## АНАЛИЗ И СИНТЕЗ СИГНАЛОВ И ИЗОБРАЖЕНИЙ

УДК 519.23

# СПЛАЙНОВАЯ РЕГРЕССИЯ С ПЕРЕМЕННЫМИ ШТРАФНЫМИ КОЭФФИЦИЕНТАМИ*

**В. И. Денисов, А. В. Фаддеенков**

*Новосибирский государственный технический университет,*
*630073, г. Новосибирск, просп. К. Маркса, 20*
*E-mail: faddeenkov@corp.nstu.ru*

Рассмотрена задача построения полупараметрической сплайновой регрессионной модели. Предложена новая модель штрафных сплайнов с переменными штрафными коэффициентами. В модели предполагается, что координаты базисных точек определяются как решение оптимизационной задачи по минимизации остаточной суммы квадратов. Выбор значений штрафных коэффициентов основан на представлении исходной модели в виде модели со случайными эффектами (модели компонент дисперсии). Методами компьютерного моделирования проведён ряд вычислительных экспериментов по восстановлению линии регрессии с различными уровнями шума и при наличии выбросов. Приведены результаты вычислительных экспериментов по восстановлению линии регрессии, демонстрирующие более высокую точность новой модели в сравнении с известными аналогами.

*Ключевые слова:* параметрические и непараметрические методы, полупараметрическая регрессия, модели штрафных сплайнов, модели компонент дисперсии.

**Введение.** В прикладном статистическом анализе в последнее время стало приобретать популярность направление, получившее название полупараметрического моделирования. Это связано с тем, что полупараметрические модели являются компромиссом между двумя крайностями: полностью параметрическим и полностью непараметрическим подходами [1–3].

В первом случае предполагается, что модель, описывающая взаимосвязи в исходных данных, известна с точностью до каких-либо параметров, к оценке которых и сводится решение задачи. Однако неверная спецификация даже некоторых компонент модели может приводить к существенным ошибкам и неправильным прогнозам. Во втором случае знание спецификации модели не обязательно, что даёт бóльшую гибкость. Но при этом непараметрические методы требуют большего количества исходных данных и при малых выборках дают низкую точность.

Преимущество полупараметрических моделей заключается в том, что они сохраняют до некоторой степени гибкость непараметрических моделей и гораздо менее подвержены проблемам неправильной спецификации по сравнению с полностью параметрическими моделями. При этом точность оценивания параметрических компонент полупараметрической модели сравнима с точностью, достигаемой при использовании верно специфицированной параметрической модели.

Целью данной работы является создание новой модели полупараметрической сплайновой регрессии с переменными штрафными коэффициентами.

*Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 13-07-00299а).

**Модель сплайновой регрессии.** Сплайновая полупараметрическая регрессионная модель может быть представлена в следующем виде:

$$y_i = \theta_1 x_{i1} + \ldots + \theta_m x_{im} + \beta_1 f_{i1} + \beta_2 f_{i2} + \ldots + \beta_k f_{ik} + \varepsilon_i, \tag{1}$$

где $y_i$ — значение отклика в $i$-м наблюдении ($i = 1, 2, \ldots, N$); $x_{ij}$ — значение $j$-го регрессора в $i$-м наблюдении ($j = 1, 2, \ldots, m$); $\theta_1, \ldots, \theta_m$, $\beta_1, \ldots, \beta_k$ — неизвестные параметры; $\theta_1 x_{i1} + \ldots + \theta_m x_{im}$ — параметрическая часть модели; $\beta_1 f_{i1} + \beta_2 f_{i2} + \ldots + \beta_k f_{ik}$ — непараметрическая часть; $f_{i1}, f_{i2}, \ldots, f_{ik}$ — значения базисных функций в $i$-м наблюдении; $\varepsilon_i$ — случайная ошибка в $i$-м наблюдении (предполагается, что все ошибки независимы и имеют одинаковое распределение с нулевым средним и дисперсией $\sigma_\varepsilon^2$: $\varepsilon_i \sim (0, \sigma_\varepsilon^2)$, $i = 1, 2, \ldots, N$).

