УДК 519.233.5

А.С. КОРХИН

О ПОСТРОЕНИИ РЕГРЕССИИ С ПЕРЕКЛЮЧЕНИЯМИ, КОГДА ТОЧКИ ПЕРЕКЛЮЧЕНИЯ НЕИЗВЕСТНЫ

Аннотация. Рассмотрена регрессия с переключениями, когда точки переключения неизвестны. Описан общий метод оценивания точек переключения и параметров линейной регрессии с переключениями общего вида. Приведены примеры его использования.

Ключевые слова: регрессия, переключения, параметры регрессии, оценивание.

ВВЕДЕНИЕ

Регрессия с переключениями — относительно новый вид регрессионных моделей, предназначенных для выявления структурных изменений в моделируемом объекте, если они происходят через некоторые интервалы времени. Такой вид регрессии нашел применение, в частности, при создании экономических моделей. В [1] приведены примеры успешного анализа экономических задач с помощью регрессий с переключениями. Частным случаем такой регрессии являются сплайны, которые в настоящее время широко используются для аппроксимации зависимостей в разных областях науки.

Основной проблемой построения регрессии с переключениями является определение точек переключения. Существует ряд публикаций, в которых исследовано нахождение точки переключения, если она единственная. Подходы к ее оцениванию основаны на методах наименьших квадратов (м.н.к.), максимального правдоподобия (см., например, [2]), а также на использовании статистического критерия Вальда [3]. При числе точек переключения больше единицы в [4] в качестве метода оценивания предлагалось использовать динамическое программирование. Однако оно применимо только тогда, когда на искомые переменные не накладываются ограничения. Например, его нельзя использовать, если априори известно, что в точке переключения линия регрессии непрерывна. Ниже формулируются задачи построения линейной регрессии с переключениями общего вида.

1. ФОРМУЛИРОВКА ЗАДАЧИ ОЦЕНИВАНИЯ

Рассмотрим регрессию вида

$$y_t = \mathbf{x}'_t \boldsymbol{\alpha}^0_{ti} + \mathbf{z}'_t \boldsymbol{\beta}^0 + \varepsilon_t,\ i = 1, \dots, k^0 + 1,\ t = 1, 2, \dots, T, \tag{1}$$

где штрих означает транспонирование; $\mathbf{x}_t \in \Re^n$ и $\mathbf{z}_t \in \Re^m$ — независимые переменные (регрессоры); $\boldsymbol{\alpha}^0_{ti} \in \Re^n$ и $\boldsymbol{\beta}^0 \in \Re^m$ — истинные неизвестные величины параметров регрессии; ε_t — случайная величина, которая вводится для учета малого влияния второстепенных величин на y_t; k^0 — число точек переключения, T — длина интервала наблюдения $[1, T]$. При этом

$$\boldsymbol{\alpha}^0_{ti} = \boldsymbol{\alpha}^0_i = \text{const} \text{ для } t \in I^0_i = [t^0_{i-1} + 1, t^0_i],\ i = 1, \dots, k^0 + 1,\ t^0_0 = 0,\ t^0_{k^0+1} = T, \tag{2}$$

где t^0_i — i-я точка переключения.

Таким образом, $\boldsymbol{\alpha}^0_{ti}$ представляет ступенчатую функцию t на интервале $[1, T]$: все компоненты $\boldsymbol{\alpha}^0_{ti}$ скачком изменяются в точках переключения. Точки переключения t_i^0, $i=1,\ldots,k^0$, считаются неизвестными, как и их число, и являются компонентами вектора $\boldsymbol{\tau}^0$. Согласно (1), (2) в регрессии имеется два вида параметров: изменяющиеся в точках переключения и неизменные на интервале наблюдения. Первый вид параметров объединим в вектор $\boldsymbol{\alpha}^0 \in \Re^{n(k^0+1)}$, состоящий из векторов $\boldsymbol{\alpha}_i^0 \in \Re^n$, $i=1,\ldots,k^0+1$. Для случайной величины ε_t в (1) примем стандартное допущение регрессионного анализа.

Допущение 1. Случайные величины ε_t, $t=1,2,\ldots$, имеют одинаковые математические ожидания и дисперсии, а именно $E\{\varepsilon_t\}=0$, $E\{\varepsilon_t^2\}=\sigma^2$, $t=1,2,\ldots$; они попарно некоррелированы: $E\{\varepsilon_t\varepsilon_\tau\}=0$; $t,\tau=1,2\ldots$; $t\neq\tau$.

Согласно данному допущению

$$E(t)=E\{y_t\}=\mathbf{x}_t'\boldsymbol{\alpha}_i^0+\mathbf{z}_t'\boldsymbol{\beta}^0,\ i=1,\ldots,k^0+1,\ t=1,2,\ldots,T. \tag{3}$$

Из (1), (2) следует, что без введения дополнительных ограничений на параметры регрессии математическое ожидание $E(t)$, рассматриваемое как функция непрерывного t, имеет разрывы в точках переключения.

Если $\boldsymbol{\beta}^0=\mathbf{O}_m$, где $\mathbf{O}_m$ — нулевой m-мерный вектор, и известны точки переключения, то регрессия (1) разбивается на k^0+1 независимую регрессию, параметры которых можно оценить независимо методом наименьших квадратов, решив систему нормальных уравнений.

Оценим параметры регрессии и точки переключения м.н.к., считая величину k заданной, решив задачу

$$S(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\tau})=\sum_{i=1}^{k+1}\sum_{t=t_{i-1}+1}^{t_i}(y_t-\mathbf{x}_t'\boldsymbol{\alpha}_i-\mathbf{z}_t'\boldsymbol{\beta})^2\to\min, \tag{4}$$

где $\boldsymbol{\alpha}_i$ $(i=1,\ldots,k+1)$ и $\boldsymbol{\beta}$ — варьируемые величины параметров, соответственно изменяющихся и не изменяющихся в точках переключения: $\boldsymbol{\alpha}=[(\boldsymbol{\alpha}_1)'\,(\boldsymbol{\alpha}_2)'\,\ldots$ $\ldots(\boldsymbol{\alpha}_{k+1})']'$; t_i, $i=1,\ldots,k$, — варьируемые точки переключения; $\boldsymbol{\tau}=[t_1\ t_2\ \ldots\ t_k]'$. Минимизация в (4) выполняется по $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\tau}$. Чтобы в (4) критерий оптимизации был явной функцией искомой переменной $\boldsymbol{\tau}$, преобразуем формулу для $S(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\tau})$. Для этого введем функции и векторы:

$$f_{ti}(t_{i-1},t_i)=\begin{cases}1, & t_{i-1}<t\le t_i,\\ 0, & t\notin(t_{i-1},t_i],\end{cases}$$

$$\mathbf{f}_i(t_{i-1},t_i)=[f_{1i}(t_{i-1},t_i)\ \ f_{2i}(t_{i-1},t_i)\ \ldots\ f_{Ti}(t_{i-1},t_i)]',\ i=1,\ldots,k+1,$$

а также положим

$$u_t(\boldsymbol{\beta})=y_t-\mathbf{z}_t'\boldsymbol{\beta},\ e_{ti}(\boldsymbol{\alpha}_i,\boldsymbol{\beta})=u_t(\boldsymbol{\beta})-\mathbf{x}_t'\boldsymbol{\alpha}_i,$$

$$\mathbf{e}_i(\boldsymbol{\alpha}_i,\boldsymbol{\beta})=[e_{1i}(\boldsymbol{\alpha}_i,\boldsymbol{\beta})\ \ e_{2i}(\boldsymbol{\alpha}_i,\boldsymbol{\beta})\ \ldots\ e_{Ti}(\boldsymbol{\alpha}_i,\boldsymbol{\beta}]',\ t=1,2,\ldots,T,\ i=1,\ldots,k+1. \tag{5}$$

С учетом введенных обозначений имеем из (4)

$$S(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\tau})=\sum_{i=1}^{k+1}(\mathbf{f}_i'(t_{i-1},t_i)\mathbf{e}_i(\boldsymbol{\alpha}_i,\boldsymbol{\beta}))^2\to\min. \tag{6}$$

Здесь $\boldsymbol{\alpha}\in\Re^{n(k+1)}$, $\boldsymbol{\beta}\in\Re^m$ — непрерывные переменные, вектор $\boldsymbol{\tau}$ имеет целочисленные компоненты t_i, $i=1,\ldots,k$, на которые наложены ограничения

$$1\le t_i\le T,\ t_i-t_{i-1}\ge\theta_i,\ i=1,\ldots,k+1,\ t_0=0,\ t_{k+1}=T, \tag{7}$$

где θ_i — минимальная длина i-го отрезка интервала $[1, T]$, на котором параметр $\boldsymbol{\alpha}_i$, $i=1,\ldots,k+1$, постоянен. Определение θ рассматривается ниже.

