ЗНАКОВЫЕ ПРОЦЕДУРЫ АНАЛИЗА РАСТУЩИХ СИСТЕМ

П. Ф. Тарасенко
Томский государственный университет, Россия
e-mail: ptara@ich.tsu.ru

An approach to the estimation of the parameters in the linear model of quantile regression is proposed for the case when indicators of residuals are used as analyzed attributes. Particular cases are presented for which unknown scale of noise is measured by the interquantile range. It is shown that indicator-based estimators have same asymptotic efficiency as weighted absolute deviation estimators, but they don't require equal distribution of noise.

Введение

Математическое описание растущих систем приводит к моделям регрессии, в рамках которых наблюдаемые величины (например, величины запасов углеводородов в природной совокупности месторождений) рассматриваются как линейная зависимость от приоритета накопления, наблюдаемая на фоне некоторых шумов [1]. Модель шумов имеет ключевое значение при выборе метода обработки наблюдений, поэтому должна соответствовать реальной ситуации. При работе с величинами запасов месторождений углеводородов приходится иметь дело с малыми объемами выборок, неизвестным распределением шумов, выбросами в наблюдениях, но при этом доступны априорные сведения, которые могут быть представлены в виде квантильной регрессии, когда теоретическая зависимость является не условным математическим ожиданием, а условной квантилью некоторого уровня.

В [1] рассматривалось два метода обработки наблюдений — метод наименьших квадратов и знаковый метод [2], который соответствует модели квантильной регрессии уровня $1/2$. В то же время запасы углеводородов являются оценками и чаще завышаются, чем занижаются. Поэтому в предлагаемой работе идея знакового анализа обобщается на случай квантильной регрессии произвольного уровня, а также на модели погрешностей, масштаб которых описывается интерквантильным размахом.

Широкую известность модель квантильной регрессии получила с выходом основополагающей статьи [3]. Среди множества последующих работ выделим результаты по исследованию свойств оценок параметров квантильных регрессий сразу нескольких уровней [4, 5], а также [6, 7], где методом наименьших взвешенных модулей одновременно оцениваются не одна, а две условных квантили. Традиционно методы оценивания

параметров квантильной регрессии используют принцип минимума взвешенных модулей остатков. Только для случая медианной регрессии широко известны процедуры, основанные не на модулях, а на знаках остатков [2]. При этом знаковые оценки одинаково с традиционными оценками наименьших модулей асимптотически эффективны, они обладают большей устойчивостью к выбросам и не требуют, чтобы случайные погрешности измерений были одинаково распределены, что является привлекательным при их применении для обработки величин запасов углеводородов природной совокупности месторождений как растущей системы.

1. Постановка задачи

Пусть наблюдения $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ подчиняются линейной модели

$$\mathbf{Y} = \mathbf{X}'\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{1}$$

где $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)' \in \mathbb{R}^T$ — неизвестные параметры; $\mathbf{X}$ — матрица плана, образованная столбцами $\mathbf{X}_i = (X_{i1}, \ldots, X_{iT})'$, $i = 1, \ldots, n$. Для независимых случайных величин ε_i, образующих вектор $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$, будем рассматривать три основные модели. Согласно первой из них, о функции распределения $F_i(x) = \mathbb{P}\{\varepsilon_i < x\}$ известно, что $F_i(-\mu) = 1 - F_i(\mu) = p$, где вероятность $p \in (0, 1/2)$ задана, а параметр $\mu > 0$ неизвестен. Во второй модели дополнительно известно, что $F_i(0) = 1/2$. На первую модель далее будем ссылаться как на модель с двумя квантилями, а вторую будем называть моделью с тремя квантилями. Вариант модели с двумя квантилями рассматривался ранее в [6, 7]. Третий интересующий нас случай — квантильная регрессия [3], когда о функции распределения F_i известно только, что $F_i(0) = p \in (0, 1)$. Знаковый анализ этой модели при $p = 1/2$ ранее рассматривался в [2].

Чтобы оперировать моделями с разным числом квантилей, используем следующее общее описание априорной информации о случайных погрешностях с квантилями заданных уровней, известных с точностью до параметров. Пусть при каких-то неизвестных параметрах $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)'$ выполняются условия $F_i(c_k + \mathbf{d}_k'\boldsymbol{\mu}) = q_k$, $k = 1, \ldots, K-1$, где $c_1 \leq \ldots \leq c_{K-1}$ и $0 < q_1 \leq \ldots \leq q_{K-1} < 1$ — заданные постоянные, а векторы $\mathbf{d}_k$ описывают способ параметризации квантилей. При этом все элементы вектора $\mathbf{d}_k$ равны нулю, кроме, возможно, одного, равного ± 1, так что каждый параметр из числа $\mu_1, \ldots, \mu_M$ отвечает за сдвиг одной или нескольких квантилей ($M < K-1$), а если $\mathbf{d}_k = \mathbf{0}$, то k-я квантиль не параметризована. Естественно, что если $c_k = c_{k+1}$, то $\mathbf{d}_k = \mathbf{d}_{k+1}$. Будем рассматривать только случаи, для которых $q_k = q_{k-1}$ тогда и только тогда, когда $c_k = c_{k-1}$.