Одним из распространённых вариантов базисных функций являются функции вида

$$f_{ij} = f_j(x_i) = (x_i - b_j)_+^p = \begin{cases} (x_i - b_j)^p & \text{при } (x_i - b_j) > 0, \\ 0 & \text{при } (x_i - b_j) \le 0, \end{cases} \tag{2}$$

где $b_j \in [a, b]$ — узловые точки ($j = 1, 2, \ldots, k$); $p$ — некоторая положительная целая константа.

В матричном виде уравнение (1) может быть представлено как

$$Y = \tilde{X}\tilde{\Theta} + E, \tag{3}$$

где

$$Y = \begin{bmatrix} y_1 & y_2 & \ldots & y_N \end{bmatrix}^T;$$

$$\tilde{X} = \begin{bmatrix} x_{11} & \ldots & x_{1m} & (x_1 - b_1)_+^p & \ldots & (x_1 - b_k)_+^p \\ x_{21} & \ldots & x_{2m} & (x_2 - b_1)_+^p & \ldots & (x_2 - b_k)_+^p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & \ldots & x_{Nm} & (x_N - b_1)_+^p & \ldots & (x_N - b_k)_+^p \end{bmatrix};$$

$$\tilde{\Theta} = \begin{bmatrix} \theta_1 & \ldots & \theta_m & \beta_1 & \ldots & \beta_k \end{bmatrix}^T; \qquad E = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \ldots & \varepsilon_N \end{bmatrix}^T.$$

Оценивание неизвестных параметров этой модели проводится с помощью методов обычного регрессионного анализа. При использовании метода наименьших квадратов оценки неизвестных параметров принимают вид

$$\hat{\tilde{\Theta}} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T Y. \tag{4}$$

Качество воспроизведения исходной зависимости напрямую обусловлено количеством базисных функций и расположением их базисных точек. Однако чрезмерное усложнение модели может приводить к излишней подгонке линии регрессии под исходные данные.

**Модель штрафных сплайнов.** Данная проблема традиционно решается дополнительным сглаживанием модели с переходом к так называемым «штрафным сплайнам» [1]. Идея этого метода заключается в том, что для снижения излишнего влияния непараметрической части на её параметры налагается ограничение (штраф) и вектор оценок параметров вычисляется следующим образом:

$$\hat{\tilde{\Theta}} = (\tilde{X}^T \tilde{X} + \lambda^2 D)^{-1} \tilde{X}^T Y, \tag{5}$$

где $\lambda^2$ — параметр сглаживания; $D$ — $(m+k)\times(m+k)$-матрица штрафа:

$$D = \begin{bmatrix} 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \dots & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0_{(m\times m)} & 0_{(m\times k)} \\ 0_{(k\times m)} & I_{(k\times k)} \end{bmatrix}.$$

При $\lambda^2 = 0$ сглаживание непараметрической части не проводится и оценка (5) совпадает с обычной МНК-оценкой (4). Чрезмерное же увеличение параметра сглаживания ($\lambda^2 \to +\infty$) приводит к тому, что регрессионная модель (3) вырождается в модель, состоящую только из параметрической части. В связи с этим выбору величины параметра сглаживания следует уделять особое внимание.

Проведённые ранее исследования [4] показали, что выбор параметра сглаживания может осуществляться решением оптимизационной задачи по минимизации критерия кроссвалидации:

$$\min_{\lambda} CV = \sum_{i=1}^{N} \Big( \frac{y_i - \hat{y}_i}{1 - S_{\lambda,\,ii}} \Big)^2, \tag{6}$$

где $S_{\lambda,\,ii}$ — $i$-й диагональный элемент матрицы $S_\lambda = \tilde{X}(\tilde{X}^T\tilde{X} + \lambda^2 D)^{-1}\tilde{X}^T$.

