УДК 681.3.07

Метод повышения качества прогнозных регрессионных моделей

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

Abstract
Smirnov A.V., Rychka O.V. Article "The method of improving accuracy of forecasting regression models". *The original method of improving accuracy of regression forecasting models and its modification is offered. The method is based on elimination of part of the anomalous and unimportant values which are out of rectangle. The rectangle is built on a line of the equation* $\hat{Y} = A \cdot x + B$ *with sides 2k·SD of residuals* e_i *and 2l|A|·SD* X_i *. The values k and l are determined after reaching* R^2 *a set quantity. Comparison new method with the known Cook's method at the use of multi-criteria approach is conducted.*

Введение

Методы статистического прогнозирования в настоящее время широко используются в науке и технике для решения задач прогнозирования, планирования, диагностики и оперативного управления. Эти методы делятся на два больших класса: регрессионные прогнозные модели и модели на основе экстраполяции трендов одномерных временных рядов. От точности предсказания этих моделей зависит корректность, действенность, надежность оперативного управления экономическими, социальными и техническими системами. Несмотря на то, что эти модели широко известны с XIX века их модернизация, уточнение продолжается до сих пор. Об этом свидетельствуют многочисленные публикации в зарубежных и отечественных специализированных научных изданиях [1,2].

К неудовлетворительно решенным на сегодняшний день задачам прикладного регрессионного анализа, по мнению авторов, относятся:

− отсутствие достаточно эффективных алгоритмов обнаружения "аномальных ошибок" или "выбросов" исходной статистики;
− отсутствие достаточно эффективных алгоритмов устранения части исходной статистики, которая посредственно поддерживает стремление регрессора объяснить поведение зависимой переменной;
− привязка известных методов выявления выбросов исходной статистики к конкретным законам распределения вероятностей (которые исследователям априорно не известны);
− остается открытым вопрос о том, что должен делать исследователь с выявленными аномальными измерениями и ненадежными данными (исключить их или корректировать их особым образом);
- все известные методы обнаружения аномальных измерений плохо формализованы и сводятся к простому перебору вычисленных статистик, что затрудняет их использование в современных компьютерных технологиях.

Разработка новых эффективных алгоритмов выявления и преобразования аномальных и не достаточно надежных измерений будет способствовать существенному повышению качества прогнозных регрессионных моделей, успешной их реализации в пакетах прикладных программ для ЭВМ и активному внедрению их в современные автоматизированные информационные системы.

Современные концепции повышения качества

Анализ источников по теме исследований показал, что в настоящее время существует две основных концепции в борьбе за повышение качества прогнозных регрессионных моделей:

− выявление с последующим исключением из анализа единственной аномальной невязки (выявление с последующим устранением нескольких аномальных невязок на основе поэтапного устранения по одному выбросу) [1,2];
− нахождение с последующим исключением большего количества невязок, которые не всегда являются аномальными и их совместное отбрасывание приводит к минимальным изменениям параметров исходного регрессионного уравнения [1].

Первую концепцию реализуют методы Эктона (Acton F.S., 1959), Титьена-Мура-Бекмана (Tietjen G.L., Moore R.H., Beckman R.J., 1973), а так же Прескотта-Лунда (Prescott P., Lund R.E., 1975). Эти методы предназначены для выявления с последующим удалением единственного аномального измерения при нормальном законе распределения случайных величин невязок и их количестве $n \geq 30$. Рассмотренные методы можно применять многократно последовательно к одной и той же экспериментальной статистике. Преимущество методов заключается в простоте алгоритмов, а недостатков у них множество:

− требование нормального распределения случайных величин невязок (на практике это требование выполняется с большой натяжкой, т.к. наличие аномальных выбросов приводит к утяжелению "хвостов" плотности вероятностей);

− ограничения на минимальный объем статистики n (ограничивает применимость перечисленных методов при $n < 30$);

− последовательное применение метода к каждому из n измерений (полный перебор) резко увеличивает общее время анализа. Это время возрастает в N раз при выборе из статистики N-кратных выбросов;

− при отбрасывании одной или нескольких невязок происходит изменение параметров нового регрессионного уравнения по сравнению с исходным. При этом возникает смещение прогнозного значения $Y_{прогн}$. Это ограничивает реальное повышение качества прогнозной модели.