Определение 1.1. При фиксированных параметрах $\boldsymbol{\mu}_0$ будем говорить об *априорном разбиении* $\mathbb{C}(\boldsymbol{\mu}_0)$, состоящем из K смежных интервалов $C_1(\boldsymbol{\mu}_0), \ldots, C_K(\boldsymbol{\mu}_0)$, разделенных границами $c_k + \mathbf{d}_k'\boldsymbol{\mu}_0$, $k = 1, \ldots, K-1$. Вопрос о том, какому из смежных интервалов — $C_k(\boldsymbol{\mu}_0)$ или $C_{k+1}(\boldsymbol{\mu}_0)$ — принадлежит граница $c_k + \mathbf{d}_k'\boldsymbol{\mu}_0$, решается в каждом конкретном случае.

Определение 1.2. *Множеством допустимых параметров априорного разбиения* будем называть множество $\mathbb{M} = \big\{\boldsymbol{\mu}_0 : c_k + \mathbf{d}_k'\boldsymbol{\mu}_0 < c_{k+1} + \mathbf{d}_{k+1}'\boldsymbol{\mu}_0 \ \forall k \in \{1, \ldots, K-2\} : c_k < c_{k+1}\big\}$.

Определение 1.3. Дискретное распределение вероятностей $p_1, \ldots, p_K$, приписанных интервалам априорного разбиения, назовем *априорным распределением* и обозна-

чим его через $\overline{p} = \{p_1, \ldots, p_K\}$. Здесь $p_k = q_k - q_{k-1}$ при дополнительных обозначениях $q_0 = 0$ и $q_K = 1$.

Определение 1.4. *Семейством априорных разбиений* будем называть параметризованный класс

$$\mathbb{C} = \{\mathbb{C}(\boldsymbol{\mu}_0) : \ \boldsymbol{\mu}_0 \in \mathbb{M}\}, \quad \mathbb{C}(\boldsymbol{\mu}_0) = \{C_1(\boldsymbol{\mu}_0), \ldots, C_K(\boldsymbol{\mu}_0)\}. \tag{2}$$

При этом, если $\boldsymbol{\mu}$ — истинные значения параметров априорного разбиения, выполняются равенства $\mathbb{P}\{\varepsilon_i \in C_k(\boldsymbol{\mu})\} = p_k$, $k = 1, \ldots, K$, $i = 1, \ldots, n$.

В частности, для модели с двумя квантилями $K = 3$, $d_1 = -1$, $d_2 = 1$ и

$$\mathbb{C} = \{\{(-\infty, -\mu_0), [-\mu_0, \mu_0], (\mu_0, \infty)\} : \ \mu_0 > 0\}, \quad \overline{p} = \{p, 1-2p, p\}.$$

Для модели с тремя квантилями $K = 5$, $d_1 = -1$, $d_2 = d_3 = 0$, $d_4 = 1$ и

$$\mathbb{C} = \{\{(-\infty, \mu_0), [-\mu_0, 0), \{0\}, (0, \mu_0], (\mu_0, \infty)\} : \ \mu_0{>}0\},$$

$$\overline{p} = \{p, (1-2p)/2, 0, (1-2p)/2, p\}.$$

Для квантильной регрессии априорное разбиение не параметризовано, $K = 3$, $d_1 = d_2 = 0$ и

$$\mathbb{C} = \{(-\infty, 0), \{0\}, (0, \infty)\}, \quad \overline{p} = \{p, 0, 1-p\}.$$

Определение 1.5. *Априорным классом* будем называть множество распределений $\mathbb{F}(\overline{p}, \mathbb{C}, \mathbb{M}) = \{\mathbb{F}_{\boldsymbol{\mu}}(\overline{p}, \mathbb{C}(\boldsymbol{\mu})) : \boldsymbol{\mu} \in \mathbb{M}\}$, где

$$\mathbb{F}_{\boldsymbol{\mu}}(\overline{p}, \mathbb{C}(\boldsymbol{\mu})) = \{F : F(c_k + \mathbf{d}_k'\boldsymbol{\mu}) = q_k \ \forall k \in \{1, \ldots, K-1\}\}.$$

Мы собираемся построить процедуру оценивания параметров $\boldsymbol{\theta}$ модели (1) и параметров $\boldsymbol{\mu}$ разбиения (2). Для этого сначала синтезируем тест для проверки простой гипотезы о параметрах, а затем применим принцип максимально достигнутого уровня значимости для получения оценок параметров. Все результаты приводятся без доказательства из-за ограниченности объема статьи.

2. Проверка гипотез о параметрах

Рассмотрим задачу проверки простой гипотезы H_0 против сложной альтернативы H_1 вида

$$H_0 : (\boldsymbol{\theta}', \boldsymbol{\mu}')' = \mathbf{0}, \quad H_1 : (\boldsymbol{\theta}', \boldsymbol{\mu}')' \neq \mathbf{0}, \tag{3}$$

где гипотетические значения всех параметров взяты нулевыми без ограничения общности. Действительно, более общую гипотезу $(\boldsymbol{\theta}', \boldsymbol{\mu}')' = (\boldsymbol{\theta}_0', \boldsymbol{\mu}_0')'$ можно свести к (3) за счет замены $\mathbf{Y}$ на $\mathbf{Y} - \mathbf{X}'\boldsymbol{\theta}_0$ и c_k на $c_k + \mathbf{d}_k'\boldsymbol{\mu}_0$.

