УДК 004.9 DOI: 10.18698/0236-3933-2017-5-114-131

ПАРНАЯ РЕГРЕССИЯ В *MICROSOFT EXCEL* С ИСПОЛЬЗОВАНИЕМ Р-СПЛАЙНОВ

В.И. Аникин[1] anikin_vi@mail.ru
О.В. Аникина[2] blue-waterfall@yandex.ru
О.М. Гущина[2] g_o_m@tltsu.ru

[1] **Поволжский государственный университет сервиса, Тольятти, Самарская обл., Российская Федерация**
[2] **Тольяттинский государственный университет, Тольятти, Самарская обл., Российская Федерация**

Аннотация

Теоретически показана возможность построения рекурсивных регрессионных моделей на базе усеченных полиномиальных Р-сплайнов. По принципу «программирование без программирования», без написания программного кода *VBA* в *Microsoft Excel* построены общие и рекурсивные модели полупараметрической Р-сплайновой регрессии с использованием усеченных полиномов первого, второго и третьего порядков, отличающиеся впечатляюще простой табличной структурой. Предложенные методы создания табличных Р-сплайновых моделей существенно расширяют возможности *Microsoft Excel* как простого и эффективного средства сглаживания и регрессионного анализа случайных выборок данных

Ключевые слова

Microsoft Excel, табличная модель, регрессионный анализ, полупараметрическая регрессия, Р-сплайн, В-сплайн

Поступила в редакцию 12.10.2016

Введение. Одной из причин, побудившей авторов к исследованиям по теме работы, является наличие в *Microsoft Excel* известного и давнего дефекта, затрудняющего строгий регрессионный анализ данных в этой программе, а именно, встроенные средства *Microsoft Excel* позволяют проводить через случайные точки данных стандартные линии тренда (линейную, полиномиальную и др.), а также показывать на графике вид их уравнения, однако в ряде случаев коэффициенты уравнения выводятся *Microsoft Excel* неверно. Например, если через случайные точки y_i, полученные по формуле $y_i = \sqrt[3]{x_i} + 2\varepsilon\left(rnd_i - 0,5\right)$, где rnd_i — случайное число с равномерным распределением вероятностей на отрезке [0, 1]; $\varepsilon = 0,2$ — амплитуда аддитивной случайной добавки; $i = 0,\ldots, 100$, $x = [0, 10]$, провести полиномиальную линию тренда пятого порядка, то *Microsoft Excel* показывает для нее уравнение:

$$y = 0,0002x^5 - 0,0063x^4 + 0,0638x^3 - 0,3134x^2 + 0,8914x + 0,3102. \quad (1)$$

Если по уравнению (1) на том же графике построить линию регрессии, то при $x > 4$ она сильно отличается от линии тренда, проведенной *Microsoft Excel* (рис. 1, *а*).

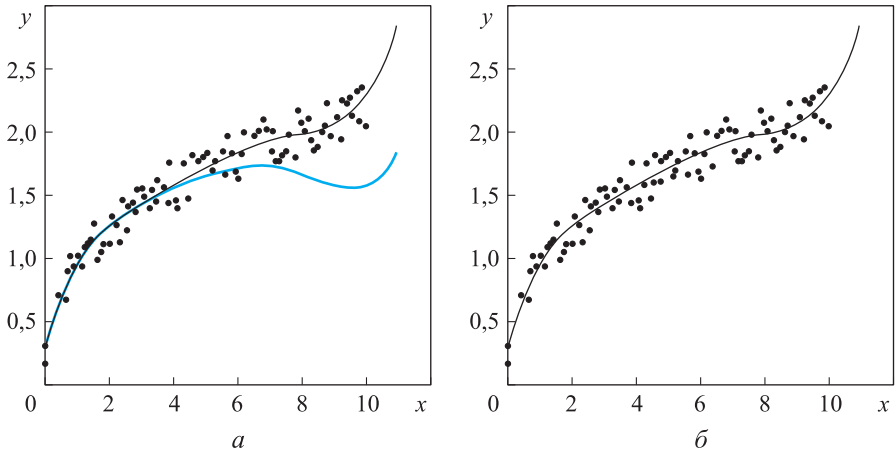

Рис. 1. Линия тренда с параметрами уравнения (1) до (*а*) и после (*б*) их коррекции

Кроме того, в реальности существуют случайные наборы данных, для которых невозможно подобрать подходящую стандартную линию тренда *Microsoft Excel*, а в более сложных случаях нельзя даже указать вид уравнения этой линии, например, для непериодических временных рядов.

В Интернете можно найти несколько бесплатных надстроек *Microsoft Excel*, например, *SRS1 Cubic Spline for Excel V2.5 Released*, *Mathanxl.3.1*, и коммерческая надстройка *XLSTAT* для проведения линии кубического сплайна через заданные точки (x_i, y_i). Эти надстройки являются удобными инструментами интерполяции и экстраполяции данных, но не позволяют сглаживать стохастические данные, т. е. вычислять для них нелинейную линию регрессии (тренда). В работе [1] предложен метод сглаживания случайных данных путем их локального усреднения, реализованный табличными средствами *Microsoft Excel*, без программирования на *VBA*, однако приведенные примеры практического применения этого метода не доказывают его эффективность.

В настоящее время широкое применение нашла мощная техника сглаживания данных с помощью штрафованных Р- и В-сплайнов [2, 3], для которых теоретически и экспериментально доказана несмещенность оценки линии математического ожидания случайных данных [4]. К сожалению, ни одной публикации по использованию штрафованных сплайнов в *Microsoft Excel* авторами обнаружено не было.

Цель настоящей работы — разработка и исследование общих и рекурсивных табличных моделей полупараметрической Р-сплайновой регрессии в *Microsoft Excel*, позволяющих преодолеть затруднения, связанные с указанной ошибкой, а также существенно расширяющих возможности *Microsoft Excel* как простого и эффективного средства сглаживания и регрессионного анализа случайных данных.

Исследования, представленные в настоящей работе, показали возможность реализации этой техники чисто табличными средствами *Microsoft Excel* (про-

стой код *VBA*, использованный в табличных моделях, предназначен исключительно для автоматизации работы пользователя и без него можно обойтись). Созданные табличные модели P-сплайновой регрессии на основе усеченных полиномов первого, второго и третьего порядков, особенно рекурсивные модели, отличаются впечатляющей простотой и высоким качеством сглаживания.

P-сплайны и метод наименьших квадратов. В литературе различают два класса нелинейных регрессий:

1) регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам;

2) регрессии, нелинейные по оцениваемым параметрам.

Здесь рассмотрены нелинейные регрессии первого класса, что позволяет при нахождении параметров уравнения регрессии эффективно использовать метод наименьших квадратов (МНК) [5, 6]. Известно, что этот метод нашел широкое практическое применение в самых различных задачах оптимизации ввиду уникальных статистических свойств: несмещенности, состоятельности и эффективности оценки неизвестных линейных параметров.