Ещё один подход заключается в рассмотрении исходной модели в виде модели компонент дисперсии [1]. В этом случае предполагается, что вместо модели (1) используется модель

$$y_i = \theta_1 x_{i1} + \ldots + \theta_m x_{im} + u_1 f_{i1} + u_2 f_{i2} + \ldots + u_k f_{ik} + \varepsilon_i, \tag{7}$$

где коэффициенты $u_j$ — независимые одинаково распределённые случайные величины $u_j \sim$ $\sim (0, \sigma_u^2)$, $j = 1, 2, \ldots, k$. В матричном виде модель (7) может быть выражена как

$$Y = X\Theta + Zu + E, \tag{8}$$

где

$$X = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{Nm} \end{bmatrix}; \qquad Z = \begin{bmatrix} (x_1 - b_1)_+^p & \dots & (x_1 - b_k)_+^p \\ (x_2 - b_1)_+^p & \dots & (x_2 - b_k)_+^p \\ \vdots & \ddots & \vdots \\ (x_N - b_1)_+^p & \dots & (x_N - b_k)_+^p \end{bmatrix};$$

$$\Theta = \begin{bmatrix} \theta_1 & \dots & \theta_m \end{bmatrix}^T; \quad u = \begin{bmatrix} u_1 & \dots & u_k \end{bmatrix}^T; \quad E = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_N \end{bmatrix}^T.$$

В рамках этого подхода предполагается, что наилучший параметр сглаживания соответствует отношению

$$\lambda^2 = \sigma_\varepsilon^2 / \sigma_u^2. \tag{9}$$

Для оценивания компонент дисперсии $\sigma_u^2$ и $\sigma_\varepsilon^2$ можно воспользоваться методами, основанными на максимизации правдоподобия. Если предположить, что все случайные эффекты модели (8) распределены по нормальному закону, то их оценки максимального правдоподобия решением оптимизационной задачи

$$\max_{\sigma_\varepsilon^2, \sigma_u^2} l(\sigma_\varepsilon^2, \sigma_u^2),$$

где

$$l(\sigma_\varepsilon^2, \sigma_u^2) = -\frac{1}{2}\{\log|V| + Y^T V^{-1}[I - X(X^T V^{-1} X)^{-1} X^T V^{-1}]Y + N\log(2\pi)\};$$

$$V = \sigma_u^2 Z Z^T + \sigma_\varepsilon^2 I.$$

При наложении дополнительных ограничений на независимость оценок компонент дисперсии от фиксированных параметров модели оценки максимального правдоподобия преобразуются в оценки ограниченного максимального правдоподобия (REMLE — Restricted Maximum Likelihood Estimations). В этом случае оптимизационная задача принимает вид

$$\max_{\sigma_\varepsilon^2, \sigma_u^2} l_R(\sigma_\varepsilon^2, \sigma_u^2), \tag{10}$$

где

$$l_R(\sigma_\varepsilon^2, \sigma_u^2) = l(\sigma_\varepsilon^2, \sigma_u^2) - \frac{1}{2}\log|X^T V^{-1} X|.$$

**Модель с переменными штрафными коэффициентами.** Ещё одним важным фактором, оказывающим существенное влияние на качество модели, является расположение узловых точек базисных функций (2). В работах [5, 6] предложено несколько алгоритмов для определения оптимальных координат узловых точек и один из подходов предполагал решение оптимизационной задачи

$$\min_{b_1, \ldots, b_k \in [a, b]} \mathrm{ESS},$$

где $\mathrm{ESS} = e^T e$, $e = Y - \tilde{X}[\tilde{X}^T \tilde{X}]^{-1} \tilde{X}^T Y$.