Статистическую проверку гипотез (3) будем строить с использованием того факта, что при гипотезе $\mathbb{P}\{Y_i \in C_k(\mathbf{0})\} = p_k$. Введем функцию

$$s(x, \mathbf{u}) = \sum_{k=1}^{K} I\{x \in \bigcup_{j=1}^{k} C_j(\mathbf{u})\},$$

которая принимает значение k, если $x \in C_k(\mathbf{u})$, в качестве анализируемых признаков используем величины $s_i = s(Y_i, \mathbf{0})$, $i = 1, \ldots, n$. Пространство признаков, таким образом,

содержит K^n элементов. Обозначим его через $\mathcal{S}$. Функция мощности произвольного индикаторного теста (основанного на признаках $\mathbf{s} = (s_1, \ldots, s_n)'$) с критической областью $\mathcal{S}_1 \subset \mathcal{S}$ для проверки гипотез (3) может быть записана в виде $\mathbb{P}(\mathcal{S}_1|\boldsymbol{\theta}, \boldsymbol{\mu}) = \sum\limits_{\boldsymbol{s}\in\mathcal{S}_1} \mathbb{P}(\mathbf{s}|\boldsymbol{\theta}, \boldsymbol{\mu})$, где $\mathbb{P}(\mathbf{s}|\boldsymbol{\theta}, \boldsymbol{\mu})$ — совместное распределение признаков $\mathbf{s}$ при истинных параметрах, причем

$$\mathbb{P}(\mathbf{s}|\boldsymbol{\theta}, \boldsymbol{\mu}) = \prod\nolimits_{i=1}^{n} \mathbb{P}(s_i|\boldsymbol{\theta}, \boldsymbol{\mu}) \tag{4}$$

и при гипотезе $\mathbb{P}(s_i = k|\mathbf{0}, \mathbf{0}) = p_k$. Мы собираемся построить тест на основе локальных свойств отношения правдоподобия $\mathbb{P}(\mathbf{s}|\boldsymbol{\theta}, \boldsymbol{\mu})/\mathbb{P}(\mathbf{s}|\mathbf{0}, \mathbf{0})$, поэтому нас будут интересовать свойства распределения $\mathbb{P}(\mathbf{s}|\boldsymbol{\theta}, \boldsymbol{\mu})$, его производные по параметрам $\boldsymbol{\theta}$ и $\boldsymbol{\mu}$. Если функцией распределения случайных погрешностей ε_i является некоторая $F_{i,\boldsymbol{\mu}} \in \mathbb{F}_{\boldsymbol{\mu}}$, то

$$\mathbb{P}(s_i = k|\boldsymbol{\theta}, \boldsymbol{\mu}) = F_{i,\boldsymbol{\mu}}(c_k - \mathbf{X}_i'\boldsymbol{\theta}) - F_{i,\boldsymbol{\mu}}(c_{k-1} - \mathbf{X}_i'\boldsymbol{\theta}) \tag{5}$$

при всех $k = 1, \ldots, K$, если дополнительно обозначить $c_0 = -\infty$ и $c_K = \infty$. Чтобы описать изменение вероятностей в зависимости от параметров $\boldsymbol{\mu}$, рассмотрим параметризованные множества альтернативных распределений.

Определение 2.1. Множество распределений $\mathbb{F}(F_{i,\mathbf{0}}) \subset \mathbb{F}(\overline{p}, \mathbb{C}, \mathbb{M})$, для которого при каждом $\boldsymbol{\mu} \in \mathbb{M}$ пересечение $\mathbb{F}(F_{i,\mathbf{0}}) \bigcap \mathbb{F}_{\boldsymbol{\mu}}$ состоит из единственного распределения и $\mathbb{F}(F_{i,\mathbf{0}}) \bigcap \mathbb{F}_{\mathbf{0}} = \{F_{i,\mathbf{0}}\}$, будем называть *траекторией альтернативных распределений.*

Значения $F_{i,\boldsymbol{\mu}}(u)$ в рамках одной траектории можно рассматривать как функцию двух аргументов — u и $\boldsymbol{\mu}$. Будем говорить в связи с этим о функции $F_{i,\boldsymbol{\mu}}(u)$ на траектории $\mathbb{F}(F_{i,\mathbf{0}})$.

Утверждение 1. Пусть на некоторой траектории $\mathbb{F}(F)$ функция $F_{\boldsymbol{\mu}}(u)$ непрерывно дифференцируема в окрестностях точек $(u, \boldsymbol{\mu}) = (c_k, \mathbf{0})$. Тогда на этой траектории

$$\begin{aligned} \boldsymbol{\nabla}_{\boldsymbol{\theta}}\mathbb{P}(s_i = k|\mathbf{0}, \mathbf{0}) &= \mathbf{X}_i[f(c_{k-1}) - f(c_k)], \\ \boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathbb{P}(s_i = k|\mathbf{0}, \mathbf{0}) &= f(c_{k-1})\mathbf{d}_{k-1} - f(c_k)\mathbf{d}_k, \end{aligned} \tag{6}$$

где f — производная функции F в соответствующих точках и $f(-\infty) = f(\infty) = 0$.