Идея P-сплайнов довольно простая. Гладкая неизвестная функция регрессии аппроксимируется функциональной параметрической формой, образованной набором базисных функций большой размерности. Размерность базиса выбирается так, чтобы достигалась достаточная гибкость аппроксимации, тогда как коэффициенты базисных функций штрафуются, обеспечивая желаемую гладкость результирующего функционала.

Пусть имеется n+1 пар точек (x_i, y_i), удовлетворяющих модели $y_i = \hat{y}(x_i) + \varepsilon_i$, где $i = 0,\ldots, n$; $\hat{y}(x)$ — неизвестная функция регрессии; ε_i — независимые случайные добавки с нулевым средним значением и постоянной дисперсией σ^2. Моделируем функцию $\hat{y}(x)$ P-сплайном степени p с усеченным полиномиальным базисом:

$$\hat{y}(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \beta_{pk} (x - \xi_k)^{p+},$$

где

$$(x - \xi_k)^{p+} = \begin{cases} (x - \xi_k)^p & \text{при} \quad x > \xi_k; \\ 0 & \text{при} \quad x \le \xi_k; \end{cases}$$

β_{pk} — вес кусочных функций $(x - \xi_k)^{p+}$; ξ_k — фиксированные координаты узлов P-сплайна.

Как правило, степень полинома, число и положение узлов P-сплайна мало влияют на точность оценки функции регрессии, и можно использовать от 35 до 40 равномерно отстоящих друг от друга узлов для большинства случайных выборок данных и всех гладких функций, не содержащих большого числа осцилляций [4].

Если по каким-либо причинам этот выбор не устраивает, то, определяя новые узлы или сдвигая существующие, P-сплайн можно легко модифицировать,

чтобы он лучше подходил под искомые данные. Нетрудно понять, что варьирование числа узлов Р-сплайна равносильно добавлению к регрессионной модели большего числа предсказывающих переменных. К сожалению, можно добавить слишком много узлов и «переобучить» модель, тем самым подчеркивая в данных случайные флуктуации, которые несущественны.

Для того чтобы избежать «переобучения», на величины весов $\beta_{p1}, \beta_{p2}, \ldots, \beta_{pK}$ базисных функций накладываются ограничения (штрафы). Существует несколько критериев штрафования, из них простейшим и одним из лучших является критерий

$$\sum_{i=1}^{K} \beta_{pk}^2 < C,$$

где C — некоторая константа.

Представленные соображения можно сформулировать математически в векторно-матричном виде, для чего определим вектор $\mathbf{y}^{\mathrm{T}} = [y_0, \ldots, y_n]$, а также матрицы $\mathbf{X}$ и $\mathbf{D}$ вида

$$\mathbf{X} = \begin{bmatrix} 1 & x_0 & \ldots & x_0^p & \ldots (x_0 - \xi_1)^{p+} & \ldots & (x_0 - \xi_K)^{p+} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & \ldots & x_n^p & \ldots (x_n - \xi_1)^{p+} & \ldots & (x_n - \xi_K)^{p+} \end{bmatrix};$$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & 1 \end{bmatrix}.$$

Тогда для заданного сглаживающего параметра λ оценка вектора неизвестных параметров $\boldsymbol{\beta}$ регрессионной модели, где $\boldsymbol{\beta}^{\mathrm{T}} = [\beta_0, \beta_1, \beta_{p1}, \ldots, \beta_{pK}]$, может быть получена путем минимизации целевой функции

$$F = \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \boldsymbol{\beta}^{\mathrm{T}} \mathbf{D} \boldsymbol{\beta}. \qquad (2)$$

Дифференцируя выражение (2) по $\boldsymbol{\beta}$ и приравнивая производную нулю, получаем

$$\lambda \mathbf{D}\boldsymbol{\beta}^* - \mathbf{X}^{\mathrm{T}} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^* \right) = 0,$$

откуда

$$\boldsymbol{\beta}^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{D} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y};$$

$$\hat{\mathbf{y}} = \mathbf{X} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{D} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y}. \qquad (3)$$

Недостатком этого приближения является то, что сглаживание контролируется одним общим, задаваемым пользователем параметром λ, что может вызвать некоторые затруднения, когда данные содержат как быстро, так и медленно осциллирующие участки.

Эффективное решение уравнения (3) табличными средствами невозможно, так как для этого к матрице $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{D})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ необходимо применить сингулярное разложение [7], встроенная поддержка которого в *Microsoft Excel* отсутствует. Отметим, что работа [7] содержит работоспособный программный код R алгоритма решения уравнения (3), который можно использовать для сравнения полученных здесь результатов с результатами выполнения этого кода.

Вместо векторно-матричного решения уравнения (3) для нахождения функции регрессии $\hat{\mathbf{y}}(\mathbf{x})$ авторами настоящей работы в *Microsoft Excel* решена оптимизационная задача:

минимизировать целевую функцию

$$F = \sum_{i=0}^{n}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right)^2 + \lambda\sum_{i=1}^{K}\beta_{pk}^2, \tag{4}$$

где $\mathbf{y}$ — случайная выборка данных; $\hat{\mathbf{y}}$ — P-сплайн с усеченным полиномиальным базисом p-го порядка.

При решении этой задачи возникает проблема подбора наилучшего значения сглаживающего параметра λ. Необходимость оптимизации обусловлена тем, что при $\lambda = 0$ сглаживание отсутствует (переобученная модель), а при больших значениях λ оно слишком велико (линия регрессии игнорирует быстрые средние локальные изменения в данных).

Существует несколько автоматических методов решения этой проблемы.

1. Оптимизация параметра λ по одному из критериев [7]:
- перекрестной проверки (*cross-validation*) $CV(\lambda)$;
- обобщенной перекрестной проверки $GCV(\lambda)$;
- информационному критерию Акаике (*Akaike*) $AIC(\lambda)$;
- скорректированному критерию Акаике $AIC_C(\lambda)$.

2. Оптимизация параметра λ по критерию максимального правдоподобия: $\lambda_{best} = \sigma / \sigma_\beta$ [4], где σ — несмещенное стандартное отклонение вектора остатков $\mathbf{y} - \hat{\mathbf{y}}$; σ_β — стандартное отклонение значений параметров $\boldsymbol{\beta}_p^m = [\beta_{p1}, \ldots, \beta_{pK}]$ полиномиального P-сплайна.

Одним из недостатков табличной реализации в *Microsoft Excel* полиномиальных P-сплайнов является громоздкость получаемой электронной таблицы, обусловленная необходимостью расчета в отдельных ячейках $N_{\text{общ}} = (K + p + 1)(n + 1)$ значений базисных функций сплайна.