Следствием оптимизации явился тот факт, что узловые точки в модели располагаются в порядке убывания значимости. Включение в модель базисных функций, соответствующих первым узловым точкам, позволяет учитывать значительные отклонения, вызванные упрощённой структурой модели, и приводит к существенному улучшению качества прогнозирования. Последние же базисные функции по большей части соответствуют незначительным корректировкам модели и могут вызывать повышенную чувствительность к выбросам.

Как отмечалось выше, снижение излишней чувствительности достигается введением параметра сглаживания $\lambda$. Но следует также отметить, что сглаживание, проводимое в рамках модели (3), (5), приводит к наложению одинаковых штрафных коэффициентов на все базисные функции. При этом никак не учитывается сила влияния каждой из базисных функций. Эти аргументы позволяют утверждать, что модель штрафных сплайнов обладает существенным недостатком и необходимо провести её модификацию.

В качестве такой модификации авторы предлагают отказаться от единого параметра сглаживания $\lambda$ для всех базисных функций модели (3) и перейти к разбиению функций на группы с присвоением каждой из них уникального штрафного коэффициента. Для осуществления этой идеи следует заменить модель (8) с одним случайным фактором многофакторной моделью (подробно ознакомиться с методами анализа многофакторных моделей со случайными эффектами можно в работе [7]). В качестве простейшего примера рассмотрим следующую модель:

$$Y = X\Theta + Z_\alpha u_\alpha + Z_\beta u_\beta + E, \tag{11}$$

где

$$X = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{Nm} \end{bmatrix}; \qquad Z_\alpha = \begin{bmatrix} (x_1 - b_1)_+^p & \dots & (x_1 - b_s)_+^p \\ (x_2 - b_1)_+^p & \dots & (x_2 - b_s)_+^p \\ \vdots & \ddots & \vdots \\ (x_N - b_1)_+^p & \dots & (x_N - b_s)_+^p \end{bmatrix};$$

$$Z_\beta = \begin{bmatrix} (x_1 - b_{s+1})_+^p & \dots & (x_1 - b_k)_+^p \\ (x_2 - b_{s+1})_+^p & \dots & (x_2 - b_k)_+^p \\ \vdots & \ddots & \vdots \\ (x_N - b_{s+1})_+^p & \dots & (x_N - b_k)_+^p \end{bmatrix};$$

$$\Theta = \begin{bmatrix} \theta_1 & \dots & \theta_m \end{bmatrix}^T; \quad u_a = \begin{bmatrix} u_1 & \dots & u_s \end{bmatrix}^T;$$

$$u_\beta = \begin{bmatrix} u_{s+1} & \dots & u_k \end{bmatrix}^T; \quad E = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_N \end{bmatrix}^T.$$

Данная модель представляет собой модель компонент дисперсии с двумя случайными факторами $\alpha$ и $\beta$, которым соответствуют два различных вектора случайных эффектов $u_a$ и $u_\beta$. Компоненты дисперсии этих факторов и случайной ошибки обозначим $\sigma_\alpha^2$, $\sigma_\beta^2$ и $\sigma_\varepsilon^2$. Для оценивания компонент дисперсии модели (11) используются те же алгоритмы, что и для модели (8).

Если базисные функции отсортированы соответственно, например, остаточной сумме квадратов, то при правильном выборе границы $s$ в факторах $\alpha$ и $\beta$ будут сосредоточены «сильные» и «слабые» функции, следовательно, для каждой из этих групп базисных функций целесообразно ввести отдельный параметр сглаживания: $\lambda_\alpha^2 = \sigma_\varepsilon^2/\sigma_\alpha^2$, $\lambda_\beta^2 = \sigma_\varepsilon^2/\sigma_\beta^2$.