Таким образом, на любой непрерывно дифференцируемой альтернативной траектории производные правдоподобия индикаторных признаков выражаются через значения плотности гипотетического распределения (принадлежащего данной траектории) на границах априорного разбиения. Этим можно воспользоваться при построении процедуры проверки гипотез на том основании, что множество распределений, отвечающих сложной альтернативе H_1, представляется в виде $\mathbb{F}(\overline{p}, \mathbb{C}, \mathbb{M}) \setminus \mathbb{F}_{\mathbf{0}}$, а априорный класс $\mathbb{F}(\overline{p}, \mathbb{C}, \mathbb{M})$ можно представить в виде объединения непрерывно дифференцируемых альтернативных траекторий. Следующий результат является простым следствием из (4) и утверждения 1.

Утверждение 2. Пусть на траекториях $\mathbb{F}(F_i)$ функции $F_{i,\boldsymbol{\mu}}(u)$ непрерывно дифференцируемы в окрестностях точек $(u, \boldsymbol{\mu}) = (c_k, \mathbf{0})$. Тогда на этих траекториях

$$\begin{aligned} \boldsymbol{\nabla}_{\boldsymbol{\theta}}\mathbb{P}(\mathbf{s}|\mathbf{0}, \mathbf{0}) &= -\mathbb{P}(\mathbf{s}|\mathbf{0}, \mathbf{0})A_1 \sum_{i=1}^{n} \mathbf{X}_i R_{1i}(s_i), \\ \boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathbb{P}(\mathbf{s}|\mathbf{0}, \mathbf{0}) &= -\mathbb{P}(\mathbf{s}|\mathbf{0}, \mathbf{0})A_2 \sum_{i=1}^{n} \mathbf{R}_{2i}(s_i), \end{aligned} \tag{7}$$

где

$$R_{1i}(k) = [f_i(c_k) - f_i(c_{k-1})] \,/\, (A_1 p_k)\,, \qquad (8)$$
$$\mathbf{R}_{2i}(k) = [f_i(c_k)\mathbf{d}_k - f_i(c_{k-1})\mathbf{d}_{k-1}] \,/\, (A_2 p_k)\,.$$

В полученных выражениях предполагается, что если $p_k = 0$, то $R_{ji}(k) = 0$. Постоянные A_1 и A_2 введены здесь потому, что при некоторых условиях их удается выбрать так, чтобы наборы значений $R_{ji} = \{R_{ji}(1), \ldots, R_{ji}(K)\}$ не зависели от неизвестных величин. Например, для квантильной регрессии, если предположить, что значения $f_i(0)$ не зависят от i, то можно взять $A_1 = -f_i(0)$, после чего $R_{1i} = \{-1/p, 0, 1/(1-p)\}$. Для модели с двумя квантилями, если $f_i(-\mu) = f_i(\mu)$ и эти значения не зависят от i, то при $A_1 = A_2 = -f_i(\mu)/p$ получаем $R_{1i} = \{-1, 0, 1\}$, $R_{2i} = \{1, -2p/(1-2p), 1\}$. В то же время для модели с тремя квантилями, даже если значения $f_i(0)$ и $f_i(-\mu) = f_i(\mu)$ не зависят от i, то с помощью масштабных множителей не удается полностью устранить зависимость R_{ji} от неизвестных величин. Так, при $A_1 = A_2 = -2f_i(\mu)/(1-2p)$, если обозначить $Q = (1-2p)/(2p)$ и $\alpha_i = [f_i(0) - f_i(\mu)]/f_i(\mu)$, то $R_{1i} = \{-Q, -\alpha_i, 0, \alpha_i, Q\}$, $R_{2i} = \{Q, -1, 0, -1, Q\}$.

Полученные наборы величин будем называть *метками* множеств априорного разбиения по аналогии с термином, который используется в ранговом анализе для обозначения весов рангов. Даже если предположения, сделанные при получении наборов меток, не выполняются, мы будем их использовать, обозначая через B_j вместо R_{ji}. Важным свойством наборов меток (8) является их нулевое математическое ожидание по априорному распределению, т. е. $\sum\limits_{k=1}^{K} R_{ji}(k)p_k = 0$. При переходе к модифицированным меткам B_j это свойство сохраняется.

Соотношения (7) дают возможность записать градиент для отношения правдоподобия $\mathbb{P}(\mathbf{s}|\boldsymbol{\theta}, \boldsymbol{\mu})/\mathbb{P}(\mathbf{s}|\mathbf{0}, \mathbf{0})$ при гипотетических параметрах, и это может служить основой для применения принципа максимума отношения правдоподобия при построении теста. Малая норма градиента отношения правдоподобия является косвенным свидетельством того, что его экстремум в пространстве параметров лежит недалеко от гипотетических значений. Это приводит к тесту, который отклоняет гипотезу на уровне значимости γ, если

$$\boldsymbol{\xi}_n' \mathbf{V}_n^{-1} \boldsymbol{\xi}_n > t_\gamma, \qquad (9)$$

где

$$\boldsymbol{\xi}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} \mathbf{X}_i B_1(s_i) \\ \mathbf{B}_2(s_i) \end{pmatrix}; \quad \mathbf{V}_n = \begin{pmatrix} d_1^2 \mathbf{V}_{X,n} & \mathbf{E}_{X,n}\mathbf{C}' \\ \mathbf{C}\mathbf{E}_{X,n}' & \mathbf{D}_2 \end{pmatrix};$$