Однако для полиномиальных P-сплайнов существует возможность строить рекурсивные табличные модели, объем эквивалентных вычислений в которых сокращается до пересчета $N_{\text{рек}} = (p + 1)(n + 1)$ ячеек, т. е. более чем на порядок

при больших значениях K. Так, при $n = 100$, $K = 40$, $p = 1$ для общей регрессионной модели имеем $N_{\text{общ}} \approx 4200$, а для рекурсивной — $N_{\text{рек}} \approx 200$.

Действительно, вводя обозначения

$$\hat{y}^{(p)} = \hat{y}^{(p)}\left(\beta_0,...,\beta_p,x\right) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} \beta_{pk}\left(x-\xi_k\right)^{p+}$$

и дважды дифференцируя это выражение по x, получаем

$$\frac{d\hat{y}^{(p)}}{dx} = \beta_1 + 2\beta_2 x + ... + p\beta_p x^{p-1} + p\sum_{k=1}^{K-1}\beta_{pk}\left(x-\xi_k\right)^{(p-1)+};$$

$$\frac{d\hat{y}^{(p)}}{dx} = p\left[\beta_1 + \beta_2 x + ... + \beta_p x^{p-1} + \sum_{k=1}^{K-1}\beta_{pk}\left(x-\xi_k\right)^{(p-1)+}\right] - \sum_{m=1}^{p-1}\left(p-m\right)\beta_m x^{m-1};$$

$$\frac{d\hat{y}^{(p)}}{dx} = p\,\hat{y}^{(p-1)}\left(\beta_1,...,\beta_p,x\right) - \sum_{m=1}^{p-1}\left(p-m\right)\beta_m x^{m-1}; \qquad (5)$$

$$\frac{d^2\hat{y}^{(p)}}{dx^2} = 2\beta_2 + ... + p\left(p-1\right)\beta_p x^{p-2} + p\left(p-1\right)\sum_{k=1}^{K-2}\beta_{pk}\left(x-\xi_k\right)^{(p-2)+};$$

$$\frac{d^2\hat{y}^{(p)}}{dx^2} = p\left(p-1\right)\hat{y}^{(p-2)}\left(\beta_2,...,\beta_p,x\right) - \sum_{m=1}^{p-2}\left(p-m\right)\left(p-m-1\right)\beta_m x^{m-2}. \qquad (6)$$

В конечных разностях равенства (5) и (6) имеют вид

$$\frac{\hat{y}_i^{(p)} - \hat{y}_{i-1}^{(p)}}{h_i} = p\,\hat{y}^{(p-1)}\left(\beta_1,...,\beta_p,\,x\right) - \sum_{m=1}^{p-1}\left(p-m\right)\beta_m x^{m-1};$$

$$\frac{\hat{y}_i^{(p)} - 2\hat{y}_{i-1}^{(p)} + \hat{y}_{i-2}^{(p)}}{h^2} = p\left(p-1\right)\hat{y}^{(p-2)}\left(\beta_2,...,\beta_p,x\right) - \sum_{m=1}^{p-2}\left(p-m\right)\left(p-m-1\right)\beta_m x^{m-2},$$

откуда следуют рекурсивные формулы

$$\hat{y}_i^{(p)} = \hat{y}_{i-1}^{(p)} + \left[p\,\hat{y}^{(p-1)}\left(\beta_1,...,\,\beta_p,\,x_i\right) - \sum_{m=1}^{p-1}\left(p-m\right)\beta_k x_i^{m-1}\right]h; \qquad (7)$$

$$\hat{y}_i^{(p)} = 2\hat{y}_{i-1}^{(p)} - \hat{y}_{i-2}^{(p)} +$$

$$+\left[p\left(p-1\right)\hat{y}^{(p-2)}\left(\beta_2,...,\beta_p,x\right) - \sum_{m=1}^{p-2}\left(p-m\right)\left(p-m-1\right)\beta_m x^{m-2}\right]h^2, \qquad (8)$$

где h — шаг дискретизации по оси x.

Следовательно, если известны значения $\hat{y}^{(p-1)}\left(\beta_1,...,\,\beta_p,\,x_i\right)$ полиномиального Р-сплайна порядка p–1, построенного для значений параметра $\beta_1,\ldots,\beta_p$, или значения $\hat{y}^{(p-2)}\left(\beta_2,...,\,\beta_p,\,x_i\right)$ полиномиального Р-сплайна порядка p–2, постро-

енного для значений параметров $\beta_2,\ldots,\beta_p$, то по формулам (7) и (8) легко найти значения Р-сплайна $\hat{y}^{(p)}(\beta_0,...,\beta_p,\ x_i)$ порядка p при условии, что построена рекурсивная модель полиномиального Р-сплайна первого порядка.

В частности, для полиномиальных Р-сплайнов второго и третьего порядков имеем

$$\hat{y}_i^{(2)} = \hat{y}_{i-1}^{(2)} + [2\hat{y}^{(1)}(\beta_1,\ \beta_2,\ x_i) - \beta_1]h_i; \tag{9}$$

$$\hat{y}_i^{(3)} = \hat{y}_{i-1}^{(3)} + [3\hat{y}^{(2)}(\beta_1,...,\ \beta_3,\ x_i) - 2\beta_1 - \beta_2 x_i]h_i; \tag{10}$$

$$\hat{y}_i^{(3)} = 2\hat{y}_{i-1}^{(3)} - \hat{y}_{i-2}^{(3)} + \left[6\hat{y}^{(1)}(\beta_2,\ \beta_3,\ x) - 2\beta_2\right]h_i^2. \tag{11}$$

К достоинствам общих (нерекурсивных) Р-сплайновых табличных моделей можно отнести их универсальность, возможность использования любых базисных функций, в частности, В-сплайнов [3], к их недостаткам — большой объем вычислений в ходе оптимизации и громоздкость табличной модели.

Достоинства рекурсивных Р-сплайновых табличных моделей — существенно меньший объем вычислений и очень простая структура табличной модели, недостаток — в качестве базиса Р-сплайнов можно использовать только усеченные полиномиальные функции.

Постановка модельных экспериментов. Модельные эксперименты включали в себя создание по формулам (8), (9), (11) общих и рекурсивных регрессионных Р-сплайновых моделей и их сравнение в *Microsoft Excel*.

Случайные выборки входных пар данных (x_i, y_i), применяемые в модельных экспериментах, генерировались по формулам $y_i = y(x_i) = \mathrm{M}(x_i) + \varepsilon N_i(0,\ 1)$, где $i = 0,...,n$; $\mathrm{M}(x)$ — заданная детерминированная нелинейная функция, равная математическому ожиданию случайной функции $y(x)$; $N_i(0,\ 1)$ — стандартное нормальное распределение; ε — амплитуда случайной аддитивной добавки.