Малая эффективность рассмотренных методов, реализующих первую концепцию повышения качества, привела к разработке принципиально новых. Здесь следует отметить первый метод принципиально новой концепции − метод Кука (Cook R.D., 1977). Несколько позднее появились методы Белсли-Ку-Уэлша (Belsley D.A., Kuh E., Welsch R.E., 1980) и Аткинсона (Askinson A.C., 1985). Однако последние перечисленные методы являются незначительной модификацией исходного метода Кука [1]. Сущность метода Кука заключается в нахождении при отбрасывании уравнивающих измерений, которые стабилизируют параметры, нового регрессионного уравнения по отношению к исходному. Это приводит к тому, что даже при отбрасывании множества невязок, эффект смещения оценок $Y_{прогн}$ будет минимальным. Здесь имеется принципиальное отличие с методами первой концепции. В случае применения метода Кука или его модификаций отбрасываются не аномальные невязки, а любые невязки, которые удовлетворяют минимуму критерия Кука. К преимуществу метода Кука следует отнести:

− высокую эффективность метода при ручном анализе из-за существенного уменьшения величины смещения $Y_{прогн}$.

К недостаткам этого метода относятся:

− сложность вычислений (особенно при вычислении статистики Кука высокого порядка);

− искусственность и непрозрачность при выборе отбрасываемых невязок;

− малая пригодность метода Кука для реализации оперативного управления системами из-за большого времени анализа.

Все перечисленные выше недостатки традиционных методов улучшения качества прогнозных регрессионных моделей дали толчок к разработке нового оригинального метода, который может быть успешно реализован в современных компьютерных технологиях.

Цель исследований

Целью исследований является разработка нового метода улучшения качества прогнозных линейных регрессионных моделей, который может успешно применяться в современных компьютерных технологиях.

Особое место в данных исследованиях занимают сравнения предложенного оригинального метода с методом − прототипом Кука по традиционным для регрессионного анализа критериям и количеству элементарных операций ЭВМ для их реализации.

Используемая модель статистических данных

В качестве моделей "засорения" исходной статистики для реализации линейных регрессионных прогнозных моделей обычно используется модель Тьюки или Шурыгина [3]. В данных исследованиях используется модель Тьюки:

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

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 модель (1) может быть существенно расширена, путем отхода от нормального закона $\varphi(y;a;\sigma^2)$. В качестве $\varphi(y)$ 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 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:

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

Величина СКО x_i оценивает их разброс относительно $\overline{x}_i$:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x}_i)^2}{n-1}}\,, \qquad (3)$$

где: $\overline{x}_i = \frac{\sum_{i=1}^{n} x_i}{n}$ − математическое ожидание случайных величин x_i.

Находим проекцию величины σ_x на линию, которая описана уравнением $\hat{Y}$:

$$\sigma_{пр}=|A|\cdot \sigma_x, \qquad (4)$$

где: $\sigma_{пр}$ – проекция величины на уравнение $\hat{Y} = A \cdot x + B$.

Используя (2) и (4), строим прямоугольник со сторонами $2k\cdot\sigma_e$ и $2l|A|\cdot\sigma_x$, где k и l числа (обычно $0{,}6\le k< 3$ и $0{,}6\le l< 3$). Построенный прямоугольник изображен на рис.1.

Прямоугольник 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. Применение прямоугольника с симметричными относительно линии регрессионного уравнения сторонами автоматически снижает величину смещения Δ при оценке нового $Y_{прогн}$ (поскольку величина Δ не может превышать геометрические размеры этого прямоугольника).

При реализации данного метода возникает вопрос: до каких пор следует уменьшать размеры прямоугольника, внутри которого исходные статистические данные считаются надежными? Неограниченное уменьшение сторон этого прямоугольника может привести к абсурдному научному результату. Эта задача в рассматриваемом методе решается путем определения значений коэффициента детерминации R^2 при поэтапном уменьшении величин k и l, т.е. задается значение, например R^2=0,9, и находятся величины k и l. При отбрасывании части статистики σ_e = varia и σ_x=varia.

Предложенный метод можно существенно упростить, учитывая, что при решении большинства практических задач величины дисперсий σ_x^2 и σ_y^2 могут существенно отличаться. В случае, когда $\sigma_x^2 > \sigma_y^2$, вместо прямоугольника можно использовать две параллельные линии, расстояние между которыми равно $2l|A|\cdot\sigma_x$ (рис. 2а). В противном случае, когда $\sigma_x^2 < \sigma_y^2$, модифицированный метод реализуется двумя параллельными линиями с расстоянием между ними $2k\sigma_e$ (рис. 2б). Естественно, что и первая и вторая модификации несколько снижают эффективность предложенного метода, хотя их реализация значительно проще.

