, $Y_{прогн}$=93,67. Доверительный прогнозный интервал составляет 24,5% от величины $Y_{прогн}$=93,67 при $P_{дов}$=1. В качестве частных критериев эффективности будем использовать, как и в [3]:

− коэффициент детерминации R^2;
− модуль величины смещения результата прогноза;
− количество элементарных операций ЭВМ для реализации преобразований;
− величину доверительного интервала прогноза.

Подробное описание этих частных критериев осуществлено в [3]. Результаты сопоставительного анализа приведены в табл.1

Таблица 1. Результаты сопоставительного анализа

Используемый метод	Параметры метода	Номера исключенных или преобразованных наблюдений	Критерий качества метода			
			R^2	Δ^2, %	Величина доверительного интервала, %	Количество элементарных операций
Метод Кука	D_k=0,015	**2, 3, 19**	0,58	0,32	14,2	~ $0{,}4\cdot10^6$
	D_k=0,002	**2, 3, 11, 14, 19**	0,64	0,18	13,18	~ $7\cdot10^6$
	D_k=0,097	**2, 3, 14, 20, 19, 11, 5, 4**	0,77	0,77	13	~ $1\cdot10^8$
Известный метод [3] (вариант рис. 2б)	k=1,27 (D_k=0,17)	**19, 3, 13**	0,7	0,12	12,7	~ $0{,}5\cdot10^3$
	k=1 (D_k=0,74)	**3, 13, 14, 20, 19**	0,83	1,96	10,2	~ $0{,}5\cdot10^3$
	k=0,8 (D_k=0,22)	**2, 3, 13, 14, 20, 19, 11, 5**	0,86	1,35	9,35	~ $0{,}5\cdot10^3$
Предложенный метод (вариант рис. 2а)	k=1,27 (D_k=0,17)	**19, 3, 13**	0,53	0,68	15,6	~ $0{,}5\cdot10^3$
	k=1 (D_k=0,74)	**3, 13, 14, 20, 19**	0,6	0,4	12,4	~ $0{,}5\cdot10^3$
	k=0,8 (D_k=0,22)	**2, 3, 13, 14, 20, 19, 11, 5**	0,67	0,13	10	~ $0{,}5\cdot10^3$

Сопоставление результатов (табл.1) методов повышения качества прогнозных регрессионных моделей для рассмотренного примера дало следующие результаты:

− по критерию R^2 наиболее предпочтителен метод [3], ему на 10% уступает метод Кука, а предложенный также на 10% проигрывает методу Кука;
− по величине модуля смещения результата прогноза лучшим оказался предложенный метод (особенно при большом количестве точек с преобразованными координатами). Он превосходит по этому показателю метод Кука в 6 раз и метод [3] в 10,3 раза. Это объясняется тем, что исходные данные с преобразованными координатами создают коридор ±u и величина Δ^2 всегда меньше его величины;

− по величине доверительного интервала прогноза предложенный метод уступает методу [3] на 6,5%, но опережает метод Кука на 30%. Следует обратить внимание читателей, что в этой работе величины доверительного интервала прогноза пересчитаны для доверительной вероятности $P_{дов}$=1. Этим и объясняются расхождения по этому параметру с [3].

− по количеству элементарных операций на ЭВМ, которые необходимы для реализации сравниваемых методов, метод Кука на три порядка уступает как методу [3] так и предлагаемому.

По мнению авторов, незначительный проигрыш предложенного оригинального метода улучшения качества прогнозных регрессионных моделей по отдельным критериям объясняется тем, что данный метод существенно нелинейный и он приводит к изменению вероятностных, спектральных (корреляционных) характеристик исходных данных.

Предложенный метод, как и метод [3] можно использовать для улучшения качества нелинейных прогнозных регрессионных моделей. Для этого требуется их свести к линейным, сделать необходимые преобразования и вновь преобразовать в нелинейные.

Выводы

На основании проделанных исследований можно сделать следующие выводы:

1. Предложенный оригинальный метод улучшения качества прогнозирования регрессионных моделей является существенно нелинейным и он сводится к преобразованию исходных невязок с помощью нелинейного безинерционного и симметричного двухстороннего ограничителя с симметричным линейным участком.

2. Предложенный в [3] способ улучшения качества прогнозных регрессионных моделей, основанный на отбрасывании аномальных и ненадежных исходных данных, является линейным.

3. Предложенный в данной статье и ранее описанный в [3] методы улучшения качества прогнозных регрессионных моделей по большинству частных критериев дают сходные результаты. Некоторые уменьшения эффекта от применения предложенного в данной работе метода связано с его нелинейностью, которая влияет на исходные вероятности, спектральные (корреляционные) характеристики исходных данных.

4. Основным преимуществом оригинальных методов повышения качества прогнозных регрессионных моделей перед методом-прототипом Кука является существенно меньшее количество элементарных операций при реализации на ЭВМ. Это объясняется уходом от простого перебора, используемого в методе Кука, и переходом к целенаправленному поиску наилучших параметров предложенных методов. При этом выигрыш по быстродействию может оказаться в 10^2…10^5 раз.

Литература

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. 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.

4. Тихонов В.Н. Статистическая радиотехника: изд. 2-е; перераб. и доп. / В.Н. Тихонов. – М.: Радио и связь, 1982. – 624 с.

Надійшла до редакції 17.01.2011