В экспериментах использовано несколько функций $\mathrm{M}(\mathbf{x})$ с равномерным шагом дискретизации по оси x и заданными параметрами a_i:

$$\mathrm{M}(x_i) = \sqrt[3]{x_i},\ x_i = [0,\ 10],\ \ i = 0,...,100;$$

$$\mathrm{M}(x_i) = a_0 + a_1 x_i + a_2 \sin(x_i),\ \ x_i = [0,\ 10],\ i = 0,...,100.$$

Разработка табличных моделей парной Р-сплайновой регрессии выполнена по предложенной технологии алгоритмического табличного моделирования (АТМ) [8, 9]. Достоинством этой технологии является то, что искомую табличную модель создают не методом проб и ошибок, а основываясь на алгоритме решения задачи, что позволяет получать логически и структурно безупречные электронные таблицы.

Общий и рекурсивный алгоритмы решения задачи парной регрессии с использованием Р-сплайна первогого порядка приведены ниже:

```
'============================================================
' Входы программы
' x0 – координата x первого элемента случайной выборки данных y
' n – число интервалов дискретизации по оси x
' h – равномерный шаг дискретизации по оси x
' y0...yn – элементы случайной выборки данных y
' nk - число элементов данных между соседними узлами Р-сплайна
' ξ0...ξk – ' координаты узлов Р-сплайна на оси x
' β0,β1,βp1...βpk – начальные значения коэффициентов базисных
функций
' λ - параметр сглаживания
' Параметры, переменные и код программы
k=div(n/nk) ' всего узлов сплайна, кроме левого начального
' === Цикл вычисления значений Р-сплайна ===
' i,j – счетчики циклов
for i=0 to n  ' цикл для общей Р-сплайновой табличной модели
  xi=ifelse(i>0;xi-1+h;x0)
  ' цикл вычисления значений базисных функций fj(xi)
  f1(xi)=β0
  f2(xi)=β1*xi
  for j=3 to k+2
    fj(xi)=ifelse(xi-ξj>0;βp(j-2)*(xi-ξj);0)
  next
  ŷi = Σ_{j=1}^{k+2} fj(xi)   ' i-е значение функции регрессии
  dyi=yi-ŷi      ' i-й остаток
next
Fs = Σ_{i=0}^{n} dyi^2   ' сумма квадратов остатков
Fλ = Σ_{i=1}^{k} βpi^2   ' сумма квадратов сплайн-коэффициентов
F=Fs+λFλ     ' значение целевой функции
' === Статистика остатков ===
R2=cor2(y,ŷ)    ' коэффициент детерминации
dy = mean(dy)    ' среднее значение остатков
s = sqrt(var(dy)n/(n-1))      ' стандартное отклонение
sβ = sqrt(var(βp))
DW=2(1-cor(ŷ,ŷт))   ' коэффициент автокорреляции Дарбина — Уотсона
'
============================================================
Цикл for i=0 to n для рекурсивной Р-сплайновой табличной модели:
'
============================================================
for i=0 to n
  xi=ifelse(i>0;xi-1+h;x0)
  kki=ifelse(i=0;β1;ifelse(xi mod h=0;βpi;0))
  ' текущий коэффициент наклона ломаной линии Р-сплайна
  kcuri=ifelse(i<2;k0;kcuri-1+ifelse(xi mod h=0;kki-1;0))
  ŷi=ifelse(i=0;β0; ŷi-1+kcuri*hi)  ' ŷ-функция регрессии
  dyi=yi-ŷi      ' i-й остаток
next
'============================================================
```

Псевдокод рекурсивного алгоритма отличается от псевдокода общего алгоритма лишь циклом вычислений по i, а структура рекурсивной табличной модели отличается от структуры простой модели парной параметрической регрессии двумя дополнительными вектор-столбцами k_k и k_{cur} высотой n+1 ячеек.

На основе приведенных алгоритмов были созданы и исследованы регрессионные табличные модели для входных модельных данных $y_i = a_0 + a_1 x_i + a_2 \sin(x_i) + + N_i(0,1)$ (табл. 1).

Таблица 1

Исследованные регрессионные P-сплайновые табличные модели

Тип модели	Уравнение регрессии
P-сплайновая первого порядка: общая	$\hat{y}_i = \sum_{m=0}^{1} \beta_m x_i^m + \sum_{k=1}^{K} \beta_{1k} (x_i - \xi_k)^+$
рекурсивная	$\hat{y}_i = \hat{y}_{i-1} + \beta_i h$
P-сплайновая второго порядка: общая	$\hat{y}_i = \sum_{m=0}^{2} \beta_m x_i^m + \sum_{k=1}^{K} \beta_{2k} (x_i - \xi_k)^{2+}$
рекурсивная	$\hat{y}_i^{(2)} = \hat{y}_{i-1}^{(2)} + \left[2\hat{y}^{(1)}(\beta_1, \beta_2, x_i) - \beta_1 \right] h$
P-сплайновая третьего порядка: общая	$\hat{y}_i = \sum_{m=0}^{3} \beta_m x_i^m + \sum_{k=1}^{K} \beta_{2k} (x_i - \xi_k)^{3+}$
рекурсивная	$\hat{y}_i^{(3)} = 2\hat{y}_{i-1}^{(3)} - \hat{y}_{i-2}^{(3)} + \left[6\hat{y}^{(1)}(\beta_2, \beta_3, x_i) - 2\beta_2 \right] h^2$
Общая B-сплайновая	$\hat{y}_i = \sum_{m=0}^{1} \beta_m x_i^m + \sum_{k=1}^{K} \beta_{1k} B(x_i - \xi_k)$

Оптимизация параметров этих уравнений выполнена с помощью инструмента *Microsoft Excel* «Поиск решения» методом обобщенного приведенного градиента (ОПГ) при диалоговых настройках, приведенных в табл. 2.

Таблица 2

Настройки диалогового окна *Microsoft Excel* «Поиск решения»

Параметр	P-сплайновая регрессия
Целевая функция	$F_{\text{цел}} = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{k=1}^{K} \beta_{pk}^2$
Изменяемые переменные	$\boldsymbol{\beta}^{\mathrm{T}} = (\beta_0, \ldots, \beta_p, \beta_{p1}, \ldots, \beta_{pK})$
Ограничения	$\lvert\beta_i\rvert, \lvert\beta_{pi}\rvert \le 50$

Начальные значения изменяемых переменных для P-сплайновой регрессии равны $\beta_0 = y_{\text{ср}}$, $\beta_1, \ldots, \beta_p = 0$, $\beta_{p1}, \ldots, \beta_{pK} = 0$.