Задача (6), (7) имеет непрерывные и дискретные искомые переменные, поэтому она называется задачей оптимизации со смешанными переменными. Она нелинейна по этим параметрам и имеет разрывные вектор-функции $\mathbf{f}_i$, $i=1,\ldots,k+1$. Естественно, задача (6), (7) является многоэкстремальной и, следовательно, достаточно сложной. Ее структура такова, что для фиксированного $\boldsymbol{\tau}$ она преобразовывается в задачу выпуклой оптимизации по $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$, которая при некоторых простых условиях, сформулированных ниже в теореме 1, имеет единственное решение. Общее количество сочетаний точек переключения, удовлетворяющих (7), конечное. Следовательно, решение задачи (6), (7) можно получить за конечное число итераций. Самое простое и точное ее решение — полный перебор всех точек переключения, когда на каждой итерации для их фиксированного сочетания легко определяется оптимальная величина функции $S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau})$.

Однако прогресс в разработке методов дискретной и смешанной оптимизации (например, [5]), создание новых методов глобальной оптимизации (например, [6]), наличие современных программных продуктов для решения различных оптимизационных задач (при этом следует отметить Поиск решения — надстройку табличного процессора MS Excel) позволяет утверждать, что и без полного перебора можно решить задачу (6), (7). Это показано на примерах в данной статье. Если $\boldsymbol{\beta}^0 = \mathbf{O}_m$, то задача (6), (7) может быть также решена методом динамического программирования [4].

Из решения ряда экономических задач априори известно, что функция (3) непрерывна [1]. Отсюда следуют равенства $\mathbf{x}'_{t_i^0}\boldsymbol{\alpha}_i^0 + \mathbf{z}'_{t_i^0}\boldsymbol{\beta}^0 = \mathbf{x}'_{t_i^0}\boldsymbol{\alpha}_{i+1}^0 + \mathbf{z}'_{t_i^0}\boldsymbol{\beta}^0$, $i=1,\ldots,k$. Переходя к варьируемым величинам параметров, получаем

$$\mathbf{x}'_{t_i}(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_{i+1}) = 0, \; i=1,\ldots,k. \tag{8}$$

Выполнение равенств (8) обеспечивает непрерывность функции $E(t)$ по t, когда t — непрерывная величина.

Запишем систему уравнений (8) в виде явной функции точек переключения:

$$\sum_{t=1}^{T}\varphi_{ti}(t_i)\mathbf{x}'_{t_i}(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_{i+1}) = 0, \; i=1,\ldots,k+1, \tag{9}$$

где $\varphi_{ti}(t_i) = \begin{cases} 1, & t = t_i, \\ 0, & t \neq t_i, \end{cases}$ $t=1,2,\ldots,T$.

Частным случаем решения (6), (7), (9) является задача сглаживания последовательности $y_1, y_2, \ldots, y_T$ линейным сплайном с неизвестными точками стыковки отрезков прямых. Тогда в (1) $n=2$, $\mathbf{x}'_t = [1 \;\; t]$, $\boldsymbol{\beta}^0 = \mathbf{O}_m$. Поэтому ограничение (9) принимает упрощенный вид:

$$[1 \;\; t_i]'(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_{i+1}) = 0, \; i=1,\ldots,k. \tag{10}$$

В более сложных задачах сглаживания, когда используются B-сплайны, компонентами вектора $\mathbf{x}_t$ могут быть степени t от первой до третьей.

Решение задачи (6), (7), (9) выполним за два шага. На первом находится регрессия y на регрессор $\mathbf{x}_t$, для чего в (5) полагается $\boldsymbol{\beta} = \mathbf{O}_m$. Также находится регрессия y на $\mathbf{z}$. Полученные оценки точек переключения и параметров являются начальными для решения задачи (6), (7), (9) в общем виде на втором шаге.

Рассмотрим теперь требования к независимым переменным и ограничениям в предложенной задаче оценивания. Введем матрицы $\mathbf{X}=[\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_T]'$ размера $T\times n$ и $\mathbf{Z}=[\mathbf{z}_1\ \mathbf{z}_2\ \ldots\ \mathbf{z}_T]'$ размера $T\times m$ и составим матрицу всех независимых переменных $[\mathbf{X} \vdots \mathbf{Z}]$ размера $T\times(n+m)$.

Теорема 1. Пусть k задано и точки переключения фиксированы, $T\geq n+m$, матрица $\mathbf{W}=[\mathbf{X} \vdots \mathbf{Z}]$ имеет полный ранг. Тогда достаточным условием существования единственного решения задачи (4), (8) является выполнение неравенства $l_{\max}\geq n+m$, где $l_{\max}=\max\limits_{1\leq i\leq k+1} l_i$; $l_i\geq n$, $i=1,\ldots,k+1$, $i\neq\arg\left(\max\limits_{1\leq i\leq k+1} l_i\right)$.

Доказательство. Запишем задачу (4) в матричном виде:

$$S(\boldsymbol{\gamma},\boldsymbol{\tau})=||\mathbf{Y}-\mathbf{W}\boldsymbol{\gamma}||^2\to\min. \tag{11}$$

В (11) минимум берется по $\boldsymbol{\gamma}=[\boldsymbol{\alpha}' \vdots \boldsymbol{\beta}']'\in\Re^N$; $\mathbf{Y}=[y_1\ y_2\ \ldots\ y_T]'$; матрица $\mathbf{W}=[\mathbf{v} \vdots \mathbf{V}]$ имеет размер $T\times N$, где $N=(k+1)n+m$. Матрица $\mathbf{W}$ имеет подматрицы: блочно-диагональную матрицу $\mathbf{v}$ размера $T\times(k+1)$, $\mathbf{v}=\mathrm{diag}\,(\mathbf{v}_i)$, $i=1,\ldots,k+1$, где $\mathbf{v}_i$ имеет размер $l_i\times n$; матрицу $\mathbf{V}$ размера $T\times m$. Здесь $l_i=t_i-t_{i-1}$ — длина отрезка, на котором параметр $\boldsymbol{\alpha}_i$ постоянен, причем $\sum\limits_{i=1}^{k+1} l_i=T$.

По условию теоремы вектор $\boldsymbol{\tau}$ точек переключения фиксирован, поэтому (11) является задачей квадратичной оптимизации. Как известно (см., например [7]), она имеет единственное решение, если матрица $\mathbf{W}'\mathbf{W}$ положительно определена. Это свойство справедливо, если и только если матрица $\mathbf{W}$ имеет полный ранг, что означает линейную независимость ее столбцов $\mathbf{W}_j$, $j=1,\ldots,N$, условием которой является выполнение равенства $\sum\limits_{j=1}^{N}\lambda_j\mathbf{W}_j=\mathbf{O}_T$ для всех и только для всех $\lambda_j=0$, $j=1,\ldots,N$. Данное условие с учетом блочной структуры $\mathbf{W}$ имеет матричный вид

$$\mathbf{W}\boldsymbol{\lambda}=\mathbf{v}\boldsymbol{\lambda}_1+\mathbf{V}\boldsymbol{\lambda}_2=\mathbf{O}_T\ ,\ \boldsymbol{\lambda}\in\mathbb{R}^N,\ \boldsymbol{\lambda}_1\in\mathbb{R}^{n(k+1)},\ \boldsymbol{\lambda}_2\in\mathbb{R}^m, \tag{12}$$

где вектор $\boldsymbol{\lambda}=[\boldsymbol{\lambda}_1'\ \boldsymbol{\lambda}_2']'$ имеет компоненты λ_j, $j=1,\ldots,N$.