Оценивание параметров исходной модели (3) в данном случае будет проводиться с использованием двух коэффициентов сглаживания:

$$\hat{\tilde{\Theta}} = (\tilde{X}^T\tilde{X} + \lambda_\alpha^2 D_\alpha + \lambda_\beta^2 D_\beta)^{-1}\tilde{X}^T Y, \tag{12}$$

где

$$D_\alpha = \begin{bmatrix} 0_{(m\times m)} & 0_{(m\times s)} & 0_{(m\times k-s)} \\ 0_{(s\times m)} & I_{(s\times s)} & 0_{(s\times k-s)} \\ 0_{(k-s\times m)} & 0_{(k-s\times s)} & 0_{(k-s\times k-s)} \end{bmatrix}; D_\beta = \begin{bmatrix} 0_{(m\times m)} & 0_{(m\times s)} & 0_{(m\times k-s)} \\ 0_{(s\times m)} & 0_{(s\times s)} & 0_{(s\times k-s)} \\ 0_{(k-s\times m)} & 0_{(k-s\times s)} & I_{(k-s\times k-s)} \end{bmatrix}.$$

Применение подобных переменных штрафных коэффициентов должно повысить гибкость модели и улучшить качество восстановления искомой зависимости.

**Вычислительный эксперимент.** Для сравнительного исследования точности и работоспособности предложенной модели был проведён ряд вычислительных экспериментов. В качестве истинной использовалась модель следующего вида:

$$y_i = y_i^0 + \varepsilon_i = \beta_0 + x_i - 1{,}8x_i^2 + x_i^3 + \beta_1(x_i - b_1)_+ + \beta_2(x_i - b_2)_+ + \varepsilon_i. \tag{13}$$

При моделировании отклика значения независимой переменной $x$ равномерно варьировались на отрезке [0, 1]. При каждой реализации набора исходных данных в функции (13) параметры $\beta_j$, $j = 0, 1, 2$, определялись псевдослучайным датчиком как случайные величины, равномерно распределённые на отрезках [1, 3], [0,25, 0,35], [0,6, 0,7] соответственно. Случайная ошибка для каждого наблюдения генерировалась из предположений о нормальном распределении: $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Дисперсия ошибки $\sigma_\varepsilon^2$ выбиралась таким образом, чтобы величина уровня шума была равна наперёд заданному значению. Уровень шума $\rho = (\sigma_\varepsilon / c) \cdot 100$ % определялся как отношение шум/сигнал в процентах, здесь $c^2 = \frac{1}{N-1} \sum\limits_{i=1}^{N} (y_i^0 - \bar{y}^0)^2$ ($y_i^0$ — незашумлённые измерения отклика).

Оценивание точности построенной оценки линии регрессии проводилось с точки зрения точности воспроизведения истинной модели. В качестве критерия точности использовалась сумма квадратов:

$$\mathrm{ESS}_0 = (Y_0 - \hat{Y})^T (Y_0 - \hat{Y}), \tag{14}$$

где $Y_0$ — вектор наблюдений, построенный по истинной модели (13) при полном отсутствии случайных ошибок.

Оценивание отклика проводилось по модели (3) при $p = 1$:

$$y_i = \theta_0 + \theta_1 x_i + \sum_{j=1}^{k} \beta_j (x_i - b_j)_+ + \varepsilon_i. \tag{15}$$

Для оценивания параметров модели (15) использовались четыре подхода. В первом оценивание параметров проводилось без сглаживания, по формуле (4). Далее результаты этого подхода будем обозначать OLS (Ordinary Least Squares). Во втором (OLS_CV) оценивание параметров осуществлялось со сглаживанием, по формуле (5), выбор параметра сглаживания происходил по критерию кроссвалидации (6). Третий подход (REMLE) — оценивание параметров со сглаживанием, по формуле (5), параметр сглаживания (9) выбирался путём оценивания компонент дисперсии модели (8) по критерию ограниченного максимального правдоподобия (10). Четвёртый подход (REMLE_2) базируется на использовании модели с переменным штрафом, оценивание параметров проводим по формуле (12), а их выбор — путём оценивания компонент дисперсии модели (11) по критерию ограниченного максимального правдоподобия (число базисных функций, входящих в первый фактор, $s = 3$).

