ОЦЕНИВАНИЕ ПАРАМЕТРОВ НЕЛИНЕЙНОЙ МОДЕЛИ КВАНТИЛЬНОЙ РЕГРЕССИИ ЗНАКОВЫМ МЕТОДОМ

П. Ф. Тарасенко, А. В. Журавлев

Томский государственный университет, ОПИ ТНЦ СО РАН
Томск, Россия
E-mail: ptara@ich.tsu.tomsk.su, avzhur@ich.tsu.tomsk.su

Рассматривается подход к оцениванию параметров нелинейной квантильной регрессии, основанный на знаковом методе. Процедура знакового оценивания синтезируется на основе принципов максимума отношения правдоподобия и максимума достигнутого уровня значимости. Сформулированы условия асимптотической нормальности знаковых оценок, предложен алгоритм их вычисления. Рассмотрен пример оценивания параметров модели диэлектрической проницаемости вещества.

Ключевые слова: квантильная регрессия, нелинейная модель, знаковый метод, статистическое оценивание параметров.

ВВЕДЕНИЕ, ПОСТАНОВКА ЗАДАЧИ

Пусть $F(y\,|\,\mathbf{X})$ – функция условного распределения с. в. Y при заданном векторе регрессоров $\mathbf{X}$. Квантильной регрессией уровня p называют [1] функцию

$$Q_p(\mathbf{X}) = F^{-1}(p\,|\,\mathbf{X}) = \inf\{y\colon F(y\,|\,\mathbf{X}) \geq p\}.$$

Если функция квантильной регрессии известна с точностью до m параметров $Q_p(\mathbf{X}) = y(\mathbf{X}\,|\,\boldsymbol{\theta})$ и ее требуется восстановить по независимым наблюдениям с. в. Y при некоторых неслучайных значениях $\mathbf{X}_1, \ldots, \mathbf{X}_n$, то мы приходим к модели наблюдений

$$Y_i = y_i(\boldsymbol{\theta}) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

где ε_i –независимые случайные величины, о ф. р. F_i которых известно, что $F_i(0) = p$, $f_i(F_i^{-1}(p)) > 0$ и для большей общности положено $y_i(\boldsymbol{\theta}) = y(\mathbf{X}_i\,|\,\boldsymbol{\theta})$.

Как метод статистического анализа данных квантильная регрессия известна с 1978 года, когда вышла в свет основополагающая работа [1]. Библиография по этому вопросу очень обширна, и интерес к приложениям со временем только усиливается. Причиной этому служат два обстоятельства. Во-первых, получив оценки для различных уровней p, можно выносить суждения об изменении условного распределения при изменении регрессоров (например, об изменении интерквантильного размаха). Во-вторых, даже при оценивании медианной регрессии для симметричных распределений (когда обычная регрессия совпадает с квантильной) оценки квантильной регрессии (основанные на модулях остатков) более устойчивы к симметричным выбросам (затягиванию хвостов распределения с. в. ε_i), чем оценки МНК.

Традиционно методы оценивания параметров квантильной регрессии основываются на минимизации суммы взвешенных модулей остатков. В то же время для частного случая линейной модели медианной регрессии разработаны методы знакового анализа [2], которые обладают той же асимптотической эффективностью, что и оценки наименьших модулей, но более устойчивы к несимметричному засорению выборки и не требуют одинаковой распределенности с. в. ε_i.

В предлагаемой работе идея знакового анализа распространяется на случай произвольного уровня p и на случай нелинейной регрессии. Затем мы конструируем алгоритм поиска оценок и демонстрируем результаты, полученные для одного практического примера.

Мы собираемся построить процедуру оценивания параметров $\boldsymbol{\theta}$ модели (1) на основе анализа не самих невязок, а их знаков $s_i(\boldsymbol{\theta}_0)=\mathrm{sign}[Y_i-y_i(\boldsymbol{\theta}_0)]$, где $\mathrm{sign}[u]=\{-1, u<0;\ 0, u=0;\ 1, u>0\}$. Для этого используем подход, впервые предложенный в [3] для рангового анализа, а в [2] – для построения знаковых оценок: сначала синтезируем тест для проверки простой гипотезы о параметрах, а затем применим принцип максимального достигнутого уровня значимости для получения оценок параметров.