Положим $\boldsymbol{\lambda}_1'=[\boldsymbol{\lambda}_{11}'\ \boldsymbol{\lambda}_{12}'\ \ldots\ \boldsymbol{\lambda}_{1,k+1}']$, $\mathbf{V}=[\mathbf{V}_1'\ \ \mathbf{V}_2'\ \ldots\ \mathbf{V}_{k+1}']'$, где $\boldsymbol{\lambda}_{1i}\in\mathbb{R}^n$, матрица $\mathbf{V}_i$ имеет размер $l_i\times m$, $i=1,\ldots,k+1$. Тогда из (12) имеем систему уравнений

$$\mathbf{v}_i\boldsymbol{\lambda}_{1i}+\mathbf{V}_i\boldsymbol{\lambda}_2=\mathbf{O}_{l_i},\ i=1,\ldots,k+1. \tag{13}$$

Пусть без потери общности первый отрезок, на котором параметр $\boldsymbol{\alpha}_i$ постоянен, имеет максимальную длину: $l_{\max}=l_1\geq n+m$ (согласно условию теоремы). Из (13) имеем

$$[\mathbf{v}_1 \vdots \mathbf{V}_1]\begin{bmatrix}\boldsymbol{\lambda}_{11}\\ \boldsymbol{\lambda}_2\end{bmatrix}=\mathbf{O}_{l_1}. \tag{14}$$

Представим матрицу размера $l_1\times(m+n)$ в левой части равенства (14) в блочном виде: $[\mathbf{v}_1 \vdots \mathbf{V}_1]=\begin{bmatrix}\mathbf{M}_1\\ \ldots\\ \mathbf{M}_2\end{bmatrix}$, подматрицы $\mathbf{M}_1$ и $\mathbf{M}_2$ имеют размеры соответственно $(m+n)\times(m+n)$ и $(l_1-m-n)\times(m+n)$. Тогда из (14) получаем уравнение

$\mathbf{M}_1\begin{bmatrix}\boldsymbol{\lambda}_{11}\\ \boldsymbol{\lambda}_2\end{bmatrix}=\mathbf{O}_{m+n}$ относительно $\boldsymbol{\lambda}_{11}$ и $\boldsymbol{\lambda}_2$. Согласно условию теоремы о ранге матрицы $[\mathbf{X} \vdots \mathbf{Z}]$ матрица $\mathbf{M}_1$ имеет полный ранг. Тогда из полученного уравнения имеем $\boldsymbol{\lambda}_{11}=\mathbf{O}_n$, $\boldsymbol{\lambda}_2=\mathbf{O}_m$. Подставив $\boldsymbol{\lambda}_2=\mathbf{O}_m$ в (13), получим $\mathbf{v}_i\boldsymbol{\lambda}_{1i}=\mathbf{O}_{l_i}$, $i=2,\ldots,k+1$. Но матрицы $\mathbf{v}_i$ по условию теоремы имеют полный ранг. Поэтому $\boldsymbol{\lambda}_{1i}=\mathbf{O}_n$, $i=2,\ldots,k+1$. Таким образом, в (12) $\boldsymbol{\lambda}_1$ и $\boldsymbol{\lambda}_2$ — нулевые векторы, т.е. имеем нулевое решение этого уравнения. Следовательно, существует единственное решение задачи (4). При добавлении к ней линейных ограничений (8) также получим задачу с единственным решением, если ограничения будут линейно независимы. Чтобы установить это свойство, запишем (8) в матричном виде: $\mathbf{Q}(\boldsymbol{\tau})\,\mathbf{L}\boldsymbol{\alpha}=\mathbf{O}_k$, где

$$\mathbf{Q}(\boldsymbol{\tau})=\begin{bmatrix}\mathbf{x}'_{t_1} & \mathbf{O}_{1n} & \mathbf{O}_{1n} & \ldots & \mathbf{O}_{1n} & \mathbf{O}_{1n}\\ \mathbf{O}_{1n} & \mathbf{x}'_{t_2} & \mathbf{O}_{1n} & \ldots & \mathbf{O}_{1n} & \mathbf{O}_{1n}\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \mathbf{O}_{1n} & \mathbf{O}_{1n} & \mathbf{O}_{1n} & \ldots & \mathbf{O}_{1n} & \mathbf{x}'_{t_k}\end{bmatrix}', \quad \mathbf{L}=\begin{bmatrix}-\mathbf{J}_n & \mathbf{J}_n & \mathbf{O}_{nn} & \ldots & \mathbf{O}_{nn} & \mathbf{O}_{nn}\\ \mathbf{O}_{nn} & -\mathbf{J}_n & \mathbf{J}_n & \ldots & \mathbf{O}_{nn} & \mathbf{O}_{nn}\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \mathbf{O}_{nn} & \mathbf{O}_{nn} & \mathbf{O}_{nn} & \ldots & -\mathbf{J}_n & \mathbf{J}_n\end{bmatrix},$$

$\mathbf{J}_n$ — единичная матрица порядка n. Матрицы $\mathbf{Q}(\boldsymbol{\tau})$ и $\mathbf{L}$ соответственно имеют размеры $k\times nk$ и $nk\times n(k+1)$.

Применив к матрице $\mathbf{Q}(\boldsymbol{\tau})\mathbf{L}$ неравенство Сильвестра, получим

$$\rho(\mathbf{Q}(\boldsymbol{\tau}))+\rho(\mathbf{L})-nk\le\rho(\mathbf{Q}(\boldsymbol{\tau})\mathbf{L})\le\min(\rho(\mathbf{Q}(\boldsymbol{\tau})),\rho(\mathbf{L})), \tag{15}$$

где $\rho(\mathbf{A})$ — ранг матрицы $\mathbf{A}$.

Из структур матриц $\mathbf{Q}(\boldsymbol{\tau})$ и $\mathbf{L}$ следует, что $\rho(\mathbf{Q}(\boldsymbol{\tau}))=k$, $\rho(\mathbf{L})=nk$. Тогда из (15) вытекает $\rho(\mathbf{Q}(\boldsymbol{\tau})\mathbf{L})=k$, что и требовалось доказать. Теорема доказана.

Чтобы выполнить условия теоремы 1, наложенные на величину θ_i в (7), следует считать θ_1 или θ_k равными $n+m$, а остальные θ_i — равными n. Для упрощения можно положить $\theta_i=n+m$, $i=1,\ldots,k+1$.

2. ОРГАНИЗАЦИЯ ВЫЧИСЛИТЕЛЬНОГО ПРОЦЕССА

Описанные методы предназначены для оценивания точек переключения и параметров регрессии, если число k этих точек задано. Однако на практике число k неизвестно или известно приближенно. Поэтому предлагается следующая процедура. Качество оценивания определим информационным критерием, например критерием Акаике. В этом случае учитывается закон распределения случайных величин ε_t, $t=1,2,\ldots$, в (1). Будем считать его нормальным и поэтому дополним допущение 1 еще одним предположением.

Допущение 2. Случайные величины ε_t, $t=1,2,\ldots$, нормально распределены.

Из допущений 1, 2 следует, что величины ε_t, $t=1,2,\ldots$, имеют одинаковые нормальные распределения и независимы. Для такого случая критерий Акаике имеет вид

$$AIC(k)=2(K(k)-L(k))/T, \tag{16}$$

где $K(k)$ — число оцениваемых параметров регрессии, $L(k)$ — логарифм функции правдоподобия для нормального распределения шума в (1), причем $K(k)=n(k+1)-M$, $L(k)=-(T/2)(1+\ln 2\pi+\ln(S(\hat{\boldsymbol{\gamma}})/T))$, где $0\le M\le k$ означает число ограничений на параметры регрессии в (8), $\hat{\boldsymbol{\gamma}}$ — решение задачи (4), (8) для заданного k и фиксированных точек переключения.