$\mathbf{V}_{X,n} = \dfrac{1}{n}\sum\limits_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i'$; $\mathbf{E}_{X,n} = \dfrac{1}{n}\sum\limits_{i=1}^{n} \mathbf{X}_i$; $d_1^2 = \sum\limits_{k=1}^{K} p_k B_1^2(k)$; $\mathbf{C} = \sum\limits_{k=1}^{K} p_k B_1(k)\mathbf{B}_2(k)$; $\mathbf{D}_2 = \sum\limits_{k=1}^{K} p_k \mathbf{B}_2(k)\mathbf{B}_2'(k)$. При гипотезе $\mathbb{E}\{\boldsymbol{\xi}_n\} = \mathbf{0}$ и $\mathbf{V}_n = \mathbb{E}\{\boldsymbol{\xi}_n\boldsymbol{\xi}_n'\}$, поэтому матрица $\mathbf{V}_n$ играет в (9) нормирующую роль. Векторная статистика $\boldsymbol{\xi}_n$ зависит от случайных величин только через метки. Такие статистики уместно называть индикаторными, так как они используют факты принадлежности остатков множествам априорного разбиения.

Кроме того, будем рассматривать только такие наборы меток, для которых при всех $k = 1, \ldots, K-1$ либо $B_1(k+1) \neq B_1(k)$, либо $\mathbf{B}_2(k+1) \neq \mathbf{B}_2(k)$. В противном случае

пару смежных интервалов априорного разбиения можно объединить, что не повлияет на значения индикаторной статистики.

Для моделей с двумя и тремя квантилями имеет место $\mathbf{C} = \mathbf{0}$, что можно трактовать как некоррелированность наборов меток B_1 и $\mathbf{B}_2$. В этих моделях число параметров априорного разбиения $M = 1$, поэтому $\mathbf{D}_2$ на самом деле является скаляром, который мы будем обозначать через d_2^2. Если обозначить через $\nu_{ik} = I\{Y_i \in C_k(\mathbf{0})\}$ индикаторы остатков, то для рассматриваемых моделей статистику теста (9) можно конкретизировать. Так, для квантильной регрессии имеем

$$\boldsymbol{\xi}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{X}_i \left(\frac{\nu_{i3}}{1-p} - \frac{\nu_{i1}}{p} \right), \quad \mathbf{V}_n = \frac{1}{p(1-p)} \mathbf{V}_{X,n}, \tag{10}$$

т. е. здесь статистика строится на суммах взвешенных знаков остатков. Для модели с двумя квантилями получаем

$$\boldsymbol{\xi}_n' \mathbf{V}_n^{-1} \boldsymbol{\xi}_n = \frac{1}{2p} \boldsymbol{\xi}_{n,1}' \mathbf{V}_{X,n}^{-1} \boldsymbol{\xi}_{n,1} + \frac{1-2p}{2p} \xi_{n,2}^2, \tag{11}$$

где

$$\boldsymbol{\xi}_{n,1} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{X}_i \left(\nu_{i3} - \nu_{i1} \right);$$

$$\xi_{n,2} = 2pn^{-\frac{1}{2}} \sum_{i=1}^{n} \left(\frac{\nu_{i1} + \nu_{i3}}{2p} - \frac{\nu_{i2}}{1-2p} \right).$$

Для модели с тремя квантилями, если вместо неизвестных величин α_i использовать в метках априорную догадку α_A, то вычисления дают

$$\boldsymbol{\xi}_n' \mathbf{V}_n^{-1} \boldsymbol{\xi}_n = \frac{1}{(1-2p)(\alpha_A^2 - \frac{1-2p}{2p})} \boldsymbol{\xi}_{n,1}' \mathbf{V}_{X,n}^{-1} \boldsymbol{\xi}_{n,1} + \frac{2p}{1-2p} \xi_{n,2}^2, \tag{12}$$

где

$$\boldsymbol{\xi}_{n,1} = \frac{1-2p}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{X}_i \left(\frac{\nu_{i5} - \nu_{i1}}{2p} + \alpha_A \frac{\nu_{i4} - \nu_{i2}}{1-2p} \right);$$

$$\xi_{n,2} = \frac{1-2p}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{\nu_{i1} + \nu_{i5}}{2p} - \frac{\nu_{i2} + \nu_{i4}}{1-2p} \right).$$

Из структуры полученных выражений видно, что статистики $\boldsymbol{\xi}_{n,1}$ накапливают информацию о сдвиге остатков за счет отклонения параметров $\boldsymbol{\theta}$, а статистики $\xi_{n,2}$ накапливают информацию об отклонениях параметра масштаба μ. Если проверяется гипотеза $(\boldsymbol{\theta}', \boldsymbol{\mu}')' = (\boldsymbol{\theta}_0', \boldsymbol{\mu}_0')'$, то в (10)–(12) достаточно положить $\nu_{ik} = I\{Y_i - \mathbf{X}_i' \boldsymbol{\theta}_0 \in C_k(\boldsymbol{\mu}_0)\}$.