ПРОВЕРКА ГИПОТЕЗ О ПАРАМЕТРАХ

Рассмотрим задачу проверки простой гипотезы H_0 против сложной альтернативы H_1 вида

$$H_0:\ \boldsymbol{\theta}=\boldsymbol{\theta}_0,\quad H_1:\ \boldsymbol{\theta}\neq\boldsymbol{\theta}_0. \tag{2}$$

Статистическую проверку гипотез (2) будем строить на основе знаков $s_i=s_i(\boldsymbol{\theta}_0)$ с использованием того факта, что при гипотезе имеет место $P\{s_i=-1\}=1-P\{s_i=1\}=p$. Пространство всех наборов признаков $\mathbf{s}=(s_1,\ldots,s_n)'$ состоит из 3^n элементов (включая априорно невероятные наборы, содержащие нули).

Самым простым способом построить тест для проверки гипотез (2) является применение принципа максимума отношения правдоподобия $L(\mathbf{s}\mid\boldsymbol{\theta},\boldsymbol{\theta}_0)=$ $=P(\mathbf{s}\mid\boldsymbol{\theta})\,/\,P(\mathbf{s}\mid\boldsymbol{\theta}_0)$. В критическую область включаются те наборы $\mathbf{s}$, для которых норма градиента отношения правдоподобия в гипотетической точке принимает наибольшие значения: $\|\nabla_{\boldsymbol{\theta}}L(\mathbf{s}\mid\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)\|>\mathrm{const}$. Малая норма градиента косвенно указывает на то, что экстремум отношения правдоподобия находится «недалеко» от гипотетических параметров. Норму и порог в таком тесте следует выбирать так, чтобы он обеспечивал заданный уровень значимости, по крайней мере в асимптотическом смысле.

Рассмотрим правдоподобие набора признаков $P(\mathbf{s}\mid\boldsymbol{\theta})=\prod_{i=1}^{n}P(s_i\mid\boldsymbol{\theta})$. При этом для непрерывных ф. р. F_i имеем:

$$P(s_i=-1\mid\boldsymbol{\theta})=F_i(y_i(\boldsymbol{\theta}_0)-y_i(\boldsymbol{\theta})),\quad P(s_i=0\mid\boldsymbol{\theta})=0,\quad P(s_i=1\mid\boldsymbol{\theta})=1-F_i(y_i(\boldsymbol{\theta}_0)-y_i(\boldsymbol{\theta})),$$

поэтому при гипотезе для дифференцируемых функций F_i и y_i получаем:

$$\nabla_{\boldsymbol{\theta}}P(s_i=-1\mid\boldsymbol{\theta}_0)=-\nabla y_i(\boldsymbol{\theta}_0)f_i(0),\ \ \nabla_{\boldsymbol{\theta}}P(s_i=0\mid\boldsymbol{\theta})=0,\ \ \nabla_{\boldsymbol{\theta}}P(s_i=1\mid\boldsymbol{\theta}_0)=\nabla y_i(\boldsymbol{\theta}_0)f_i(0),$$

$$\nabla_{\boldsymbol\theta} P(\mathbf{s}\,|\,\boldsymbol\theta_0) = P(\mathbf{s}\,|\,\boldsymbol\theta_0)\Sigma_{i=1}^{n}\nabla y_i(\boldsymbol\theta_0) f_i(0) B(s_i),$$

где $B(-1) = -1/p$, $B(0) = 0$, $B(1) = 1/(1-p)$ – взвешенные знаки. Если не принимать во внимание возможные различия значений $f_i(0)$, то мы приходим к следующему нормированному варианту теста максимального отношения правдоподобия:

$$p(1-p)\xi_n'(\boldsymbol\theta_0)\mathbf{V}_n^{-1}(\boldsymbol\theta_0)\xi_n(\boldsymbol\theta_0) > t_\alpha, \tag{3}$$

где $$\xi_n(\boldsymbol\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla y_i(\boldsymbol\theta_0)B(s_i) = \frac{1}{\sqrt{n}}\left[\sum_{\substack{i=1\\ s_i>0}}^{n}\frac{\nabla y_i(\boldsymbol\theta_0)}{1-p} - \sum_{\substack{i=1\\ s_i<0}}^{n}\frac{\nabla y_i(\boldsymbol\theta_0)}{p}\right],$$

$$\mathbf{V}_n(\boldsymbol\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\nabla y_i(\boldsymbol\theta_0)\nabla' y_i(\boldsymbol\theta_0).$$

При гипотезе $E\{\xi_n(\boldsymbol\theta_0)\} = 0$ и $E\{\xi_n(\boldsymbol\theta_0)\xi_n'(\boldsymbol\theta_0)\} = \dfrac{1}{p(1-p)}\mathbf{V}_n(\boldsymbol\theta_0)$, поэтому матрица $\mathbf{V}_n(\boldsymbol\theta_0)$ играет в (3) нормирующую роль.