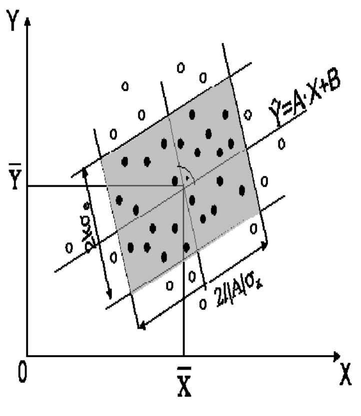

Рисунок 1 – Реализация предложенного метода

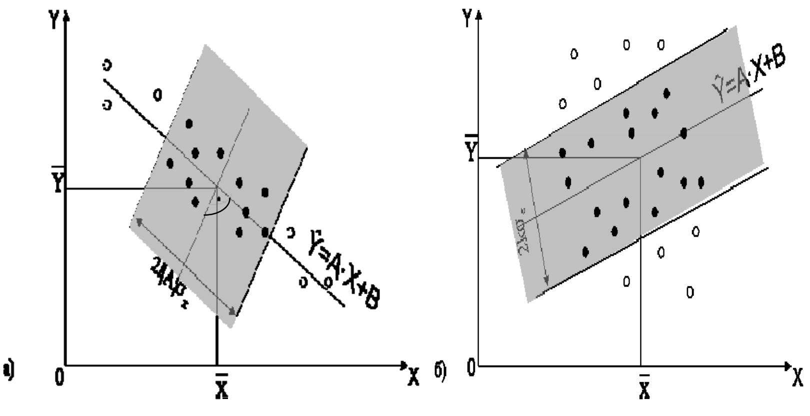

Рисунок 2 − Упрощенные модификации метода а) $\sigma_x^2 > \sigma_y^2$ и б) $\sigma_x^2 < \sigma_y^2$

Сопоставление методов

Сравним между собой метод-прототип Кука [1] и предложенный нами оригинальный метод (вариант рис. 2б). Для взаимного тестирования будем использовать известный пример [1] о взаимосвязи возраста ребенка, в котором он произнес свое первое слово (X), и результаты адаптивного теста Геселля (Y).

На рис. 3 представлены данные этого примера с нумерацией экспериментальных измерений для наглядности. При использовании 100% исходных данных линейное регрессионное уравнение имеет вид: $\hat{Y}=109{,}87 - 1{,}127 \cdot X$ (R^2= 0,41). При $X_{прогн}$=14,381, $Y_{прогн}$=93,67. Доверительный прогнозный интервал составляет 4,8% от величины $Y_{прогн}$=93,67 при $P_{дов}$=0,9.