Качество регрессии было оценено путем статистического анализа остатков $dy_i = y_i - \hat{y}_i$, $i = 0,..., n$ [10]. Для каждой табличной регрессионной модели вычислено среднее значение $\overline{y} = \frac{1}{n+1}\sum_{i=0}^{n} y_i$, коэффициент детерминации $R^2 = \mathrm{cor}^2(y, \hat{y})$, несмещенное стандартное отклонение остатков $s = \sqrt{\mathrm{var}(dy)\frac{n}{n-1}}$, стандартное отклонение s_β коэффициентов $\boldsymbol{\beta}_p$ и наилучшее значение параметра $\lambda_{best} = s / s_\beta$ (для Р-сплайнов), коэффициент автокорреляции Дарбина — Уотсона $DW = 2[1-\mathrm{cor}\,(dy(x), dy(x+h)]$.

Наилучшее значение параметра λ_{best} в Р-сплайновых моделях регрессии оценивалось не только по формуле $\lambda_{best} = s / s_\beta$, но и визуально, следуя принципу: в нескольких последовательных итерациях подбирается такое наименьшее значение λ, при котором график линии регрессии становится гладким, и при этом обеспечивается равномерность и симметричность графика разброса остатков относительно оси x.

Результаты моделирования. Прежде чем представить результаты моделирования, поясним особенности построения регрессионных табличных моделей в *Microsoft Excel*. Для примера рассмотрим простую, но важную рекурсивную Р-сплайновую модель первого порядка, так как без нее, как отмечалось, невозможно создание рекурсивных Р-сплайновых моделей более высоких порядков. Структура этой модели показана на рис. 2. В соответствии с технологией АТМ константы и параметры модели сохраняются в изолированных ячейках, а изменяемые переменные модели — в вектор-столбцах. Интервал ячеек \$U\$3:\$DR\$5, содержащий текущие и начальные значения параметров $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_0, \beta_1, \beta_{p1},\ldots, \beta_{p100})$ и положения узлов $\boldsymbol{\xi} = (\xi_1,\ldots, \xi_{100})$ Р-сплайна максимального размера K =100, не показан.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1																
2		a_0	5		n	100		n_k	4			y_{cp}	20,989		DW	2,177
3		a_1	3		ε	1,5		K	25			R^2	0,976		d_u	1,69
4		a_2	6		h	0,100						s	1,481		$4-d_u$	2,31
5												s_β	1,615		d_L	1,65
6		λ	0,9	67,209	Подгонка			$F_{цел}$	217,13	277,62		s/s_β	0,917		1,69 < DW < 2,31	
7																
8		i	i_k	k	$k_{тек}$	x	y_{rnd}	y	ŷ	y-ŷ	M(y)					
9		0	0	8,7656	8,766	0,0	4,154	3,714	4,922	-1,21	5,000					
10		1		0	8,766	0,1	6,689	4,260	5,798	-1,54	5,899					
11		2		0	8,766	0,2	9,268	4,976	6,675	-1,70	6,792					
12		3		0	8,766	0,3	6,589	8,911	7,551	1,36	7,673					
13		4	1	-1,23821	8,766	0,4	9,343	9,784	8,428	1,36	8,537					
14		5		0	7,527	0,5	8,060	9,029	9,181	-0,15	9,377					
15		6		0	7,527	0,6	8,340	10,614	9,933	0,68	10,188					
16		7		0	7,527	0,7	11,612	11,814	10,686	1,13	10,965					

Рис. 2. Структура рекурсивной Р-сплайновой модели первого порядка

Важная особенность электронных таблиц — формулы и результаты расчета по ним сохраняются в одних и тех же ячейках. Это обеспечивает *Microsoft Excel* принципиально новые возможности визуализации данных по сравнению с другими средами моделирования, даже такими мощными, как *MATLAB* и *R*. В частности, с помощью заливки ячеек различными цветами и мощного инструмента условного форматирования: голубым цветом выделены входные ячейки модели, оранжевым — выходные ячейки, важные для пользователя, белым — ячейки с промежуточными формулами и значениями, которые изменять нельзя, зеленым — ячейки с какими-либо особенностями в формулах, желтым — ячейки с комментариями и полезными подсказками. Кроме того, во всех табличных моделях применен стиль ссылок R1C1, естественный для технологии алгоритмического табличного моделирования в *Microsoft Excel*.

В табличной модели (см. рис. 2) используют следующие параметры, переменные и формулы.

1. Входы модели:

R2C3:R5C3 — параметры a_i функции $M(x_i) = a_0 + a_1x_i + a_2\sin(x_i)$; R2C6 — число интервалов n по координате x; n+1 — число случайных точек (x_i, y_i); R3C6 — амплитуда случайной добавки ε; R3C6 — значение равномерного шага h по координате x; R2C9 — число точек (x_i, y_i) между соседними узлами P-сплайна; R6C3 — значение параметра сглаживания λ; R9C2 — начальное значение i счетчика точек (x_i, y_i); R9C6 — значение x_0; R9C8:R109C8 — входной вектор-столбец данных $\mathbf{y} = (y_0, y_1, \ldots, y_n)$.

2. Основные формулы модели:

```
R3C9 := ЦЕЛОЕ(R2C6/R2C9);
R4C6 := СУММКВ(СМЕЩ(base;0;2;1;R3C9));
R6C9 := СУММКВ(СМЕЩ(tblBase;0;8;n+1;1));
R6C10 := R6C9+R6C3*R6C4;
R9C3:R109C3 := ЕСЛИ(ОСТАТ(RC2;R2C9)=0;ЦЕЛОЕ(RC2/R2C9);"");
R9C4 := R3C22;
R10C4:R109C4 := ЕСЛИ(ОСТАТ(RC2;R2C9)=0;СМЕЩ(base;0;ЦЕЛОЕ(RC2/R2C9)+1);0)
R9C5:R10C5 := R9C4;
R11C5:R109C5 := R[−1]C5+ЕСЛИ(ОСТАТ(RC2;R2C9)=1;R[−1]C4;0);
R10C6:R109C6 := R[−1]C+R4C6;
R9C7:R109C7 := RC11+R3C6*НОРМОБР(СЛЧИС();0;1);
R9C9 := base;
R10C9 := base+R3C22*RC6;
R11C9:R109C9 := R[−1]C9+RC5*R4C6;
R9C10:R109C10 := RC8−RC9;
R9C11:R109C11 := R2C3+R3C3*RC6+R4C3*SIN(RC6).
```

3. Выходы модели:

R6C10 — значение целевой функции; R2C13 — среднее значение y_i; R3C13 — коэффициент детерминации R^2; R4C13 — выборочное несмещенное стандартное отклонение остатков s; R5C13 — выборочное стандартное отклонение s_β параметров $\boldsymbol{\beta}_p$; R6C13 — оценка наилучшего значения параметра сглаживания λ; R2C16 — коэффициент Дарбина — Уотсона.