Для определения порога t_α в тесте (3) необходимо знать распределение его статистики при гипотезе. Это распределение может быть указано точно, так как при гипотезе с. в. s_i независимы и принимают значения -1 и $+1$ с вероятностями p и $1-p$ соответственно. Однако вычислить это распределение аналитически не представляется возможным, поэтому его квантиль t_α уровня $1-\alpha$ придется получать методом Монте-Карло. При больших n можно воспользоваться следующей нормальной предельной аппроксимацией, которую несложно получить с помощью аппарата характеристических функций.

Утверждение 1. Пусть имеет место гипотеза, элементы векторов $\nabla y_i(\boldsymbol\theta_0)$ ограничены равномерно по i и n, для последовательности матриц $\mathbf{V}_n(\boldsymbol\theta_0)$ существует невырожденная предельная матрица $\mathbf{V}(\boldsymbol\theta_0)$. Тогда случайный вектор $\xi_n(\boldsymbol\theta_0)$ сходится по распределению к нормальному закону $N(0,\mathrm{V}(\theta_0))$, а с. в. $\xi_n'(\boldsymbol\theta_0)\mathbf{V}_n^{-1}(\boldsymbol\theta_0)\xi_n(\boldsymbol\theta_0)$ сходится к χ_m^2.

На этом мы ограничимся в изучении свойств теста (3), считая главной целью построение процедуры оценивания параметров.

ОЦЕНИВАНИЕ ПАРАМЕТРОВ

Утверждение 1 позволяет использовать для получения оценок принцип максимального достигнутого уровня значимости. Согласно ему, поскольку предельное распределение статистики теста (3) при гипотезе не зависит от гипотетических значений параметров, знаковая оценка вычисляется по правилу:

$$\boldsymbol\theta_n = \arg\min_{\boldsymbol\theta_0\in\Theta} \xi_n'(\boldsymbol\theta_0)\mathbf{V}_n^{-1}(\boldsymbol\theta_0)\xi_n(\boldsymbol\theta_0), \tag{4}$$

где Θ – априорная область изменения параметров $\boldsymbol\theta$. Асимптотическую доверительную область для параметров с уровнем доверия β можно определить в виде

$$\Theta_n(\beta) = \left\{ \boldsymbol{\theta}_0 :\ p(1-p)\xi'_n(\boldsymbol{\theta}_0)\mathbf{V}_n^{-1}(\boldsymbol{\theta}_0)\xi_n(\boldsymbol{\theta}_0) < F_H^{-1}(\beta) \right\}, \tag{5}$$

где F_H – ф. р. тестовой статистики, в качестве которой можно взять ф. р. с. в. χ^2_m.

Опуская вспомогательные результаты, приведем без доказательства теорему об асимптотической нормальности оценки (4), которая содержит достаточное условие ее состоятельности, накладываемое на проверочную функцию

$$\Psi_n(\boldsymbol{\theta}_0 \mid \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[F_i(y_i(\boldsymbol{\theta}_0) - y_i(\boldsymbol{\theta})) - p\right] \nabla' y_i(\boldsymbol{\theta}_0)\,(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) \,/\, \|\,\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|\,. \tag{6}$$

Утверждение 2. Пусть выполнены следующие условия:

а) функции $y_i(\boldsymbol{\theta}_0)$ дважды непрерывно дифференцируемы в некоторой окрестности точки истинных параметров $\boldsymbol{\theta}$ и удовлетворяют в этой окрестности равномерно по i и n условию Липшица, т. е. $\exists L_y > 0$, $\delta_y > 0$, для которых при всех $i \geq 1$ и $\|\mathbf{w}_1 - \mathbf{w}_2\| < \delta_y$ выполняется $|\,y_i(\boldsymbol{\theta} + \mathbf{w}_1) - y_i(\boldsymbol{\theta} + \mathbf{w}_2)\,| \leq L_y \|\mathbf{w}_1 - \mathbf{w}_2\|$;

б) случайные величины ε_i независимы, их ф. р. F_i непрерывны и в окрестности нуля удовлетворяют равномерно по i и n условию Липшица, т. е. $\exists L_F > 0$, $\delta_F > 0$, для которых при всех $i \geq 1$ и $|u_1 - u_2| < \delta_F$ выполняется $|\,F_i(u_1) - F_i(u_2)\,| \leq L_F |u_1 - u_2|$;