3. Свойства статистик и оценок

Для определения порога t_γ в тесте (9) необходимо знать распределение его статистики при гипотезе. Это распределение может быть указано точно, так как при гипотезе случайные величины s_i независимы и принимают значения 1, ..., K с вероятностями $p_1, \ldots, p_K$. Однако вычислить функцию распределения индикаторной статистики аналитически не представляется возможным, поэтому ее квантиль t_γ уровня $1 - \gamma$

приходится получать методом Монте-Карло. При больших n можно воспользоваться нормальной предельной аппроксимацией. Далее при изучении асимптотических свойств индикаторных статистик и оценок мы будем в разных сочетаниях ссылаться на следующие условия.

(1а) Элементы матрицы плана $\mathbf{X}$ ограничены равномерно по n.

(1б) $\mathbf{V}_X = \lim_{n\to\infty} \mathbf{V}_{X,n} > 0$.

(1в) $\mathbf{V} = \lim_{n\to\infty} \mathbf{V}_n > 0$.

(2а) $\sum\limits_{k=1}^{K} B_1(k)p_k = 0$ и $\sum\limits_{k=1}^{K} \mathbf{B}_2(k)p_k = \mathbf{0}$.

(3а) Случайные векторы $\varepsilon_1, \ldots, \varepsilon_n$ независимы, $\mathbb{P}\{\varepsilon_i < x\} = F_i(x)$, $F_i \in \mathbb{F}(\overline{p}, \mathbb{C}, \mathbb{M})$.

(3б) Существуют $L > 0$ и $\delta > 0$ такие, что для любых достаточно больших n и любых $i \in \{1, \ldots, n\}$ выполняется условие $|F_i(u_1) - F_i(u_2)| < L|u_1 - u_2|$, если $|u_1 - u_2| < \delta$.

(3в) Для всех $k \in \{1, \ldots, K-1\}$ в некоторой окрестности границы интервалов априорного разбиения c_k существуют непрерывные плотности $f_i(x) = \dfrac{d}{dx}F_i(x)$, причем $f_i(c_k)$ ограничены равномерно по n, а в окрестности точки c_k плотности $f_i(x)$ удовлетворяют условию $\exists L_0{>}0$ и $\delta_0{>}0$ (общие для всех n) такие, что $|f_i(u_1) - f_i(u_2)| < L_0|u_1 - u_2|$, если $|c_k - u_1| < \delta_0$ и $|c_k - u_2| < \delta_0$.

Без ограничения общности истинные параметры будем считать нулевыми. Введем характеристики, описывающие средний отклик меток на отклонение гипотетических параметров от истинных. Функции отклика меток определим (пользуясь условием (2а)) в виде

$$\Psi_{1,i}(u_0, \boldsymbol{\mu}_0) = \mathbb{E}B_1(s(\varepsilon_i - u_0, \boldsymbol{\mu}_0)) = B_1(K) - \sum_{k=1}^{K-1} F_i(c_k{+}\mathbf{d}_k'\boldsymbol{\mu}_0{+}u_0)\,[B_1(k{+}1) - B_1(k)]\,,$$

$$\boldsymbol{\Psi}_{2,i}(u_0, \boldsymbol{\mu}_0) = \mathbb{E}\mathbf{B}_2(s(\varepsilon_i - u_0, \boldsymbol{\mu}_0)) = \mathbf{B}_2(K) - \sum_{k=1}^{K-1} F_i(c_k{+}\mathbf{d}_k'\boldsymbol{\mu}_0{+}u_0)\,[\mathbf{B}_2(k{+}1) - \mathbf{B}_2(k)]\,.$$

Достаточные условия состоятельности индикаторных оценок будем накладывать на проверочную функцию вида

$$\boldsymbol{\Psi}_i(u_0, \boldsymbol{\mu}_0) = u_0\Psi_{1i}(u_0, \boldsymbol{\mu}_0) + \boldsymbol{\mu}_0'\boldsymbol{\Psi}_{2i}(u_0, \boldsymbol{\mu}_0). \tag{13}$$

При описании условий состоятельности и асимптотической нормальности будем использовать чувствительность меток (производные от функций отклика)

$$\psi_{11,i}(u_0, \boldsymbol{\mu}_0) = \sum_{k=1}^{K-1}[B_1(k) - B_1(k+1)]f_i(c_k + \mathbf{d}_k'\boldsymbol{\mu}_0 + u_0),$$

$$\boldsymbol{\psi}_{21,i}(u_0, \boldsymbol{\mu}_0) = \sum_{k=1}^{K-1}[\mathbf{B}_2(k) - \mathbf{B}_2(k+1)]f_i(c_k + \mathbf{d}_k'\boldsymbol{\mu}_0 + u_0),$$

$$\boldsymbol{\psi}_{12,i}(u_0, \boldsymbol{\mu}_0) = \sum_{k=1}^{K-1}[B_1(k) - B_1(k+1)]\mathbf{d}_k f_i(c_k + \mathbf{d}_k'\boldsymbol{\mu}_0 + u_0),$$

$$\boldsymbol{\psi}_{22,i}(u_0, \boldsymbol{\mu}_0) = \sum_{k=1}^{K-1}[\mathbf{B}_2(k) - \mathbf{B}_2(k+1)]\mathbf{d}_k' f_i(c_k + \mathbf{d}_k'\boldsymbol{\mu}_0 + u_0),$$

$$\boldsymbol{\psi}_i(u_0, \boldsymbol{\mu}_0) = \begin{pmatrix} \psi_{11,i}(u_0, \boldsymbol{\mu}_0) & \boldsymbol{\psi}_{12,i}'(u_0, \boldsymbol{\mu}_0) \\ \boldsymbol{\psi}_{21,i}(u_0, \boldsymbol{\mu}_0) & \boldsymbol{\psi}_{22,i}(u_0, \boldsymbol{\mu}_0) \end{pmatrix},$$