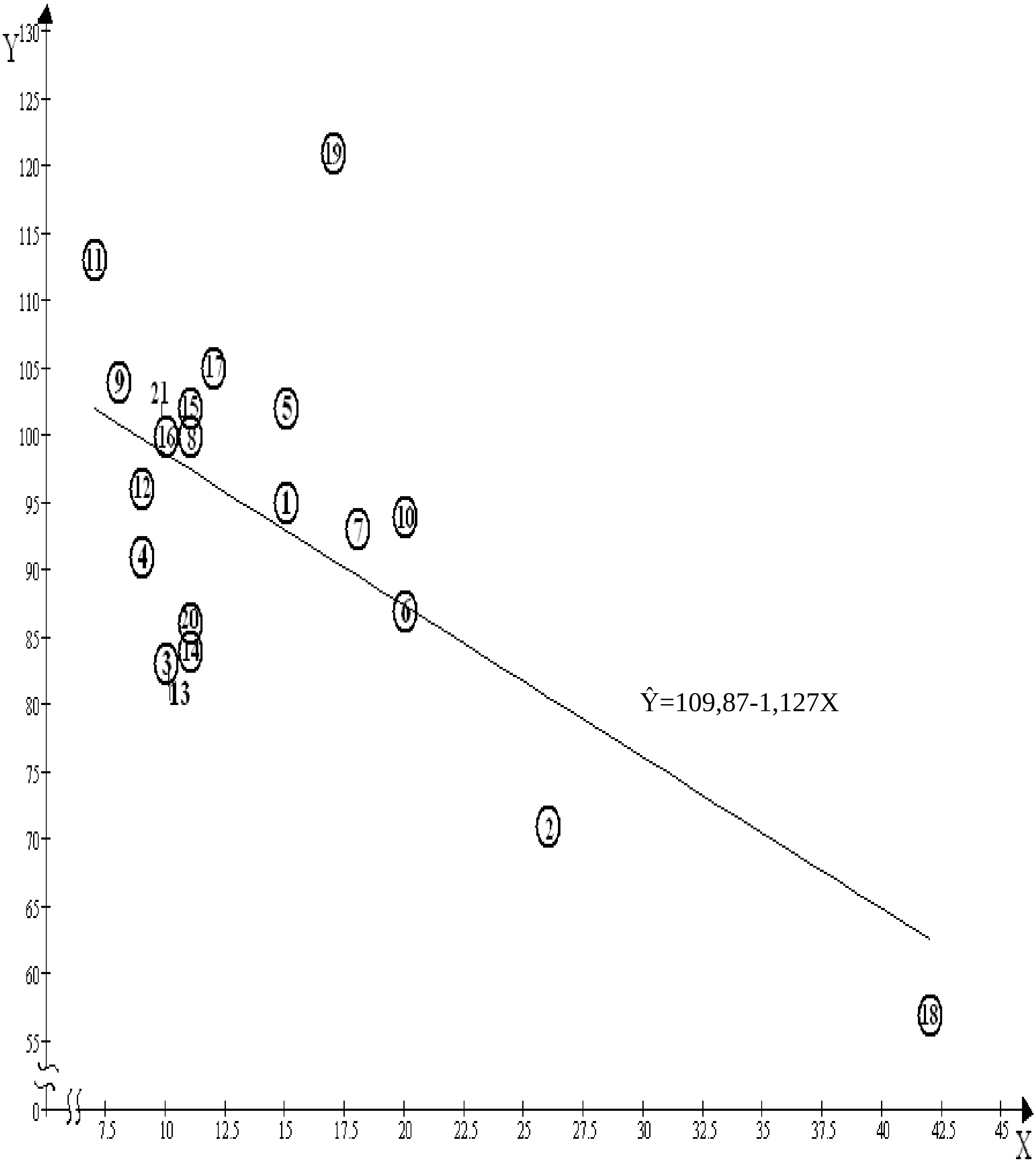

Рисунок 3 − Пример регрессионной модели для тестирования

Смещение $Y_{прогн}$ отсутствует и Δ=0%. Будем для сравнения использовать следующие частные критерии эффективности:

− коэффициент детерминации R^2, который демонстрирует исследователю силу взаимосвязи между Y и X (процент случайных величин x_i, который полностью объясняет поведение случайных величин y_i данным линейным регрессионным уравнением);

− модуль величины смещения результата прогноза (является следствием изменения положения нового регрессионного уравнения при отбрасывании части исходных измерений);

− 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, которое необходимо для реализации отбрасывания аномальных и ненадежных измерений.

Первый частный критерий традиционный для регрессионного анализа и определяется известным соотношением [1]:

$$R^2 = \frac{\left\{\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)\Big/ n\right\}^2}{\left\{\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\Big/ n\right\}\left\{\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\Big/ n\right\}} \quad (5)$$

Второй частный критерий достаточно прост:

$$\Delta_{прогн}=|(A\cdot X_{прогн} + B) - (A'\cdot X_{прогн} + B')|, \quad (6)$$

где: первое и второе слагаемое, соответственно, линейное регрессионное уравнение до отбрасывания части статистики и линейное регрессионное уравнение после отбрасывания части статистики (A, B и A', B' находятся с использованием традиционного метода наименьших квадратов).

Наибольшие проблемы возникли с третьим критерием − доверительным интервалом прогноза. Дело в том, что используемая статистика невязок имеет не нормальное распределение (1). Усечение аномальных и значительных по величине невязок еще более усугубляет эту проблему. В этой связи использовать квантили Стьюдента $t(P_{дов}$; k=n-2) не корректно. Нами использовались доверительные интервалы свободные от закона распределения случайных величин невязок [4]. Пусть e_1, e_2, e_3, …, e_i порядковые статистики и $e_1<e_2<e_3$ … $e_{i-1}<$ e_i. При этом доверительный интервал (e_1, e_i) симметричен относительно своего медианного значения e_k и определяется доверительной вероятностью:

$$P_{дов} = I_{1/2}(1,i) - I_{1/2}(i,1), \quad (7)$$

$$I_x(a,b) = 1 - I_{1-x}(b,a),\ (0 \le x \le 1)\ [5], \quad (8)$$

$$I_x(a,b) = \frac{1}{B(a,b)}\int_0^x t^{a-1}\cdot(1-t)^{b-1}dt\ \ [5], \quad (9)$$

$$B(z,\omega) = \frac{\Gamma(z)\cdot\Gamma(\omega)}{\Gamma(z+\omega)}\ [5], \quad (10)$$