в) в окрестности нуля ф. р. F_i имеют производные f_i, причем $\exists H_f > 0$ $0 < f_i(0) < H_f$ равномерно по i и n, а также $\exists L_f > 0$, $\delta_f > 0$, для которых при всех $i \geq 1$ и $|u_1 - u_2| < \delta_f$ выполняется $|\,f_i(u_1) - f_i(u_2)\,| \leq L_f |u_1 - u_2|$;

г) для последовательностей матриц

$$\mathbf{V}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\nabla y_i(\boldsymbol{\theta}_0)\nabla' y_i(\boldsymbol{\theta}_0) \quad \text{и} \quad \boldsymbol{\psi}_n = \frac{1}{n}\sum_{i=1}^{n} f_i(0)\nabla y_i(\boldsymbol{\theta})\nabla' y_i(\boldsymbol{\theta})$$

существуют невырожденные предельные матрицы $\mathbf{V}(\boldsymbol{\theta}_0)$ и ψ соответственно, а собственные значения матриц $\mathbf{V}(\boldsymbol{\theta}_0)$ ограничены сверху равномерно по $\theta_0 \in \Theta$;

д) элементы векторов $\nabla y_i(\boldsymbol{\theta})$ ограничены равномерно по i и n;

е) $\forall R > 0\ \exists C(R) > 0$: $\Psi_n(\boldsymbol{\theta}_0 \mid \boldsymbol{\theta}) \geq C(R)$ при больших n и всех $\boldsymbol{\theta}_0 \in \Theta$ таких, что $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\| > R$.

Тогда если $\boldsymbol{\theta}$ – внутренняя точка множества Θ, то распределение случайного вектора $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta})$ сходится к нормальному закону с нулевым математическим ожиданием и матрицей ковариаций $p(1-p)\boldsymbol{\psi}^{-1}\mathbf{V}(\boldsymbol{\psi}^{-1})'$.

Применяя это утверждение к случаю, когда $f_i(0) = f(0)$, имеем асимптотические ковариации $p(1-p)\mathbf{V}^{-1}/f^2(0)$, что совпадает с результатом, приведенным в [4] для оценок по методу наименьших взвешенных модулей. Для дальнейших сравнений целевую функцию метода модулей запишем в сопоставимой с (3) и (4) форме:

$$\arg\min_{\boldsymbol{\theta}_0 \in \Theta} \sum_{\substack{i=1 \\ s_i > 0}}^{n} \frac{|Y_i - y_i(\boldsymbol{\theta}_0)|}{1-p} - \sum_{\substack{i=1 \\ s_i < 0}}^{n} \frac{|Y_i - y_i(\boldsymbol{\theta}_0)|}{p}. \tag{7}$$

Условие е) является аналогом условия состоятельности из теории М-оценок параметров нелинейной регрессии. Действительно, если ввести обозначения $\mathbf{e}_0 = (\boldsymbol{\theta}_0 - \boldsymbol{\theta}) / \|\,\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|$, $\boldsymbol{\rho}_i(\boldsymbol{\theta}_0) = \nabla y_i(\boldsymbol{\theta}_0)B(s_i(\boldsymbol{\theta}_0))$, то (е) можно записать в виде неравенства

$$\Psi_n(\boldsymbol{\theta}_0 \mid \boldsymbol{\theta}) = \mathbf{e}'_0\ E\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{\rho}_i(\boldsymbol{\theta}_0) - \boldsymbol{\rho}_i(\boldsymbol{\theta})\right]\right\} \geq C(R) > 0,$$

которое должно выполняться за пределами любой R-окрестности точки истинных параметров $\boldsymbol{\theta}$. В [5] похожее достаточное условие получено для состоятельности оценки параметров медианной регрессии ($p = 1/2$) по методу наименьших модулей (7):

$$\frac{1}{n}\sum_{i=1}^{n} h_i(\boldsymbol{\theta}_0,\boldsymbol{\theta}) \min\left[F_i(h_i(\boldsymbol{\theta}_0,\boldsymbol{\theta})/2)-\tfrac{1}{2},\ \tfrac{1}{2}-F_i(-h_i(\boldsymbol{\theta}_0,\boldsymbol{\theta})/2)\right] \geq C(R) > 0,$$

$$h_i(\boldsymbol{\theta}_0,\boldsymbol{\theta}) = \left|y_i(\boldsymbol{\theta}_0) - y_i(\boldsymbol{\theta})\right|.$$