матрицу локальной (в точке нулевых аргументов, соответствующих истинным параметрам) чувствительности меток $\boldsymbol{\psi}_i = \boldsymbol{\psi}_i(0, \mathbf{0})$, состоящую из блоков $\boldsymbol{\psi}_{kl,i} = \boldsymbol{\psi}_{kl,i}(0, \mathbf{0})$, а также матрицу локальной чувствительности индикаторной статистики

$$\boldsymbol{\psi}_{X,n} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \psi_{11,i} \mathbf{X}_i \mathbf{X}_i' & \mathbf{X}_i \boldsymbol{\psi}_{12,i}' \\ \boldsymbol{\psi}_{21,i} \mathbf{X}_i' & \boldsymbol{\psi}_{22,i} \end{pmatrix}.$$

Далее будем ссылаться на следующие условия регулярности отклика меток.

(4а) Для всех $u_0 \in \mathbb{R}^1$ и $\boldsymbol{\mu}_0 \in [\mathbb{M}]$ выполняется $\Psi_i(u_0, \boldsymbol{\mu}_0) \leq 0$ и равенство здесь достигается, если и только если $u_0 = 0$ и $\boldsymbol{\mu}_0 = \mathbf{0}$.

(4б) Существуют постоянные $L_0 > 0$, $\delta_0 > 0$ такие, что при $\boldsymbol{\mu}_0 \in \mathbb{M}$ и $u_0^2 + \|\boldsymbol{\mu}_0\|^2 < \delta_0^2$ равномерно по n выполняется $\Psi_i(u_0, \boldsymbol{\mu}_0) \leq -L_0(u_0^2 + \|\boldsymbol{\mu}_0\|^2)$.

(4в) Для любого $R > 0$ существует $c(R) > 0$, такое, что при всех u_0 и $\boldsymbol{\mu}_0$, удовлетворяющих ограничениям $\boldsymbol{\mu}_0 \in \mathbb{M}$ и $u_0^2 + \|\boldsymbol{\mu}_0\|^2 > R^2$, равномерно по n выполняется $\Psi_i(u_0, \boldsymbol{\mu}_0) \leq -c(R)\sqrt{u_0^2 + \|\boldsymbol{\mu}_0\|^2}$.

(4г) При достаточно больших n для всех $i \in \{1, \ldots, n\}$ матрицы $\dfrac{1}{2}[\boldsymbol{\psi}_i + \boldsymbol{\psi}_i']$ отрицательно определены, а их собственные значения отделимы от нуля равномерно по n.

(4д) Существует невырожденная предельная матрица $\boldsymbol{\psi}_X = \lim_{n \to \infty} \boldsymbol{\psi}_{X,n}$.

Начнем с утверждения об асимптотической нормальности $\boldsymbol{\xi}_n$ из (9), которое благодаря ограниченности всех моментов суммируемых случайных векторов нетрудно получить классическим способом — с помощью аппарата характеристических функций.

Утверждение 3. Пусть выполнены условия (1а), (1в) и (2а). Тогда при гипотезе

$$\boldsymbol{\xi}_n \xrightarrow[n \to \infty]{d} N(0, \mathbf{V}), \quad \boldsymbol{\xi}_n' \mathbf{V}_n^{-1} \boldsymbol{\xi}_n \xrightarrow[n \to \infty]{d} \chi^2_{T+M}.$$

Условие (1в) следует из (1б) для всех трех рассмотренных нами моделей. Таким образом, для обеспечения асимптотического уровня значимости γ в (9) достаточно в качестве t_γ взять квантиль уровня $1 - \gamma$ распределения χ^2_{T+M}. Утверждение 3 позволяет использовать для получения оценок принцип максимального достигнутого уровня значимости, согласно которому оценками являются параметры $\boldsymbol{\theta}_0$ и $\boldsymbol{\mu}_0$, обеспечивающие наибольший достигнутый уровень значимости при проверке гипотезы $(\boldsymbol{\theta}', \boldsymbol{\mu}')' = (\boldsymbol{\theta}_0', \boldsymbol{\mu}_0')'$. Используя введенную ранее функцию $s(u, \boldsymbol{v})$, явно выразим зависимость индикаторных признаков от гипотетических параметров: $s_i = s_i(\boldsymbol{\theta}_0, \boldsymbol{\mu}_0) = s(Y_i - \mathbf{X}_i' \boldsymbol{\theta}_0, \boldsymbol{\mu}_0)$. Тогда $\boldsymbol{\xi}_n = \boldsymbol{\xi}_n(\boldsymbol{\theta}_0, \boldsymbol{\mu}_0)$ и принцип максимального достигнутого уровня значимости приводит к определению оценок в виде

$$(\boldsymbol{\theta}_n', \boldsymbol{\mu}_n')' = \arg \min_{\boldsymbol{\theta}_0 \in \mathbb{R}^T,\ \boldsymbol{\mu}_0 \in \mathbb{M}} \boldsymbol{\xi}_n'(\boldsymbol{\theta}_0, \boldsymbol{\mu}_0) \mathbf{V}_n^{-1} \boldsymbol{\xi}_n(\boldsymbol{\theta}_0, \boldsymbol{\mu}_0). \tag{14}$$