Γ(n+1) = 1·2·3…(n-1)·n = n! для целых значений аргумента [5] (11)

где: $I_x(a,b)$ − неполная Бета-функция; B(z, ω) – Бета-функция; Г(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 по (7-11) заданное значение $P_{дов}$. Найденная пара и представляет размах доверительного интервала. Естественно, что точное значение величины доверительного интервала таким способом найти не удается из-за дискретного характера порядковой статистики, но с этим приходится мириться.

В заключение, в силу необходимости, остановимся на критерии Кука (в англоязычной литературе его часто называют расстоянием Кука) [1]:

$$D_i = \left\{\frac{e_i}{\sigma_e\sqrt{1-h_{ii}}}\right\}^2\cdot\left\{\frac{h_{ii}}{1-h_{ii}}\right\}\cdot\frac{1}{p}\Big|_{max}, \quad (12)$$

где: e_i – i-я невязка, полученная при использовании полного набора исходных данных; h_{ii} − рычаг, оценивающий влияние i-й точки; σ_e − СКО ошибки оценки (2); p − общее количество параметров подбираемых в регрессионной модели (p=2).

В выражении (12)

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \quad (13)$$

Величины h_{ii} ограничены $0< h_{ii}<1$.

Из (12-13) видно, что D_i − фактически определяет влияние i-го остатка на исходное регрессионное уравнение. Таким образом выявляются одиночные аномальные измерения. В случае использования статистики Кука высокого порядка, т.е. оценки одновременного влияния нескольких невязок e_i, наблюдается следующий эффект. Аномальный выброс e_i компенсируется при удалении его из статистических данных несколькими "добротными" измерениями, которые близко расположены к точке с координатами $(\overline{X}, \overline{Y})$. Это резко снижает качество регрессионного уравнения, т.к. при этом существенно уменьшается величина R^2. Справедливости ради отметим, что при этом минимизируется величина смещения Δ.

В случае отбрасывания k невязок одновременно, критерий Кука имеет вид:

$$D_k = \frac{\sum_{i=1}^{n}[(A\cdot X + B) - (A'\cdot X + B')]^2}{p\cdot\sigma_e^2}\Bigg|_{min} \qquad (14)$$

Заметим, что в квадратных скобках (14) приведены значения смещений Δ_i (6), соответствующих n значениям Y. Здесь k − массив отброшенных значений Y.

Результаты сопоставительного анализа метода-прототипа Кука и нового оригинального метода (вариант рис. 2б) для рассматриваемого примера (рис.3) приведены в табл.1.

Таблица 1. Результаты сопоставительного анализа

Используемый метод	Параметры метода	Номера исключенных 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	~ 4	~ $0,4\cdot10^6$
	D_k=0,002	**2, 3, 11, 14, 19**	0,64	0,18	3,8	~ $7\cdot10^6$
	D_k=0,097	**2, 3, 14, 20, 19, 11, 5, 4**	0,77	0,77	1,9	~ $102\cdot10^6$
Предложенный метод (вариант рис.2б)	k=1,27 (D_k=0,17)	**19, 3, 13**	0,7	0,12	2,4	~ $0,5\cdot10^3$
	k=1 (D_k=0,74)	**3, 13, 14, 20, 19**	0,83	1,96	2,2	~ $0,5\cdot10^3$
	k=0,8 (D_k=0,22)	**2, 3, 13, 14, 20, 19, 11, 5**	0,86	1,35	1,8	~ $0,5\cdot10^3$

Сопоставление результатов (табл.1) методов повышения качества прогнозных регрессионных моделей для рассмотренного примера и при одинаковом количестве исключаемых наблюдений свелось к следующему:

- оба метода существенно повышают качество исходной регрессионной модели по критерию R^2, однако предложенный метод здесь выигрывает на 10,5%;

− метод-прототип Кука лучше на ~0,5% сохраняет параметры исходного регрессионного уравнения по критерию Δ_i;

− по величине доверительного интервала ($P_{дов}$ =0,9) незначительно выигрывает предложенный метод;

− по количеству элементарных операций для его реализации метод-прототип полностью проигрывает предложенному и не может быть рекомендован к использованию в современных компьютерных технологиях и в современных автоматизированных информационных системах для ведения оперативного управления (проигрыш достигает $2\cdot10^5$ раз для данного примера при числе отбрасывания точек k=8).