Для выполнения условия е) достаточно, чтобы отдельные слагаемые в (6) были отделимы от нуля за пределами окрестности точки $\boldsymbol{\theta}$ или чтобы такая отделимость выполнялась для сумм пар, троек и т. д. слагаемых. Это позволяет проверять данное условие численным методом для $\boldsymbol{\theta}_0 = \boldsymbol{\theta} + r\mathbf{e}$, строя графики зависимостей $\Phi_i(r \mid \boldsymbol{\theta},\mathbf{e}) = \mathbf{e}'\,\nabla y_i(\boldsymbol{\theta} + r\mathbf{e})\,[F_i(y_i(\boldsymbol{\theta} + r\mathbf{e}) - y_i(\boldsymbol{\theta})) - p]$ при конкретных распределениях F_i (из достаточно широкого с точки зрения практических применений класса). Следует заметить, что для линейных моделей, когда $y_i(\boldsymbol{\theta}) = \mathbf{X}_i'\boldsymbol{\theta}$, условие е) является следствием условия в) и выполняется [6] для всех распределений, у которых $f_i(0) > 0$.

АЛГОРИТМ ПОИСКА ОЦЕНОК

Задача поиска оценки (4) является достаточно сложной проблемой. Характер целевой функции таков, что она совершает скачки на гиперповерхностях смены знаков $y_i(\boldsymbol{\theta}_0) = Y_i$. Кроме того, внутри областей своей непрерывности она не постоянна. При умеренных объемах выборки это приводит к наличию локальных минимумов, мешающих поиску оптимального решения, а точное решение задачи (4) предполагает решение большого числа нелинейных задач оптимизации с ограничениями.

Для смягчения этих проблем предлагается свести задачу минимизации разрывной функции к последовательности задач минимизации непрерывных функций за счет постепенно убывающего сглаживания. Кроме этого, сглаживание позволяет устранить слабо выраженные локальные минимумы на начальных этапах движения к оптимальной точке. Сглаживание целевой функции проведем за счет замены взвешенных знаков $B(s_i)$ на функции

$$\widetilde{B}(Y_i - y_i(\boldsymbol{\theta}_0) \mid h) = \frac{G(G^{-1}(1-p) + (Y_i - y_i(\boldsymbol{\theta}_0))/h)}{p(1-p)} - \frac{1}{p},$$

где $G(u)$ имеет свойства функции распределения, h – параметр сглаживания. В результате статистика $\xi_n(\boldsymbol{\theta}_0)$ заменяется функцией

$$\widetilde{\xi}_n(\boldsymbol{\theta}_0 \mid h) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla y_i(\boldsymbol{\theta}_0)\widetilde{B}(Y_i - y_i(\boldsymbol{\theta}_0) \mid h),$$

а задача оптимизации (4) со сглаженной целевой функцией принимает вид

$$\boldsymbol{\theta}_n(h) = \arg\min_{\boldsymbol{\theta}_0 \in \Theta} \widetilde{\xi}_n'(\boldsymbol{\theta}_0 \mid h)\mathbf{V}_n^{-1}(\boldsymbol{\theta}_0)\widetilde{\xi}_n(\boldsymbol{\theta}_0 \mid h). \qquad (8)$$

При $h \to 0$ в каждой точке $\boldsymbol{\theta}_0$ целевая функция задачи (8) сходится к целевой функции из (4), поэтому $\lim_{h\to 0}\boldsymbol{\theta}_n(h) = \boldsymbol{\theta}_n$. Основанная на этом идея состоит в том, чтобы организовать последовательность процедур оптимизации функций с убывающим сглаживанием. При этом решение, полученное на предыдущем шаге (при большем сглаживании), используется на следующем шаге в качестве начального приближения.

Для сглаживания удобно использовать легко обращаемую логистическую функцию $G(u) = 1/(1 + e^{-u})$. В любом случае симметричная, непрерывная и возрастающая от 0 до 1 сглаживающая функция $G(u)$ производит $q\cdot 100$ % своего изменения на симметричном относительно нуля интервале длины $2G^{-1}\left(\frac{1+q}{2}\right)$. Для логистической функции этот интервал равен $2\ln\frac{1+q}{1-q}$.