Вычисления начинаются с $k=k_0$, где $k_0\ge 1$ — минимальное число переключений, известное априори. Затем k последовательно увеличивается на единицу.

При этом сумма квадратов остатков уменьшается до минимума, которому соответствует максимальное число точек переключения, определяемое величиной T и ограничением (7) на расстояние между этими точками. Согласно (16) этот критерий может монотонно уменьшаться, а затем увеличиваться ввиду роста числа оцениваемых параметров. Другой вариант поведения $AIC(k)$ — его монотонное уменьшение. В случае последующего увеличения критерия искомая оценка числа переключений $\hat{k}$ равна $\arg\min_k AIC(k)$. При монотонном уменьшении $AIC(k)$ предлагается на каждом шаге при фиксированном k последовательно для всех точек переключения проверять возможность их исключения. Например, на втором шаге $(k=2)$ проверяется возможность исключения двух точек. Если на k-м шаге нельзя исключить ни одной точки, принимается $\hat{k} \geq k$ и k увеличивается на единицу.

Пусть некоторому k соответствует малое уменьшение информационного критерия, например на 5 % величины $AIC(k-1)$. Если из k точек можно исключить одну, тогда полагаем $\hat{k} = k-1$. Если ни одной точки исключить нельзя, k увеличивается на единицу и процедура оценивания продолжается. В случае, когда на k-м шаге можно исключить более одной точки, необходимо проверить полученные решения задачи оценивания на $(k-1)$-м и k-м шагах. Причиной такой ситуации может быть непопадание в глобальный минимум задачи оценивания на обоих шагах.

Рассмотрим подробнее проверку гипотезы об исключении некоторой точки переключения из k оцененных точек. Пусть для некоторого числа k получено решение задачи (6), (7), (9); $\hat{\boldsymbol{\alpha}} \in \Re^{n(k+1)}$, $\hat{\boldsymbol{\beta}} \in \Re^m$, $\hat{\boldsymbol{\tau}} \in \Re^k$ — оценки соответственно $\boldsymbol{\alpha}^0$, $\boldsymbol{\beta}^0$, $\boldsymbol{\tau}^0$. Вектор $\hat{\boldsymbol{\alpha}}$ состоит из $k+1$ подвектора:

$$\hat{\boldsymbol{\alpha}}_{ti} = \hat{\boldsymbol{\alpha}}_i = \text{const для } t \in I_i = [\hat{t}_{i-1}+1, \hat{t}_i], \ i=1,\ldots,k+1, \ \hat{t}_0 = 0, \ \hat{t}_{k+1} = T. \quad (17)$$

Размерность вектора $\hat{\boldsymbol{\alpha}}_i \in \Re^{\hat{t}_i - \hat{t}_{i-1}}$, т.е. длина интервала I_i в (17), зависит от числа предполагаемых точек переключения и их расположения. Интервал I_j может включать в себя полностью или частично некоторые интервалы I_j^0, $j=1,\ldots,k+1$, определенные в (2). Обозначим $\Omega_i = \{j : I_j^0 \bigcap I_i, \ j=1,\ldots,k+1\}$. Таким образом, Ω_i — множество индексов интервалов I_j^0, $j=1,\ldots,k+1$, которые пересекаются с интервалом I_i.

Рассмотрим регрессию вида

$$y_t = \mathbf{x}'_t \mathbf{a}_i^0 + \mathbf{z}'_t \boldsymbol{\beta}^0 + \xi_t, \ i=1,\ldots,k+1, \ t=1,2,\ldots,T. \quad (18)$$

Для $i=1,\ldots,k+1$ вектор $\mathbf{a}_i^0$ считается постоянным на интервале, определенном в (17), концы которого фиксированы и определяются выборочными оценками точек переключения: $\hat{t}_i = t_i^* = \text{const}$, $i=1,\ldots,k$.

Регрессия (18) с фиксированными k точками переключения аппроксимирует регрессию (1) заменой вектором $\mathbf{a}_i^0 = \text{const}$ на интервале I_i ступенчатой функции $\boldsymbol{\alpha}_{tj}^0$, $t \in I_i$, $j \in \Omega_i$, где $\boldsymbol{\alpha}_{tj}^0 = \text{const}$, $t \in I_j^0$, $j=1,\ldots,k+1$ согласно формуле (2). Если $k=k^0$ и оценки точек переключений совпадают с их истинными величинами, тогда $\mathbf{a}_i^0 = \boldsymbol{\alpha}_i^0$, что означает совпадение $\mathbf{x}'_t \mathbf{a}_i^0 + \mathbf{z}'_t \boldsymbol{\beta}^0$ и $\mathbf{x}'_t \boldsymbol{\alpha}_i^0 + \mathbf{z}'_t \boldsymbol{\beta}^0$ на интервале $I_i = I_i^0$, $i=1,\ldots,k^0+1$. Поэтому чем k ближе к k^0, тем точнее аппроксимация.

Из (18) имеем $y_t = \mathbf{x}'_t \boldsymbol{\alpha}_{tj}^0 + \mathbf{z}'_t \boldsymbol{\beta}^0 + \delta_t + \xi_t$, $j \in \Omega_i$, $i=1,\ldots,k+1$, $t=1,2,\ldots,T$, где $\delta_t = \mathbf{x}'_t \mathbf{a}_i^0 - \mathbf{x}'_t \boldsymbol{\alpha}_{tj}^0$, $t=1,2,\ldots,T$. Следовательно, $\xi_t = \varepsilon_t - \delta_t$, $t=1,2,\ldots,T$. Если δ_t достаточно мало, то свойства ε_t, определенные допущениями 1, 2, будут справедливы и для ξ_t, $t=1,2,\ldots,T$.

Обозначим $\mathbf{a}^0 = [(\mathbf{a}_1^0)' \ \ldots \ (\mathbf{a}_{k+1}^0)']'$ вектор истинных параметров регрессии (18), изменяющихся в точках переключения. Оценки параметров регрессии (18) $\mathbf{a}^0$ и $\boldsymbol{\beta}^0$, обозначив их соответственно $\hat{\mathbf{a}} \in \Re^{n(k+1)}$ и $\hat{\mathbf{b}} \in \Re^m$, найдем при условии непрерывности функции $E(t)$ в (3), решив задачу

$$S(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{k+1} \sum_{t=t_i^*+1}^{t_i^*} (y_t - \mathbf{x}'_t \mathbf{a}_i - \mathbf{z}'_t \mathbf{b})^2 \to \min, \ \mathbf{A}(\boldsymbol{\tau}^*) \mathbf{a} = \mathbf{O}_K, \quad (19)$$
$$\mathbf{A}(\boldsymbol{\tau}^*) = \mathbf{O}(\boldsymbol{\tau}^*) \mathbf{L},$$

где минимизация проводится по $\mathbf{a} \in \Re^{n(k+1)}$ и $\mathbf{b} \in \Re^m$, а $\boldsymbol{\tau}^* = [t_1^* \ t_2^* \ \ldots \ t_k^*]'$ — вектор выборочных оценок точек переключения.

Для рассматриваемой выборки полученные оценки совпадают с результатом решения задачи (6), (7), (9):

$$\mathbf{a}^* = \boldsymbol{\alpha}^*, \ \mathbf{b}^* = \boldsymbol{\beta}^*, \quad (20)$$

где $\mathbf{a}^*$ и $\boldsymbol{\alpha}^*$ — выборочные оценки $\mathbf{a}^0$ и $\boldsymbol{\alpha}^0$ соответственно, $\mathbf{b}^*$ и $\boldsymbol{\beta}^*$ — выборочные оценки $\boldsymbol{\beta}^0$ (для регрессий (18) и (1)).

Из (20) вытекает, что остатки, соответствующие парам выборочных оценок $(\mathbf{a}^*, \mathbf{b}^*)$ и $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, совпадают, так как зависимая и независимые переменные в этих регрессиях равны. Следовательно, свойства случайных величин ξ_t, $t = 1, 2, \ldots, T$, можно определить по выборочным остаткам в задаче (6), (7), (9).