Вычислительные эксперименты проводились для дисперсий ошибок, соответствующих уровням шума от 5 до 20 %. Для каждого уровня шума строились модели (15) с числом узлов $k$ от 6 до 15. Координаты базисных точек определялись по алгоритму, основанному на минимизации остаточной суммы квадратов. Для каждой комбинации исходных параметров проводилось по 100 экспериментов с вычислением суммы квадратов (14). Результаты показали, что в среднем наилучшее восстановление исходной линии регрессии достигается с использованием моделей, включающих сглаживание непараметрической части. При этом выбор параметра сглаживания с помощью критерия кроссвалидации даёт наилучшие результаты только при относительно малых ошибках и небольшом количестве узлов. При увеличении числа узлов и возрастании уровня шума бóльшую точность демонстрирует подход, основанный на модели с переменным штрафом. В частности, на

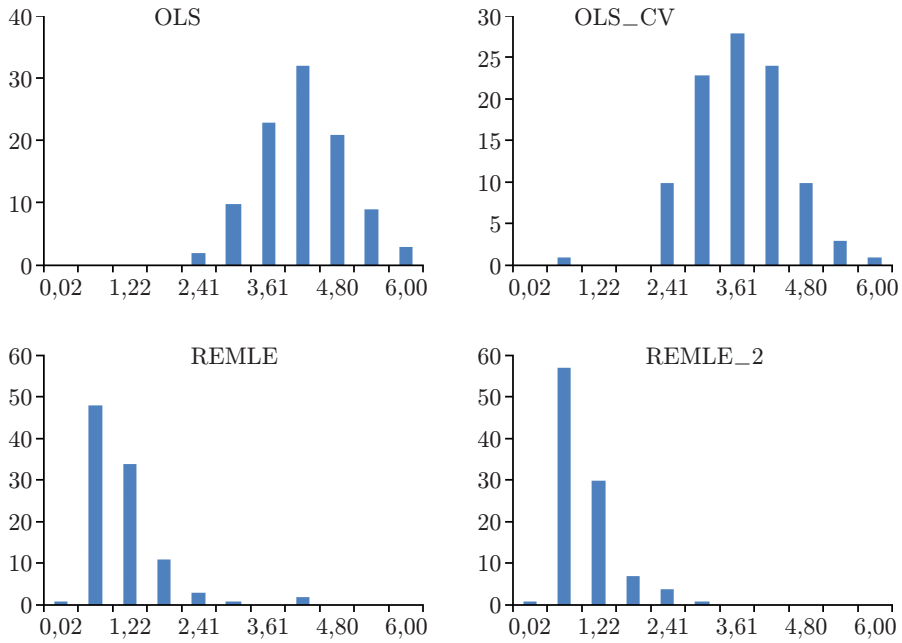


*Рис. 1*

рис. 1 представлены гистограммы, характеризующие разброс $\mathrm{ESS}_0$ при уровне шума 20 % в модели с числом узлов 15 без выбросов.

Для анализа устойчивости рассматриваемых подходов построения линии регрессии был проведён ряд вычислительных экспериментов с искусственным введением в выборку аномальных наблюдений (выбросов). Доля выбросов в исходных данных составляла 3 %, их роль исполняли нормально распределённые случайные величины с десятикратно увеличенной дисперсией ($10\sigma_\varepsilon^2$). Координаты точек, содержащих выбросы, при каждой генерации исходных данных определялись случайно. На рис. 2 представлены гистограммы, характеризующие результаты при уровне шума 20 % для модели с числом узлов 15 при наличии выбросов.