Основной вопрос, который остается решить, – это выбор убывающей последовательности параметров сглаживания. Получим величину параметра h, необходимую для того, чтобы сгладить скачки целевой функции. Она определяется размерами областей непрерывности целевой функции в пространстве параметров, границы которых имеют вид $y_i(\boldsymbol{\theta}_0) = Y_i$. Для определения параметра h, необходимого для сглаживания в районе некоторой точки $\boldsymbol{\theta}_0$, предложим следующий подход. Пусть $\Delta_i(\boldsymbol{\theta}_0) = Y_i - y_i(\boldsymbol{\theta}_0)$ $(i = 1, \dots, n)$ – набор невязок. Тогда интересующая нас величина сглаживания имеет вид $h(\boldsymbol{\theta}_0) = \max\{Z^-_{(L)}, Z^+_{(L)}\}/G^{-1}((1 + q)/2)$, где $L = [1 + (r/2)]$, Z^+_j $(j = 1, \dots, n^+)$ – модули положительных невязок, Z^-_j $(j = 1, \dots, n^-)$ – модули отрицательных невязок $(n^+ + n^- = n)$. Для вычисления $h(\boldsymbol{\theta}_0)$ надо селектировать L положительных и L отрицательных невязок, ближайших к нулю. Если $n^+ < L$, то полагаем $L^+ = n^+$ и $L^- = 2L - n^+$, а если $n^- < L$, то $L^- = n^-$ и $L^+ = 2L - n^-$. Здесь можно взять $q = 0{,}9$ и $r = m + 1$, а в общем случае q и r выступают как параметры настройки алгоритма. Данный подход исходит из двух эвристических предположений. Во-первых, на начальных этапах процедуры излишнее сглаживание не приносит вреда, если сохраняется общая тенденция изменения целевой функции. Второе предположение состоит в том, что в районе минимума целевой функции области непрерывности имеют в среднем меньшие размеры, чем в других местах пространства параметров (последствия унимодальности распределения шумов). В то же время предложенный способ прямо не рассчитан на то, чтобы сглаживать локальные минимумы, появляющиеся за счет минимумов целевой функции (4) внутри областей непрерывности.

Таким образом, предлагаемый алгоритм поиска индикаторной оценки состоит из следующих этапов:

1. Определяется приемлемое начальное приближение $\boldsymbol{\theta}_0$. Это удобно делать на основе оценки $\boldsymbol{\theta}_M$, полученной по методу наименьших модулей (7) или наименьших квадратов. При наличии в модели (1) свободного параметра и при $p\neq 1/2$ оценку МНК лучше скорректировать: $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_M - \mathbf{e}_0\Delta_{([np])}(\boldsymbol{\theta}_M)$, где вектор $\mathbf{e}_0 = \|\delta_{kj}: j = 1, \dots, m\|$, k – номер свободного параметра, $\Delta_{([np])}(\boldsymbol{\theta}_M)$ – порядковая статистика для ряда величин $\Delta_i(\boldsymbol{\theta}_M) = Y_i - y_i(\boldsymbol{\theta}_M)$.

2. На j-м шаге, имея начальное приближение $\boldsymbol{\theta}_{j-1}$ и $h_j = R_{j-1}\cdot h(\boldsymbol{\theta}_{j-1})$, получаем $\boldsymbol{\theta}_j$ как решение $\boldsymbol{\theta}_n(h_j)$ задачи оптимизации сглаженной целевой функции (8), где $R_j = R_0/Q^j$.

3. Повторяем предыдущий шаг, если $(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1})'\mathbf{V}_n(\boldsymbol{\theta}_j)(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}) > \varepsilon_0^2$. Это условие опирается на результат об асимптотической нормальности оценок параметров и призвано скомпенсировать различия в их шкалах.

Данный алгоритм имеет несколько параметров, которые могут использоваться для его настройки: R_0 – начальный масштаб сглаживания (рекомендуемое значение

R_0=1, но при наличии локальных минимумов можно его увеличить); Q – делитель сглаживания (рекомендуемое значение Q=10); ε_0 – точность поиска приближения; q – масштаб сглаживающей функции (рекомендуемое значение q=0,9); r – глубина просмотра соседних границ смены знаков (рекомендуемое значение $r \geq m+1$, где m – число параметров). Кроме того, параметры имеются у алгоритма оптимизации, который используется на этапе 2. Практически у всех алгоритмов есть два параметра – величина начального шага (рекомендуемое значение должно быть сравнимо с h_j) и величина минимального шага (рекомендуемое значение должно быть меньше ε_0 при том же способе вычисления величины последнего шага, что и на этапе 3).

В следующем разделе мы иллюстрируем одно из применений предлагаемого подхода на примере обработки данных физического эксперимента.