Проверим гипотезу об исключении точки переключения t_i^* для регрессии (18), что равносильно проверке для заданного k линейной нулевой гипотезы

$$H_{0i}: \mathbf{a}_i^0 = \mathbf{a}_{i+1}^0, \ 1 \le i \le k. \quad (21)$$

Для этого вычислим статистику

$$F^* = \frac{S(\bar{\mathbf{a}}^*, \bar{\mathbf{b}}^*) - S(\mathbf{a}^*, \mathbf{b}^*)}{S(\mathbf{a}^*, \mathbf{b}^*)} \frac{T - N + k}{\nu},$$

где ν — число ограничений, которые надо добавить к задаче (19), чтобы выполнить условие (21); $\bar{\mathbf{a}}^* \in \Re^{n(k+1)}$, $\bar{\mathbf{b}}^* \in \Re^m$ — оценки соответственно $\mathbf{a}^0$ и $\boldsymbol{\beta}^0$, полученные в результате решении задачи (19) с добавленными ограничениями

$$a_{ij} = a_{i+1,j}, \ j = j_0, j_0 + 1, \ldots, n. \quad (22)$$

Здесь a_{ij}, $j = 1, \ldots, n$, — компоненты вектора $\mathbf{a}_i$ в (19).

Чтобы определить величину ν в формуле для F^* и j_0 в (22), рассмотрим i-е $(i = 1, \ldots, k)$ ограничение в (19): $\mathbf{x}'_{t_i^*} \mathbf{a}_i - \mathbf{x}'_{t_i^*} \mathbf{a}_{i+1} = 0$. Запишем его через компоненты $\mathbf{a}_i$ и $\mathbf{a}_{i+1}$:

$$x_{t_i^* 1}(a_{i1} - a_{i+1,1}) + \sum_{j=2}^{n} x_{t_i^* 1}(a_{ij} - a_{i+1,j}) = 0.$$

Пусть в (22) $j_0 = 2$. Тогда из этого равенства следует $a_{i1} = a_{i+1,1}$. Таким образом, $n-1$ ограничение (22) позволяет реализовать условие (21): n компонент векторов $\mathbf{a}_i$ и $\mathbf{a}_{i+1}$ должны быть равны. Поэтому $\nu = n - 1$ в формуле для F^*. Если не требуется непрерывность функции $E(t)$ в точках переключения, тогда $\nu = n$, $j_0 = 1$.

Используя равенства (20), имеем

$$F^* = \frac{S(\overline{\mathbf{a}}^*, \overline{\mathbf{b}}^*) - S(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\tau}^*)}{S(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\tau}^*)} \frac{T - N + k}{\nu},$$

где $S(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\tau}^*)$ — сумма квадратов выборочных остатков для задачи (6), (7), (9).

Если в результате исследования этих остатков принимается гипотеза о соответствии шума в регрессии (18) допущениям 1, 2, то согласно [8, п. 4.6] величину F^* можно рассматривать как реализацию случайной величины F, имеющей распределение Фишера с числом степеней свободы $q_1 = \nu$ и $q_2 = T - N$. Поэтому гипотеза H_{0i} принимается на $100 \cdot p$ %-м уровне, если

$$F \le F_p(q_1, q_2), \tag{23}$$

где $F_p(q_1, q_2)$ — верхняя $100 \cdot p$ %-я точка распределения Фишера.

Описанную процедуру проверки нулевой гипотезы об исключении точки переключения можно упростить, если не требуется непрерывность по t математического ожидания $E(t)$ и $\boldsymbol{\beta}^0 = \mathbf{O}_m$. В этом случае нет необходимости в ограничениях (10).

Пусть из k оцененных точек необходимо проверить гипотезу об исключении i-й точки $(1 \le i \le k)$. Нулевая гипотеза не изменяется: она определяется формулой (21). Согласно [1] при выполнении допущений 1, 2 применительно к случайной компоненте в (18) отношение

$$F = \frac{s(\hat{\mathbf{h}}_i) - s(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\tau}})}{s(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\tau}})} \frac{t_{i+1}^* - t_{i-1}^* - 2n}{n} \tag{24}$$

при фиксированных точках переключения t_{i-1}^*, t_i^*, t_{i+1}^* имеет распределение Фишера с числом степеней свободы $q_1 = n$ и $q_2 = \hat{t}_{i+1} - \hat{t}_{i-1} - 2n$.

В (24) $s(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\tau}})$ — сумма квадратов отклонений от двух регрессий слева и справа от точки t_i^*, определенная в результате решения задачи (6), (7), (9) при $\boldsymbol{\beta}^0 = \mathbf{O}_m$,

$$s(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\tau}}) = \sum_{t = t_{i-1}^* + 1}^{t_i^*} (y_t - \mathbf{x}'_t \hat{\boldsymbol{\alpha}}_i)^2 + \sum_{t = t_i^* + 1}^{t_{i+1}^*} (y_t - \mathbf{x}'_t \hat{\boldsymbol{\alpha}}_{i+1})^2;$$

$s(\hat{\mathbf{h}}_i)$ — сумма квадратов отклонений на интервале $[t_{i-1}^* + 1, t_{i+1}^*]$ от линейной регрессии, $s(\hat{\mathbf{h}}_i) = \sum_{t = t_{i-1}^* + 1}^{t_{i+1}^*} (y_t - \mathbf{x}'_t \hat{\mathbf{h}}_i)^2$, $\hat{\mathbf{h}}_i \in \Re^n$, $\hat{\mathbf{h}}_i$ — решение задачи оценивания $s(\mathbf{h}_i) = \sum_{t = t_{i-1}^* + 1}^{t_{i+1}^*} (y_t - \mathbf{x}'_t \mathbf{h}_i)^2 \to \min$.

Проверка гипотезы сводится к проверке выполнения неравенства (23) при $q_1 = n$ и $q_2 = t_{i+1}^* - t_{i-1}^* - 2n$ и подстановке в него величины F^*, вычисленной по формуле (24) для выборочных оценок параметров регрессии.

Описанные процедуры проверки нулевой гипотезы H_{0i} осуществляются для $i = 1, \ldots, k$ в (21). Предположим, что для a точек переключения $(a \le k)$ нулевая гипотеза отвергнута. Тогда с большой вероятностью можно предположить, что для другой выборки того же объема будет оценено не менее a точек переключения.

Чтобы убедиться в том, что в результате решения задачи (6), (7), (9) по алгоритму вычисления локального минимума достигнут глобальный минимум, необходимо начинать решение для разных начальных значений вектора точек переключения $\boldsymbol{\tau}$.

Эти точки можно сформировать следующим образом. Интервал $[1, T]$ разбивается на $k+1$ отрезок приблизительно равной длины $[a_i, b_i]$, $i=1, \ldots, k+1$, причем $a_1 = 1$, $b_{k+1} = T$, $b_{i-1} = a_i$. Считаем, что начальное значение первой точки переключения равномерно распределено на отрезке $[1, b_1]$, что позволяет найти начальную величину t_1. Исходя из того, что на отрезке $[t_1 + 1 + \theta_2, b_2]$ равномерно распределена величина t_2, находим начальное значение второй точки переключения и т.д. В результате получаем начальный вектор $\boldsymbol{\tau}$. Таким образом, можно сгенерировать необходимое количество векторов $\boldsymbol{\tau}$. Начальные величины параметров $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ выбираем произвольно.

Целесообразно также использовать информацию о расположении точек переключения, рассматривая график зависимой переменной y_t, $t = 1, 2, \ldots, T$, как функции t, откуда можно получить ограничения вида $t_i \ge c$, $d \ge t_i$, где c и d — заданные числа, и добавить ограничения к (7).

3. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Рассмотрим решение нескольких задач построения регрессии с переключениями. Оценки точек переключения и параметров регрессии получены с помощью надстройки Поиск решения табличного процессора MS Excel. Во всех описываемых задачах число наблюдений $T = 58$, число точек переключения $k^0 = 5$, распределение $\varepsilon_t \sim N(0, \sigma^2)$, $t = 1, 2, \ldots, T$. Уровень шума в регрессии (1) определялся отношением $(\bar{y} / \sigma) \cdot 100\%$, где $\bar{y}$ — среднее значение зависимой переменной на интервале наблюдения $[1, T]$, σ — среднеквадратическое отклонение шума (СКО). В рассматриваемых ниже задачах гипотеза о нормальном распределении шума в регрессии не была отвергнута. Проверка гипотезы проводилась по критерию с использованием оценок коэффициентов асимметрии и эксцесса [9]. Применение критерия подробно рассмотрено в [10].

Задача 1. Описание тренда временного ряда линейным сплайном. Для этого случая в (1) $n = 2$, $\mathbf{x}_t = [1 \;\; t]'$, $\boldsymbol{\beta}^0 = \mathbf{O}_m$, ограничения имеют вид (10). Матрица независимых переменных $\mathbf{X} = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & T \end{bmatrix}$. Ее первый столбец состоит из единиц, а второй — натуральный ряд от единицы до T. Поэтому $\mathbf{X}$ размера $T \times 2$ имеет полный ранг. Тогда согласно теореме 1 минимальное расстояние между точками переключения должно быть равным двум. Положим в (7) $\theta_i = 3$, $i = 1, \ldots, k+1$, чтобы обеспечить некоторую избыточность данных.

Рассмотрим вычисления для $\sigma = 0{,}5$. Этой величине соответствует высокий уровень шума — 30,81 %. В табл. 1 приведены информационный критерий, сумма квадратов остатков $S(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\tau}^*)$, полученная при пошаговом увеличении k в ре-

Таблица 1

Показатели	Результаты оценивания точек переключения и параметров регрессии для задачи 1					
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
$S(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\tau}^*)$	29,68	20,91	15,02	13,54	10,99	10,46
$AIC(k)$	2,271	1,956	1,659	1,590	1,416	1,400
$\dfrac{AIC(k-1) - AIC(k)}{AIC(k-1)}$	—	0,139	0,152	0,042	0,109	0,011
Критерий Дарбина–Уотсона d	0,870	1,221	1,559	1,672	2,001	2,043
d_L; d_U	1,490; 1641	1,452; 1,681	1,414; 1,724	1,374; 1,768	1,334; 1,814	1,294; 1,861

Таблица 2

k	Результаты вычисления F-критерия — статистики F_{kl}^* для задачи 1						q_2	$F_{0,05}(1, q_2)$
	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$		
4	79,89	21,48	35,91	13,23	—	—	52	4,027
5	85,46	26,16	16,50	9,42	9,19	—	51	4,030
6	86,69	28,42	14,22	4,52	12,32	10,53	50	4,034

зультате решения задачи (6), (7), (10), и критерий Дарбина–Уотсона (DW). Его критические точки d_L, d_U определены для 5 %-го уровня значимости по таблицам из [11] по ближайшему меньшему числу наблюдений 55 и числу независимых переменных $n(k+1)-k-1$ (без свободного члена).

Согласно табл. 1 уменьшение информационного критерия для $k=4,5,6$ незначительно. Причем для $k=6$ снижение практически отсутствует. Поэтому предположительно оценка числа точек переключения $\hat{k}$ равна 5. Этот вывод подтверждает критерий DW, согласно которому для $k=1,2$ имеется положительная автокорреляция шума в модели (18), числу точек переключения $k=3,4$ соответствует область неопределенности критерия, начиная с $k=5$ гипотеза об отсутствии автокорреляции не отвергается, причем $k=5$ соответствует практически идеальное значение DW, что свидетельствует об отсутствии автокорреляции.

Чтобы исключить случай $\hat{k}=3$, проверим числа точек переключения для $k=4,5,6$ в соответствии с процедурой, описанной в разд. 2. Результаты вычисления F-критерия приведены в табл. 2, где F_{kl}^* означает F-статистику, полученную для проверки гипотезы об удалении l-й точки переключения из общего их числа k. Из табл. 2 следует, что нулевая гипотеза отвергается во всех случаях, кроме $k=6$, $l=4$. Для этого случая на 4 %-м уровне принимается гипотеза о том, что четвертую точку из общего числа $k=6$ точек можно исключить. Таким образом, получаем $\hat{k} \geq 4$ (для $k=4$), $\hat{k} \geq 5$ (для $k=5$), $\hat{k} \leq 5$ (для $k=6$). Следовательно, принимается гипотеза $\hat{k}=5$, что в точности совпадает с истинной величиной k^0. Полученные оценки точек переключения даны в табл. 3. Исходные данные и сплайн приведены на рис. 1. Точность оценивания точек переключения определялась величиной $\Delta = \left(\sum_{i=1}^{k} | t_i^* - t_i^0 | / t_i^0 \right) \cdot 100\,\%$. Для $\sigma=$ 0,5 получено $\Delta = 9{,}03\,\%$, причем макси-

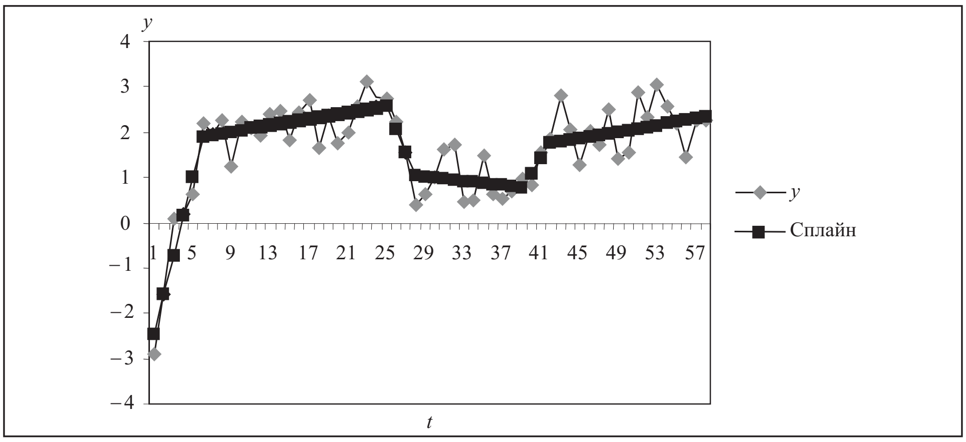

Рис. 1. Тренд в виде сплайна для $\sigma = 0{,}5$ (задача 1)

Таблица 3

Параметры	Оценка точек переключения				
	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
Истинные величины	6	22	34	38	44
Оценки для $\sigma=0{,}5$	6	26	28	39	42
Оценки для $\sigma=0{,}1$	6	23	29	38	43

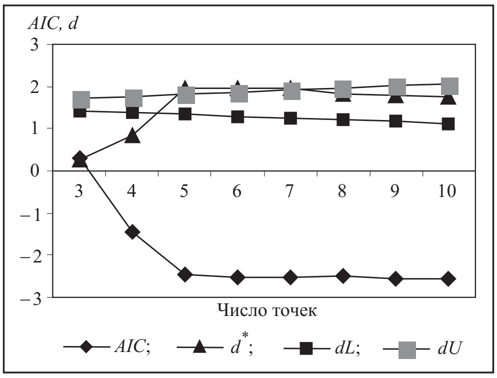

Рис. 2. Графики критериев *AIC* и *d* в зависимости от числа точек переключения для $\sigma=0{,}1$ (задача 1)

мальная ошибка оценивания точки переключения δ составляет 18,18 %. Она соответствует $i=2$.

Для определения влияния уровня шума на точность оценивания были проведены аналогичные расчеты для $\sigma=0{,}1$, чему соответствует уровень шума 6,48 %. Оценки точек переключения приведены в табл. 3. Средняя точность $\Delta=4{,}86$ %, максимальная ошибка $\delta=14{,}71$ % соответствует $i=3$. Таким образом, уменьшение уровня шума в пять раз привело к уменьшению Δ приблизительно в два раза, а максимальная ошибка δ уменьшилась на 30 %, что свидетельствует о помехоустойчивости метода оценивания.