4. Именованные ячейки модели:
base = R3C21, x0 = R9C6, n = R2C6, h = R4C6, K = R3C9, target = R6C10, tblBase = = R9C2.

5. Начальные значения параметров $\boldsymbol{\beta}$ перед поиском оптимального решения: $\boldsymbol{\beta}_{\text{нач}}$ = (R2C13, 0, …, 0).

6. Настройки диалогового окна «Параметры поиска решения»:
target — целевая ячейка; R3C21:R3C47 — изменяемые ячейки; R3C21:R3C47 ≤ 50, R3C21:R3C47 ≥ −50 — ограничения; поиск решения задачи выполняется методом ОПГ.

Отметим, что табличные модели парной параметрической регрессии очень похожи на описанную модель и отличаются от нее параметрами, отсутствием вектор-столбцов $\mathbf{i}_{\text{k}}$, $\mathbf{k}$, $\mathbf{k}_{\text{тек}}$ и видом формул в вектор-столбцах $\hat{\mathbf{y}}$ и M($\mathbf{x}$).

Работа пользователя по оптимизации созданной Р-сплайновой табличной модели сводится к следующим действиям:

1) задать требуемое число точек между узлами Р-сплайна в ячейке R2C9;

2) скопировать вектор-строку начальных значений $\boldsymbol{\beta}_{\text{нач}}$ из ячеек R5C21:R5C122 в ячейки R3C21:R3C122 вектор-строки $\boldsymbol{\beta}$;

3) выполнить команду меню «Данные → Поиск решения», убедиться в правильности настроек диалогового окна «Параметры поиска решения», при необходимости исправить их и щелкнуть кнопку «Найти решение».

Следует отметить, что ввиду наличия случайной добавки значения ячеек вектор-столбца $\mathbf{y}_{rnd}$ изменяются при каждом пересчете рабочего листа с табличной моделью. Поэтому значения (не формулы) ячеек из столбца $\mathbf{y}_{rnd}$ необходимо один раз скопировать в ячейки вектор-столбца $\mathbf{y}$ и затем работать с этой фиксированной выборкой данных. То же самое требуется выполнить при изменении значения параметров a_i во входных ячейках R2C3:R4C3, параметра ε в ячейке R3C6 и вида функции M($\mathbf{x}$) в ячейках R9C11:R109C11.

Изменение значений входных параметров n и h в ячейках R2C6, R4C6 требует следующей более серьезной модификации модели:

1) заменить значения n или h в ячейках R2C6, R4C6 требуемыми;

2) удалить лишние (при уменьшении n) или добавить с помощью маркера заполнения новые (при увеличении n) строки во все вектор-столбцы табличной модели;

3) скопировать значения ячеек из измененного вектор-столбца $\mathbf{y}_{rnd}$ в ячейки вектор-столбца $\mathbf{y}$;

4) скорректировать границы диапазонов данных, выводимых на графики, в диалоговом окне «Выбор источника данных»;

5) скорректировать диапазоны изменяемых ячеек и ограничений в диалоговом окне «Параметры поиска решения».

Для автоматизации работы пользователя в коде *VBA* были разработаны обработчик события *Worksheet_Change*(), выполняющий перечисленные действия (кроме действия 5) при любом изменении содержимого ячеек x_0, n или h, а так-

же ассоциированный с внедренной кнопкой «Подгонка» макрос *Solve*(), который выполняет действие 5, затем копирует в вектор-строку $\boldsymbol{\beta}$ начальные значения $\boldsymbol{\beta}_{\text{нач}}$ и запускает поиск оптимального решения.

Используя специально созданную табличную модель парной параметрической регрессии для выборки данных $y_i = \sqrt[3]{x_i} + N_i(0,1)$, уточняем коэффициенты полинома пятой степени (см. ниже), определяющего стандартную линию тренда *Microsoft Excel*, и построим график линии регрессии (рис. 1, *б*) по этим коэффициентам. Полное совпадение кривых, приведенных на рис. 1, *б*, свидетельствует о том, что для нахождения стандартных линий тренда *Microsoft Excel* применяет МНК.

Корректировка параметров уравнения линии тренда в *Microsoft Excel*

	θ_5	θ_4	θ_3	θ_2	θ_1	θ_0
Параметры *Microsoft Excel*	0,0002	−0,0063	0,0638	−0,3134	0,8914	0,3102
Оптимальные параметры	0,0002330	−0,006230	0,06307	−0,30971	0,88398	0,3137

Результаты P-сплайновой регрессии первого порядка для случайной выборки данных $y_i = a_0 + a_1 x_i + a_2 \sin(x_i) + \varepsilon N_i(0, 1)$, где $\mathbf{a} = [5, 3, 6]$; $\varepsilon = 1{,}5$; $x = [0, 10]$; $h = 0{,}1$, и трех значений параметра сглаживания λ (штриховой линией показана функция $M(x)$) приведены на рис. 3.

Для наилучшего значения параметра $\lambda = 0{,}9$ (рис. 3, *в*–*д*) коэффициент детерминации равен $R^2 = 0{,}970$, т. е. кривая регрессии объясняет 97 % вариаций входных случайных данных. Вид функции остатков показывает, что их дисперсия постоянна, а автокорреляция между соседними значениями y_i отсутствует, что подтверждается тестом Дарбина — Уотсона: $DW = 2{,}17$.

Эксперименты по нахождению оптимального значения параметра λ при изменении случайной добавки ε показали, что оценка $\lambda_{best} \approx s / s_\beta$ является хорошим приближением и может использоваться вместо сложного алгоритма поиска наилучшего значения λ_{best}, реализованного в программе *R* [7].

Стандартные отклонения остатков и коэффициентов β_p (n=100)

ε	0,5	1,0	1,5	2,0	2,5	3,0	4,0
s	0,439	0,945	1,481	1,805	2,802	2,855	3,829
s_β	1,642	1,517	1,615	1,755	1,600	1,564	1,464

С увеличением добавки ε стандартные отклонения остатков, как и следует ожидать, линейно возрастают, а значения s_β остаются постоянными. Последнее объясняется тем, что физически параметр s_β представляет собой оценку среднего значения приращений $k_{\text{тек}}$ тангенсов углов наклона функции $M(x)$ к оси x, которая для фиксированной выборки данных (x_i, y_i) должна оставаться приблизительно постоянной. С увеличением амплитуды случайных добавок ε наилучшее значение λ также возрастает почти линейно.