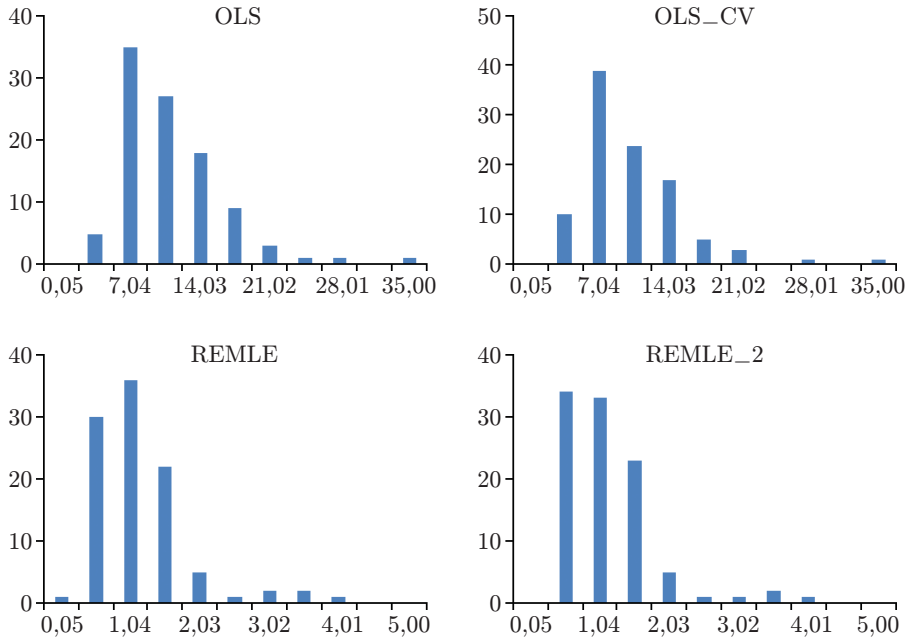


*Рис. 2*

Анализ полученных результатов показывает, что выбор параметра сглаживания, основанный на критерии кроссвалидации, в большинстве случаев лишь незначительно повышает качество прогнозирования и обладает плохой устойчивостью к засорению исходных данных выбросами. Модели, построенные с использованием метода максимального правдоподобия, продемонстрировали значительно бóльшую устойчивость, при этом модель с переменными уровнями штрафа оказалась наиболее точной.

**Заключение.** Исследования, представленные в данной работе, показали, что известная полупараметрическая регрессионная модель штрафных сплайнов обладает существенными недостатками. Была предложена модификация этой модели, основанная на более гибком подходе к выбору штрафных коэффициентов сглаживания. Разбиение базисных функций на подгруппы и определение индивидуального коэффициента сглаживания для каждой из них позволило повысить точность восстановления исходной регрессионной зависимости, что наглядно продемонстрировали проведённые вычислительные эксперименты.

## СПИСОК ЛИТЕРАТУРЫ

1. **Ruppert D., Wand M. P., Carroll R. J.** Semiparametric Regression. N. Y.: Cambridge University Press, 2003. 404 p.
2. **Horowitz J. L.** Semiparametric and Nonparametric Methods in Econometrics. N. Y.: Springer, 2009. 286 p.
3. **Ichimura H., Todd P. E.** Implementing nonparametric and semiparametric estimators // Handbook of Econometrics. Elsevier Science, 2007. Vol. 6, Pt. B. P. 5369–5468.
4. **Денисов В. И., Тимофеев В. С., Бузмакова О. И.** Штрафные сплайны в задаче идентификации полупараметрической регрессии // Науч. вестн. НГТУ. 2011. № 4(45). С. 11–24.
5. **Денисов В. И., Фаддеенков А. В.** К вопросу выбора оптимальных координат узловых точек в моделях полупараметрической регрессии // Науч. вестн. НГТУ. 2012. № 4(49). С. 3–11.
6. **Денисов В. И., Тимофеев В. С., Фаддеенков А. В.** Исследование алгоритмов выбора оптимальных координат узловых точек в полупараметрических моделях штрафных сплайнов // Науч. вестн. НГТУ. 2013. № 2(51). С. 35–44.
7. **Rao C. R., Kleffe J.** Estimation of Variance Components and Applications. N. Y.: North-Holland, 1988. 374 p.

*Поступила в редакцию 6 мая 2014 г.*