На рис. 2 приведены графики критериев Дарбина–Уотсона d и *AIC*, где величина $d^*=d$, если $d<2$; $d^*=4-d$, если $d>2$, так как именно d^* сравнивается с табличными величинами d_L и d_U, которые даны для 5 %-го уровня значимости. Наибольшему значению d^* соответствует $k=5$, для $k>5$ d^* медленно уменьшается. Только для $k=5, 6, 7$ можно принять гипотезу об отсутствии корреляции остатков, причем 5 — минимальная величина k, для которой остатки не коррелированны. Информационный критерий резко снижается до точки $k=5$, а затем медленно убывает, это также свидетельствует, что $\hat{k}\geq 5$. Проверка всех точек переключения для $k=6$, как это выполнялось для $\sigma=0{,}5$, на 1 %-м уровне показала возможность исключения четвертой точки, что позволяет сделать вывод в пользу $\hat{k}=5$. График сплайна приведен на рис. 3. Данный пример показывает, что критерий DW может быть использован при определении нижней границы для оценки

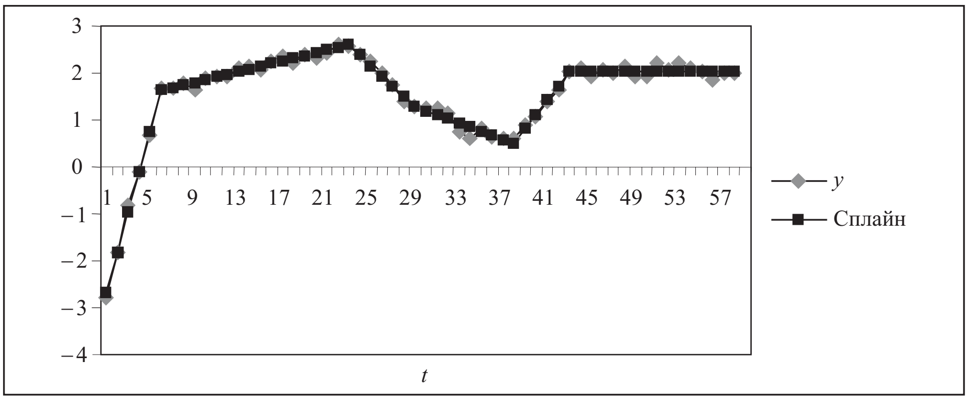

Рис. 3. График переменной y и сплайна для $\sigma=0{,}1$ (задача 1)

Таблица 4

k	Результаты вычисления F-критерия — статистики F_{kl}^{*} для задачи 2						q_2	$F_{0,05}(1, q_2)$
	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$	$l = 6$		
5	73,80	37,42	13,64	7,23	6,45	—	46	3,200
6	62,42	40,77	9,17	9,90	15,79	1,07	44	3,209

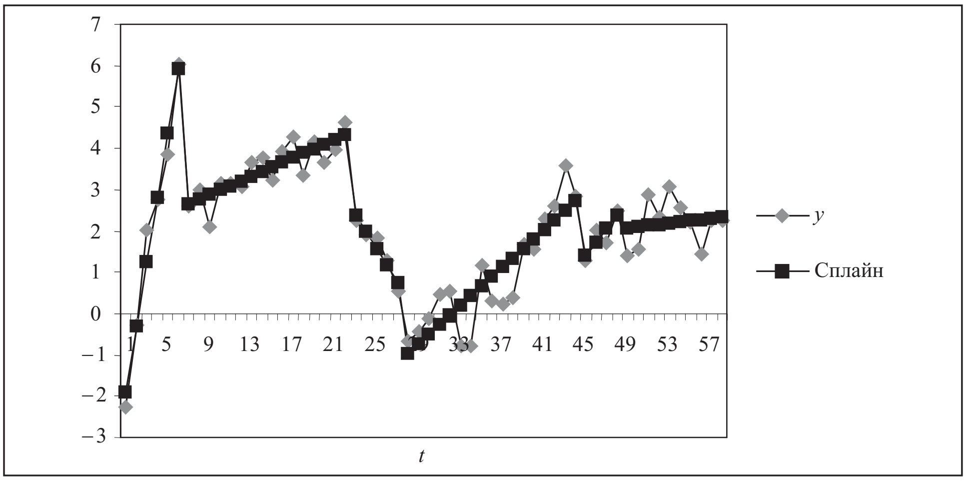

Рис. 4. График сглаживания временного ряда кусочно-линейным трендом с разрывами, $\sigma = 0{,}5$ (задача 2)

числа точек переключения $\hat{k}$, если он находится вне области неопределенности. В последующих двух задачах, в которых в точках переключения имеются разрывы, критерий DW всегда находился в области неопределенности.

Задача 2. Описание тренда временного ряда прямыми, которые в точках переключения могут иметь разрыв. Задача аналогична задаче 1, однако в этом случае не надо вводить в задачу оценивания ограничения (10). Уровень шума составляет 24,18 %. Были получены значения информационного критерия: $AIC(4) = 1{,}755$, $AIC(5) = 1{,}489$, $AIC(6) = 1{,}512$. Таким образом, критерий минимален для $k = 5$. Однако его возрастание незначительно для $k = 6$ (на 1,5 %). Поэтому все полученные точки с помощью F-критерия были проверены на возможность их исключения для $k = 5, 6$ (табл. 4). Использовать остатки в данной задаче оказалось невозможным, так как для $k = 5, 6$ величина критерия DW попала в область неопределенности. Согласно табл. 4 шестая из шести точек на 5 %-м уровне значимости может быть исключена. Таким образом, $\hat{k} = 5$. Представим оценки точек переключения в задаче 2 для $\hat{k} = 5$.

Точки переключения	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
Оценки для $\sigma = 0{,}5$	6	22	27	38	44

Сопоставление их с истинными величинами в табл. 3 определяет $\Delta = 9{,}03\ \%$, $\delta = 20{,}60\ \%$ (соответствует $i = 3$). Таким образом, вычисленные оценки точек переключения близки по точности к оценкам в задаче 1 для $\sigma = 0{,}5$. Полученная регрессия приведена на рис. 4.

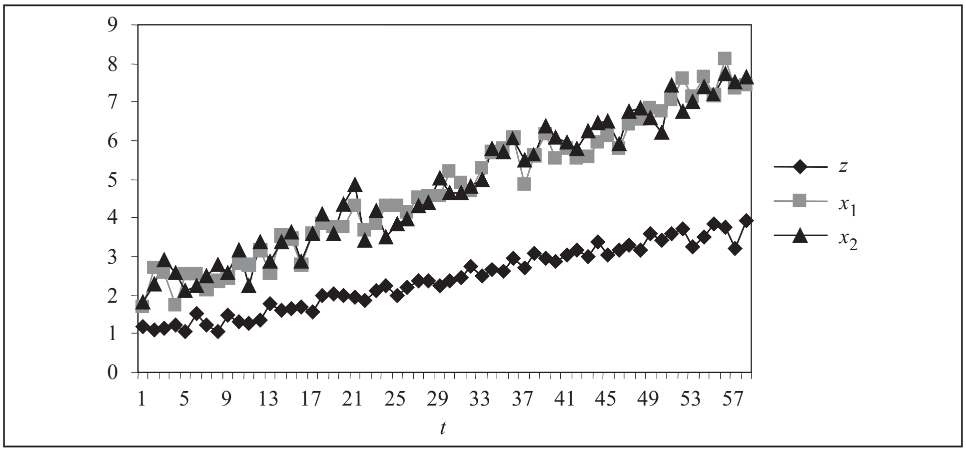

Рис. 5. Регрессоры в задаче 3

Задача 3. Построение регрессии с переключениями вида (1) для случая $n=2$, $m=1$, $\mathbf{x}_t=[x_{t1}\ x_{t2}]'$, $\mathbf{z}_t=z_t\in\Re^1$, $\boldsymbol{\beta}^0=\beta^0=7$, $\sigma=0{,}5$, что соответствует уровню шума 9,25 %. Регрессия имеет разрывы в точках переключения. Каждая независимая переменная представляет сумму линейного тренда с нормально распределенным шумом (рис. 5). Все переменные взаимно сильно коррелированны — коэффициенты корреляции равны 0,996, что свидетельствует о сильной мультиколлинеарности.