Исследования показали (эти данные отсутствуют в табл.1), что при использовании метода-прототипа Кука при исключении большого количества исходных статистических данных, количество возможных вариантов отбрасывания, по крайней мере, составляет $C_n^2 + C_n^3 + \ldots + C_n^{n/2}$, где C_n^m − количество сочетаний из n элементов по m. В этой ситуации отдельные варианты отбрасывания, при достаточно малых значениях D_k, являются явно ошибочными. При отбрасывании таких наборов измерений резко снижается величина R^2 (а критерий Кука об этом умалчивает). Это связано с сущностью критерия Кука, который направлен на слепое (в варианте использования ЭВМ) отслеживание минимума величины суммарного Δ_i. В машинном варианте, по нашему мнению, критерий Кука возможно применять только при совместном использовании с критерием R^2. На этот факт следует обратить внимание разработчиков системы MINITAB [6] и других систем, где используется статистика Кука. При ручной обработке данных эту проблему решает исследователь путем отбраковки некоторых вариантов, полученных по Куку.

Следует отметить, что и метод-прототип Кука и предложенный метод, и его модификации можно применять для нелинейных регрессионных прогнозных уравнений с внутренней нелинейностью. При этом исходное нелинейное регрессионное уравнение путем специальных преобразований приводится к линейному. Производится отбрасывание части статистики и далее путем обратного преобразования возвращаются к исходному нелинейному уравнению.

Выводы

На основании проделанных исследований можно сделать следующие выводы:

1. Метод-прототип Кука и предложенный оригинальный метод повышения качества линейных прогнозных регрессионных уравнений обеспечивают выявление и отбрасывание не только аномальных, но и не достаточно весомых при регрессионном анализе исходных данных.

2. Детальное сопоставление двух рассматриваемых метода на конкретном примере показало, что метод-прототип Кука обеспечивает выигрыш качества за счет более эффективной стабилизации параметров исходного уравнения при отбрасывании части статистики, а предложенный метод − за счет повышения значения R^2 и уменьшения величины доверительного интервала прогноза при заданной доверительной вероятности.

3. Следует обратить внимание исследователей и разработчиков программного обеспечения при реализации метода Кука на тот факт, что при достаточно малых и мало отличающихся между собой значениях D_k возможны варианты отбрасывания исходной статистики, которая существенно снижает величину R^2 регрессионного уравнения после отбрасывания.

Это объясняется "нацеленностью" критерия Кука любой ценой минимизировать суммарную величину Δ при отбрасывании. При этом аномальные и малозначимые измерения слепо уравниваются с весомыми и отбрасываются совместно.

4. На основании п.3 рекомендуем использовать известный метод Кука, только для ручного анализа. В случае регрессионного анализа на ЭВМ целесообразно или полностью отказаться от метода Кука или его использовать совместно с критерием максимума R^2 при отбрасывании, не учитывая значительное время анализа.

5. Авторы рекомендуют предложенный оригинальный метод повышения качества регрессионных прогнозных моделей для применения в пакетах программ уже существующих и новых вычислительных систем, а также в автоматизированных информационных системах различного назначения. При этом выигрыш в сокращении времени анализа по сравнению с методом-прототипом при $n \geq 20$ может составить 10^5 и более раз.

Литература

1. Дрейпер Н.Р., Смит Г. Прикладной регрессионный анализ. 3-е изд.: Пер. с англ. − М.: Вильямс, 2007. − 912 с.
2. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. − М.: ФИЗМАТЛИТ, 2006. − 816 с.
3. Айвазян С.А. и др. 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. 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.
5. 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.
6. Ханк Дж. Э, Райтс А. Дж., Уичерн Д.У. Бизнес-прогнозирование, 7-е изд.: Пер с англ. − М.: Вильямс, 2003. − 656 с.

Поступила в редакцию 15.03.2010