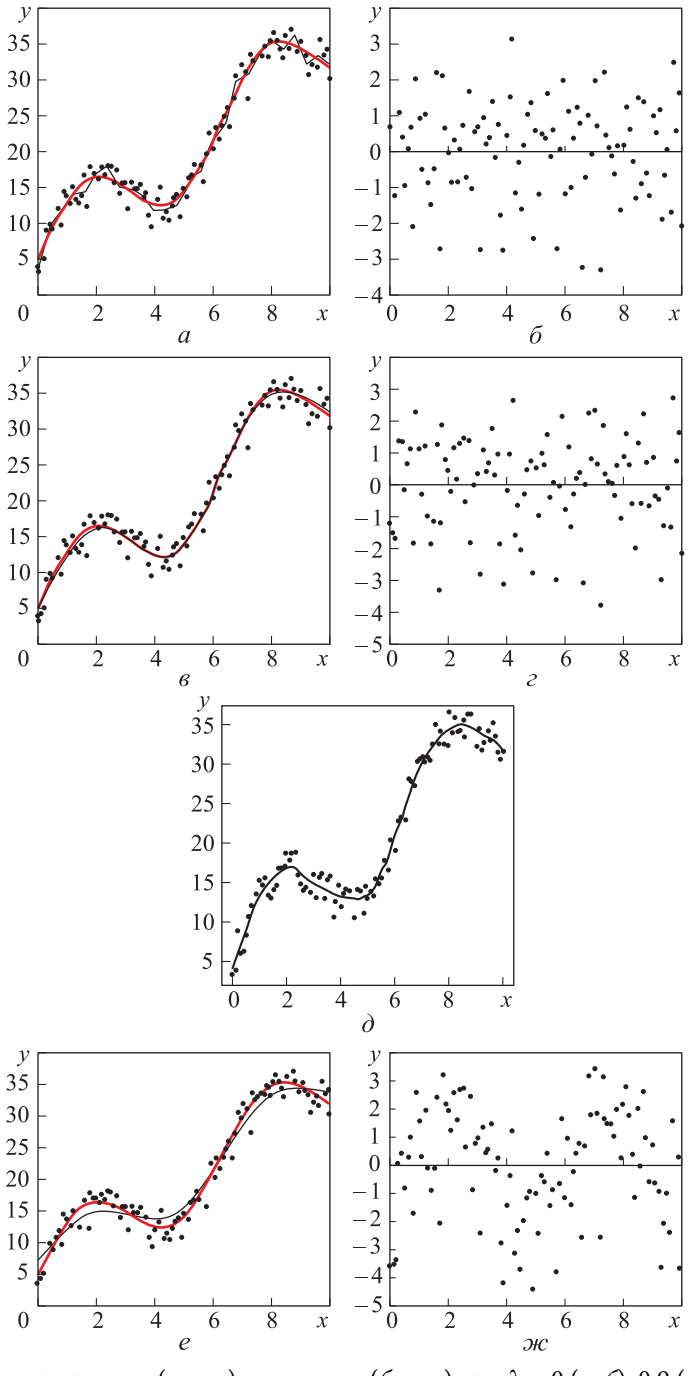

Рис. 3. Функции регрессии (*а, в, е*) и остатков (*б, г, ж*) при $\lambda = 0$ (*а, б*), 0,9 (*в–д*), 10 (*е, ж*), функция регрессии, полученная в программе *R* при $\lambda_{AICC} = 0{,}26$ (*д*)

Функции, приведенные на рис. 4, показывают возможность сглаживания в *Microsoft Excel* случайных выборок данных не только с помощью P-сплайнов, но и B-сплайнов.

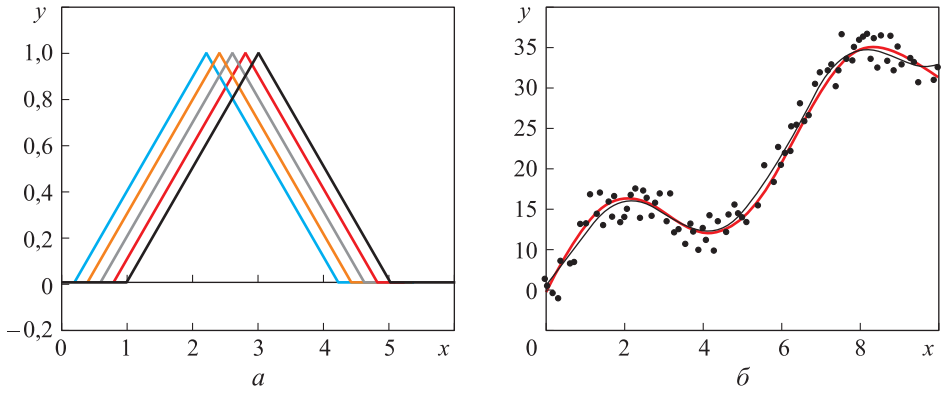

Рис. 4. Базисные функции (*а*) и функции регрессии (*б*) при сглаживании случайных данных треугольными B-сплайнами ($\lambda = 3$, $\varepsilon = 1{,}5$, $R^2 = 0{,}989$, $s = 1{,}43$)

Выводы. Аналитически доказана и экспериментально подтверждена возможность построения в *Microsoft Excel* общих, а также предложенных авторами рекурсивных регрессионных моделей на базе полиномиальных P-сплайнов первого, второго и третьего порядков.

Полученные результаты открывают перспективу широкого и эффективного применения электронных таблиц *Microsoft Excel* для сглаживания случайных данных и решения регрессионных задач, интерполяции и прогноза в экономике, социологии, биологии, медицине, экологии, физике и других областях.

ЛИТЕРАТУРА

1. *Klasson K.T.* Construction of spline functions in spreadsheets to smooth experimental data // Advances in Engineering Software. 2008. Vol. 39. Iss. 5. P. 422–429.
DOI: 10.1016/j.advengsoft.2007.03.006

2. *Eilers P.H.C., Marx B.D., Durban M.* Twenty years of P-splines // SORT. 2015. Vol. 39. No. 2. P. 149–186.

3. *Ruppert D., Wand M.P., Carroll R.J.* Semiparametric regression during 2003–2007 // Electronic Journal of Statistics. 2009. No. 3. P. 1193–1256. DOI: 10.1214/09-EJS525
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2841361

4. *Ruppert D., Wand M.P., Carroll R.J.* Semiparametric regression. New York: Cambridge Univ. Press, 2003. 404 p.

5. *Jang J.S.R., Sun C.T., Mizutani E.* Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence. New York: Prentice-Hall, 1997. 614 p.

6. *Van de Geer S.A.* Least squares estimation // Encyclopedia of Statistics in Behavioral Science. Vol. 2. Wiley, 2005. P. 1041–1045.

7. *Griggs W.* Penalized spline regression and its applications. 2013. 51 p.
URL: https://www.whitman.edu/Documents/Academics/Mathematics/Griggs.pdf
(дата обращения: 16.09.2016).

8. *Аникин В.И., Аникина О.В.* Эффективная техника создания табличных моделей в *Microsoft Excel* // Информационные технологии. 2008. № 10. С. 74–77.