Для $k=4,5,6$ в результате двухшагового решения задачи были получены следующие величины критерия Акаике: $AIC(4)=4{,}561$, $AIC(5)=4{,}390$, $AIC(6)=4{,}406$. Отсюда следует, что $\hat{k}=5$. Для этого числа точек переключения результаты оценивания приведены в табл. 5. Последние две строки представляют решение задачи (6), (7) без априорных ограничений и с ограничениями $t_1\geq 5$, $t_2\geq 21$, $t_4\geq 31$, добавленными на основе анализа графика зависимой переменной (рис. 6). Эти ограничения согласно табл. 5 не активны. Их использование позволило получить наибольшую точность оценивания точек переключения: максимальная погрешность не превышает 12 %, средняя ошибка Δ составляет 4,86 %.

Однако наименьшее значение функции цели было достигнуто с помощью двухшагового оценивания без априорных ограничений, для него $\delta=23{,}53$ %, $\Delta=7{,}64$ %. Результат оценивания представлен на рис. 6. Его преимуществом является получение удовлетворительной средней ошибки без привлечения априорной информации, эффективность использования которой на практике трудно определить.

Таблица 5

Параметры	Результаты двухшагового оценивания точек переключения					Сумма квадратов остатков
	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	
Истинные величины	6	22	34	38	44	—
Первый шаг решения	6	22	26	38	45	271,451
Второй шаг решения	6	22	26	38	47	180,996
Решение за один шаг без ограничений	4	10	22	38	50	191,377
Решение за один шаг с ограничениями	6	22	35	38	50	186,579

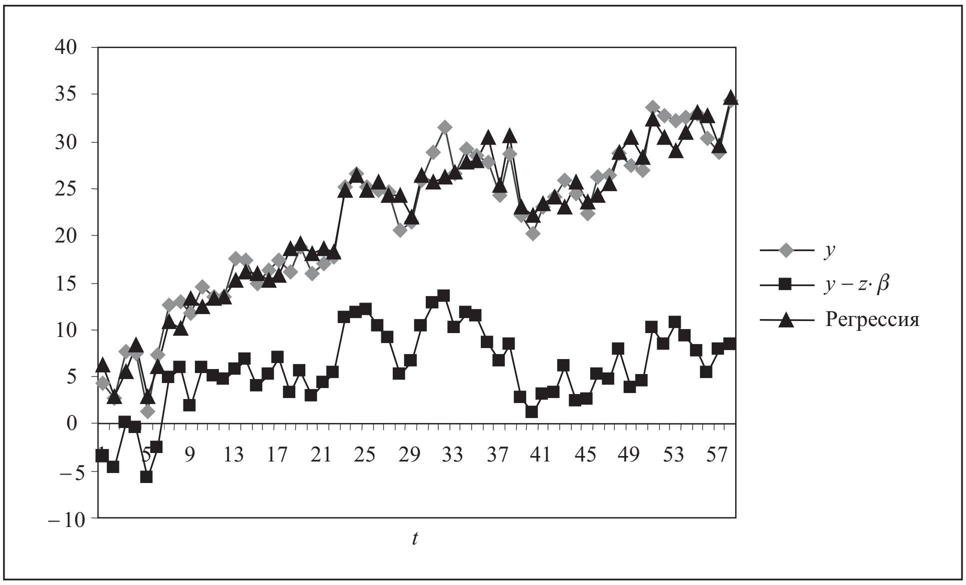

Рис. 6. Построение регрессии в задаче 3

Полученная двухшаговым методом оценка $\hat{\beta} = 6{,}585$ отличается от β^0 на 5,93 %. Погрешность при оценивании β^0 другими методами из табл. 5 ненамного больше. Однако погрешность при оценивании переключающихся параметров ввиду мультиколлинеарности оказалась значительно больше — пять оценок имели знаки, противоположные знакам истинных параметров. Результаты двухшагового оценивания, когда имеются переключающиеся и постоянный параметры (задача 3), даны в табл. 5.

Для получения оценок параметров с требуемыми знаками использовалась гребневая регрессия, для чего решалась задача оценивания $S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}) + r\,||\,\boldsymbol{\alpha}\,||^2 \to \min$, $r \geq 0$, с ограничением (7), где функция $S(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau})$ определена в (6), варьируемые переменные — $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\tau}$. Подбором r был изменен знак оценки одного параметра. При этом точки переключения соответствуют точкам в табл. 5. Данный расчет показал возможность получения различных обобщений предложенной задачи оценивания.

ЗАКЛЮЧЕНИЕ

Предложенный метод оценивания точек переключения и параметоров линейной регрессии позволяет учитывать требования к регрессии, вводить дополнительные ограничения, изменять функцию цели, чтобы получить регрессию с заданными свойствами. Приведенные примеры разной сложности иллюстрируют различные применения предлагаемого метода.

СПИСОК ЛИТЕРАТУРЫ

1. Розин Б.Б., Котюков В.И., Ягольницер М.А. Экономико-статистические модели с переменной структурой. Новосибирск: Наука, 1984. 242 с.

2. Yao Yi-Ching. Approximating the distribution of the ML estimate of the change-point in a sequence of independent R.V.'s. *Annals of Statistics*. 1987. Vol. 3. P. 1321–1328.

3. Bai J. Estimation of a change point in multiple regression models. *The Review of Economics and Statistics.* 1977. Vol. 79, Iss. 4. P. 551–563.

4. Bai J., Perron P. Estimating and testing linear models with multiple structural changes. *Econometrica*. 1998. Vol. 66, N 1. P. 47–78.
5. Grossmann I.E., Sahinidis N.V. (Eds.) Special Issues on Mixed-integer Programming and Its Applications to Engineering. Part I; II. *Optimization and Engineering*. 2002. Vol. 3, N 3; 2003. Vol. 4, N 1–2.
6. Chang Y., Sahinidis N.V. Global optimization in stabilizing controller design. *Journal of Global Optimization.* 2007. Vol. 38, N 4. P. 509–526.
7. Поляк Б.Т. Введение в оптимизацию. Москва: Наука. 1983. 384 с.
8. Себер Дж. Линейный регрессионный анализ. Москва: Мир, 1980. 456 с.
9. Кокс Д., Хинкли Д. Теоретическая статистика. Москва: Мир, 1978. 560 с.
10. Корхин А.С., Минакова Е.П. Компьютерная статистика. Ч. 1. Днепропетровск: Національний гірничий університет, 2008. 150 с.
11. Savin N.E., White K.J. The Durbin–Watson test for serial correlation with extreme sample sizes or many regressors. *Econometrica*. 1977. Vol. 45, N 8. P. 1989–1996.

Надійшла до редакції 07.07.2017

А.С. Корхін

ПРО ПОБУДОВУ РЕГРЕСІЇ З ПЕРЕМИКАННЯМИ, ЯКЩО ТОЧКИ ПЕРЕМИКАННЯ НЕВІДОМІ

Анотація. Розглянуто регресію з перемиканнями, якщо точки перемикання невідомі. Описано загальний метод оцінювання точок перемикання і параметрів лінійної регресії з перемиканнями загального вигляду. Наведено приклади його використання.

Ключові слова: регресія, перемикання, параметри регресії, оцінювання.

A.S. Korkhin

ON CONSTRUCTING A SWITCHING REGRESSION WITH UNKNOWN SWITCHING POINTS

Abstract. Regression with switching is considered. Switching points are unknown. A general method is described for estimating switching points and parameters of linear regression with switching. Examples of its use are given.

Keywords: regression, switching, regression parameters, evaluation.

Корхин Арнольд Самуилович,
доктор физ.-мат. наук, профессор кафедры Национального горного университета, Днепр,
e-mail: a.s.korkhin@gmail.com.