ОЦЕНИВАНИЕ ПАРАМЕТРОВ МОДЕЛИ ДИЭЛЕКТРИЧЕСКОЙ ПРОНИЦАЕМОСТИ ВЕЩЕСТВА

Модель Гаврильяка – Негами для спектров комплексной диэлектрической проницаемости вещества в традиционных обозначениях имеет вид

$$\varepsilon^*(\omega,\boldsymbol{\theta}) = \varepsilon'(\omega,\boldsymbol{\theta}) - i\varepsilon''(\omega,\boldsymbol{\theta}) = \varepsilon_\infty + (\varepsilon_0 - \varepsilon_\infty)\left(1 + (i\omega\tau)^{1-\alpha}\right)^{-\beta}, \tag{9}$$

где i – мнимая единица, параметры $\boldsymbol{\theta}' = (\alpha, \beta, \tau, \varepsilon_0, \varepsilon_\infty)$ содержат статическую ε_0 и оптическую ε_∞ константы, время релаксации τ, а также параметры α и β, которые определяют преобладающий механизм релаксации. При $\alpha = 0$, $\beta = 1$ имеем модель Дебая с одним временем релаксации; при $\beta = 1$, $0 < \alpha < 1$ – модель Коул-Коула с симметричным распределением времен релаксации; при $\alpha = 0$ и $0 < \beta < 1$ – модель Коул – Давидсона с асимметричным распределением времен релаксации [7].

Для определения параметров модели (9) проводится серия экспериментов на разных частотах $\omega_1, \ldots, \omega_n$ приложенного к веществу электромагнитного поля. Их результатами являются значения $\varepsilon'_1, \ldots, \varepsilon'_n$ и $\varepsilon''_1, \ldots, \varepsilon''_n$ действительной и мнимой частей комплексной диэлектрической проницаемости, определяемые с некоторой погрешностью. Распределения этих погрешностей плохо изучены, но предположения об их независимости и нулевой медиане представляются достаточно естественными. В результате мы приходим к следующей нелинейной модели двумерной регрессии

$$\varepsilon'_i = \varepsilon_\infty + r_i^{-\beta/2}(\varepsilon_0 - \varepsilon_\infty)\cos\beta\varphi_i + \varepsilon_{1i},$$

$$\varepsilon''_i = r_i^{-\beta/2}(\varepsilon_0 - \varepsilon_\infty)\sin\beta\varphi_i + \varepsilon_{2i},$$

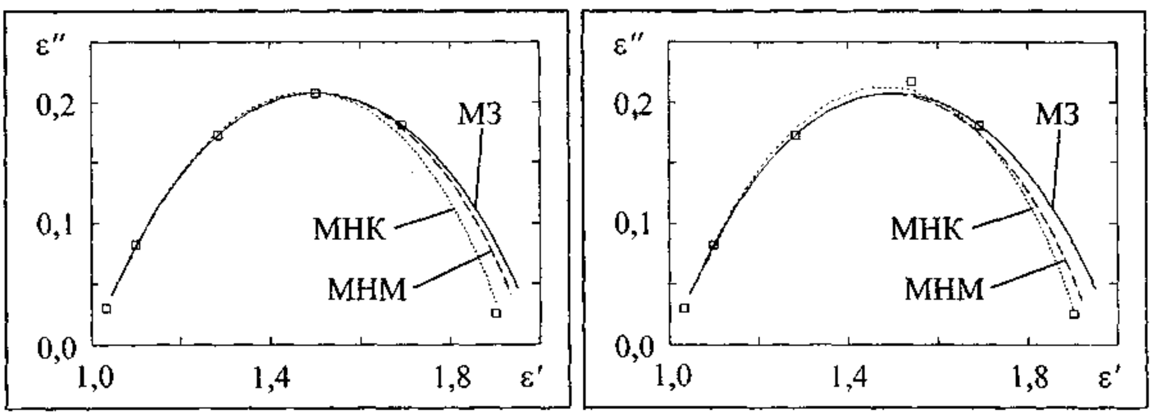
где $r_i^2 = r_{1i}^2 + r_{2i}^2$, $\varphi_i = \mathrm{arctg}(r_{1i}/r_{2i})$, $r_{1i} = 1 + (\omega_i\tau)^{1-\alpha}\sin\frac{\pi}{2}\alpha$, $r_{2i} = (\omega_i\tau)^{1-\alpha}\cos\frac{\pi}{2}\alpha$,

ε_{kj} – независимые погрешности измерений с нулевой медианой.