9. *Аникин В.И., Аникина О.В., Зибров П.Ф.* Информационные технологии имитационного моделирования // Формирование современного информационного общества — проблемы, перспективы, инновационные подходы. Материалы международного форума. СПб.: ГОУ ВПО СПбГУАП, 2011. С. 181–189.

10. *Kutner M.H., Nachtsheim C.J., Neter J., Li W.* Applied linear statistical models. New York: McGraw-Hill, 2004. 1396 p.

Аникин Валерий Иванович — д-р техн. наук, профессор, профессор кафедры «Информационный и электронный сервис» Поволжского государственного университета сервиса (Российская Федерация, 445017, Самарская обл., Тольятти, ул. Гагарина, д. 4).

Аникина Оксана Владимировна — канд. техн. наук, доцент кафедры «Прикладная математика и информатика» Тольяттинского государственного университета (Российская Федерация, 445020, Самарская обл., Тольятти, ул. Белорусская, д. 14).

Гущина Оксана Михайловна — канд. пед. наук, доцент, доцент кафедры «Прикладная математика и информатика» Тольяттинского государственного университета (Российская Федерация, 445020, Самарская обл., Тольятти, ул. Белорусская, д. 14).

Просьба ссылаться на эту статью следующим образом:

Аникин В.И., Аникина О.В., Гущина О.М. Парная регрессия в *Microsoft Excel* с использованием Р-сплайнов // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2017. № 5. С. 114–131. DOI: 10.18698/0236-3933-2017-5-114-131

PAIR REGRESSION IN *MICROSOFT EXCEL* BY USING OF P-SPLINES

V.I. Anikin[1] anikin_vi@mail.ru
O.V. Anikina[2] blue-waterfall@yandex.ru
O.M. Gushchina[2] g_o_m@tltsu.ru

[1] Volga Region State University of Servise, Togliatti, Samara Region, Russian Federation
[2] Togliatti State University, Togliatti, Samara Region, Russian Federation

Abstract

The study theoretically shows the possibility of building recursive P-spline regression models based on the truncated polynomial P-splines. Without writing *VBA* code according to the principle of "programming without programming", we created general and recursive models of semi-parametric regression in *Excel* using polynomial P-splines of the first, second and third orders, having a dramatic simplicity in the spreadsheet structure. We optimized the parameters of the models by the generalized reduced gradient method using the *Excel* Solver tool. The simulation experiments confirmed high quality of the regression and computational efficiency of the spreadsheet models. The regression quality was evaluated by statistical analysis of the residuals. For each spreadsheet regression

Keywords

Microsoft Excel, spreadsheet model, regression analysis, semiparametric regression, P-spline, B-spline

model we calculated the coefficient of determination, the unbiased standard deviation of the residuals, the standard deviation of the parameters of the basis functions, the best value of smoothing parameter and the Durbin — Watson's autocorrelation coefficient. The advantages of the general (non-recursive) P-spline spreadsheet models are their versatility, the ability to use all the basic functions, in particular, B-splines, while their disadvantages are a large amount of computations in the optimization process and a bulky spreadsheet model. The recursive P-spline spreadsheet models have such advantages as much smaller amount of computation required in the optimization process and a very simple structure of the spreadsheet model, while the disadvantage is that only polynomial functions can be used as the basis of P-splines. The proposed methods for creating the spreadsheet P-spline models significantly expand capabilities of *Excel* as a simple and effective tool for smoothing and regression analysis of random data samples

Received 12.10.2016

REFERENCES

[1] Klasson K.T. Construction of spline functions in spreadsheets to smooth experimental data. *Advances in Engineering Software*, 2008, vol. 39, iss. 5, pp. 422–429.
DOI: 10.1016/j.advengsoft.2007.03.006

[2] Eilers P.H.C., Marx B.D., Durban M. Twenty years of P-splines. *SORT*, 2015, vol. 39, no. 2, pp. 149–186.

[3] Ruppert D., Wand M.P., Carroll R.J. Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, 2009, no. 3, pp. 1193–1256. DOI: 10.1214/09-EJS525
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2841361

[4] Ruppert D., Wand M.P., Carroll R.J. Semiparametric regression. New York, Cambridge Univ. Press, 2003. 404 p.

[5] Jang J.S.R., Sun C.T., Mizutani E. Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence. New York, Prentice-Hall, 1997. 614 p.

[6] Van de Geer S.A. Least squares estimation. *Encyclopedia of Statistics in Behavioral Science*. Vol. 2. Wiley, 2005, pp. 1041–1045.

[7] Griggs W. Penalized spline regression and its applications. 2013. 51 p.
Available at: https://www.whitman.edu/Documents/Academics/Mathematics/Griggs.pdf (accessed: 16.09.2016).

[8] Anikin V.I., Anikina O.V. Effective technology of making table models in Excel. *Informatsionnye tekhnologii* [Information Technologies], 2008, no. 10, pp. 74–77 (in Russ.).

[9] Anikin V.I., Anikina O.V., Zibrov P.F. Informatsionnye tekhnologii imitatsionnogo modelirovaniya [Informational technologies of imitational simulation]. *Formirovanie sovremennogo informatsionnogo obshchestva — problemy, perspektivy, innovatsionnye podkhody. Materialy mezhdunarodnogo foruma* [Forming modern informational society — problems, prospects, innovative approaches. Proc. int. forum]. Saint-Petersburg, GOU VPO SPbGUAP Publ., 2011, pp. 181–189 (in Russ.).

[10] Kutner M.H., Nachtsheim C.J., Neter J., Li W. Applied linear statistical models. New York, McGraw-Hill, 2004. 1396 p.

Anikin V.I. — Dr. Sc. (Eng.), Professor of Information and Electronic Service Department, Volga Region State University of Service (Gagarina ul. 4, Togliatti, Samara Region, 445017 Russian Federation).

Anikina O.V. — Cand. Sc. (Eng.), Assoc. Professor of Applied Mathematics and Computer Science Department, Togliatti State University (Belorusskaya ul. 14, Togliatti, Samara Region, 445020 Russian Federation).

Gushchina O.M. — Cand. Sc. (Ped.), Assoc. Professor of Applied Mathematics and Computer Science Department, Togliatti State University (Belorusskaya ul. 14, Togliatti, Samara Region, 445020 Russian Federation).

Please cite this article in English as:

Anikin V.I., Anikina O.V., Gushchina O.M. Pair Regression in *Microsoft Excel* by using of P-Splines. *Vestn. Mosk. Gos. Tekh. Univ. im. N.E. Baumana, Priborostr.* [Herald of the Bauman Moscow State Tech. Univ., Instrum. Eng.], 2017, no. 5, pp. 114–131.
DOI: 10.18698/0236-3933-2017-5-114-131