В [8] задача оценивания параметров этой модели решалась на основе метода наименьших квадратов. При $p = 1/2$ многомерную модель квантильной регрессии вида $Y_{kj} = y_{kj}(\boldsymbol{\theta}) + \varepsilon_{kj}$, $j = 1, \ldots, n$, $k = 1, \ldots, K$, можно свести [9] к модели (1), перенумеровав уравнения регрессии индексом $i = K(j-1) + k$. Это позволяет применить процедуру (4) для оценивания параметров модели (9).

Проверка условия состоятельности е) выполнялась для суммы пар слагаемых из (6) для разных распределений с различными масштабами в диапазонах частот ω_i, которые обычно используются для измерений. Как и следовало ожидать, на больших и малых частотах условие состоятельности (проверяемое для суммы пар слагаемых) может не выполняться, когда оценивается параметр α. Однако если в план эксперимента включено достаточное количество частот из среднего диапазона, то условие состоятельности (проверяемое для всей суммы) выполняется. Это обстоятельство может быть использовано для количественного анализа планов $\omega_1, \ldots, \omega_n$ с точки зрения состоятельности. Добавим к сказанному, что проверка проводилась на широком классе распределений, таких как распределения Стьюдента (включая Коши), нормальное, логистическое, равномерное, Лапласа, на семействе распределений Тьюки (модель симметричного нормального засорения).

Основной ожидаемый эффект от применения предложенного метода по сравнению с методом взвешенных модулей состоит в большей устойчивости получаемых оценок к несимметричному засорению наблюдений на краях плана эксперимента. В задачах регрессии это явление принято называть «эффектом рычага» (leverage). Одним из способов оценить этот эффект количественно является вычисление функции чувствительности. В данной работе мы не ставим цель теоретического или количественного изучения вопросов такой устойчивости. Однако, сравнивая (3) и (7), можно заметить, что критерий знакового метода не зависит от величины невязки, а только от ее знака, в то время как критерий модулей будет изменяться под влиянием увеличения выброса.



Графики проекций кривых модели (9), оцененные методами наименьших квадратов (МНК), модулей (МНМ) и знаковым методом (МЗ)

Указанный эффект проиллюстрирован на рисунке, где приведен модельный пример обработки результатов наблюдений, содержащих существенное отклонение от модели (9) на малой частоте (левая диаграмма) и дополнительно на средней частоте (правая диаграмма). Здесь подгонка, осуществленная знаковым методом, демонстрирует большую устойчивость к «эффекту рычага» за счет того, что критерий (4) не реагирует на величину выбросов, а опирается на те наблюдения, которые лучше соответствуют модели. В то же время следует отметить наличие локального минимума у знакового критерия в районе оценки по методу наименьших модулей.

ЛИТЕРАТУРА

1. *Koenker R., Bassett G.* Regression quantiles // Econometrica. 1978. V. 46. N. 1. P. 33–50.

2. *Болдин М. В., Симонова Г. И., Тюрин Ю. Н.* Знаковый статистический анализ линейных моделей. М.: Наука. Физматлит, 1997.

3. *Hodges J. L. Jr, Lehmann E. L.* Estimates of Location Based on Rank Tests // Ann. Math. Statist. 1963. V. 34. N. 2. P. 598–611.

4. *Chen L. A., Tran L. T., Lin L. C.* Symmetric Regression Quantile and its Application to Robust Estimation for the Nonlinear Regression Model // J. Statist. Planning and Inference. 2004. V. 126. N. 2. P. 423–440.

5. *Oberholder W.* The Consistency of Nonlinear Regression Minimizing the L_1–Norm // Ann. Statist. V. 10. N. 1. P. 316–319.

6. *Тарасенко П. Ф.* Квантильная регрессия на основе знаков невязок // Вестн. ТГУ. 2003. № 6. С. 277–281.

7. *Челидзе Т. Л., Деревянко А. И., Куриленко О. Д.* Электрическая спектроскопия гетерогенных систем. Киев: Наукова думка, 1977.

8. *Сусляев В. И., Тарасенко П. Ф., Журавлев А. В., Журавлев В. А.* Выбор модели диэлектрической релаксации вещества на основе проверки гипотез // Изв. вузов. Физика. 1999. Т. 42. № 11. С. 15–22.

9. *Koenker R. Ng. P., Portnoy S.* M-estimation of Multivariate Regressions // J. Amer. Statist. Assoc. 1990. V. 85. N. 412. P. 1060–1068.