Целевая функция задачи оптимизации (14) кусочно-постоянна в пространстве параметров и испытывает скачки на гиперплоскостях, которые определяются уравнениями вида $\mathbf{X}_i' \boldsymbol{\theta}_0 + \mathbf{d}_k' \boldsymbol{\mu}_0 = Y_i - c_k$, $i = 1, \ldots, n$, $k = 1, \ldots, K-1$. Кроме того, минимум в (14) может быть не единственным, а достигаться на одном или нескольких многогранниках. В связи с этим для поиска оценок необходимо применять специальные методы, которые здесь не рассматриваются. Далее под оценками мы будем понимать любые параметры, доставляющие минимум в (14).

Доверительную область для параметров с уровнем доверия β можно определить в виде

$$\Theta_n(\beta) = \left\{ (\boldsymbol{\theta}_n', \boldsymbol{\mu}_n')' : \ \boldsymbol{\xi}_n'(\boldsymbol{\theta}_0, \boldsymbol{\mu}_0) \mathbf{V}_n^{-1} \boldsymbol{\xi}_n(\boldsymbol{\theta}_0, \boldsymbol{\mu}_0) < F_H^{-1}(\beta) \right\}, \tag{15}$$

где F_H — функция распределения тестовой статистики, в качестве которой можно взять функцию распределения случайной величины χ^2_{T+M}, получив асимптотическую доверительную область. В любом случае такая доверительная область состоит из объединения многогранников, на которых целевая функция в (14) постоянна и принимает достаточно малые значения. Добавим к этому, что если доказать асимптотическую нормальность оценки (14), то можно построить эллиптическую доверительную область для параметров.

Достаточные условия состоятельности и асимптотической нормальности сформулируем для оценки

$$(\boldsymbol{\theta}'_n, \boldsymbol{\mu}'_n)' = \arg \min_{(\boldsymbol{\theta}'_0, \boldsymbol{\mu}'_0)' \in V(D)} \boldsymbol{\xi}'_n(\boldsymbol{\theta}_0, \boldsymbol{\mu}_0) \mathbf{V}_n^{-1} \boldsymbol{\xi}_n(\boldsymbol{\theta}_0, \boldsymbol{\mu}_0), \tag{16}$$

где $V(D) = \mathbb{R}^T \times \mathbb{M}(D)$ и величина $D{>}0$ настолько мала, что $(\boldsymbol{\theta}', \boldsymbol{\mu})' \in V(D)$. Для этого достаточно, чтобы

$$D < \min_{\substack{k=1,\ldots,K-2 \\ c_{k+1} \neq c_k,\ \boldsymbol{d}_{k+1} \neq \boldsymbol{d}_k}} \frac{(\mathbf{d}_{k+1} - \mathbf{d}_k)'\boldsymbol{\mu} + (c_{k+1} - c_k)}{\|\mathbf{d}_{k+1} - \mathbf{d}_k\|^2} \ .$$

Поскольку $V(D) \uparrow V = \mathbb{R}^T \times \mathbb{M}$ при $D \to 0$ и D произвольно мало, факт состоятельности оценки (16) пригоден для описания свойств индикаторных оценок по крайней мере с точки зрения практического применения. При отсутствии параметризации априорного разбиения оценки (14) и (16) совпадают.

Теорема 3.1. Состоятельность индикаторной оценки. *Пусть для модели (1) выполнены условия (1а), (1б), (1в), (2а), (3а), (3б), (4а), (4б), (4в), а величина D достаточно мала, чтобы $(\boldsymbol{\theta}', \boldsymbol{\mu})' \in V(D)$. Тогда оценка (16) является состоятельной.*

Теорема 3.2. Асимптотическая нормальность индикаторной оценки. *Пусть для модели (1) выполнены условия (1а), (1б), (2а), (3а), (3б), (3в), (4а), (4б), (4в), (4г) и (4д). Тогда при достаточно малых $D > 0$ оценка (16) является асимптотически нормальной:* $\sqrt{n}(\boldsymbol{\theta}'_n, \boldsymbol{\mu}'_n)' \xrightarrow[n\to\infty]{d} N\left(\mathbf{0}, \boldsymbol{\psi}_X^{-1} \mathbf{V} (\boldsymbol{\psi}_X^{-1})'\right)$.

Основные условия состоятельности и асимптотической нормальности индикаторной оценки параметров выполняются для модели квантильной регрессии. В модели с двумя квантилями они выполняются для любых непрерывно дифференцируемых симметричных функций распределения случайных погрешностей. Для модели с тремя квантилями численная проверка показывает, что эти условия выполняются для распределений из семейства Стьюдента (в том числе Коши), нормального, равномерного распределения Лапласа, для распределений семейства Тьюки (модель симметричного нормального засорения).

Запишем асимптотические ковариации оценок параметров моделей в условиях одинаковой распределенности случайных погрешностей ($F_i = F$). Для квантильной регрессии асимптотическая ковариация индикаторных оценок по теореме 3.2 равна

$$\mathbf{D}_{\boldsymbol{\theta}} = \frac{p(1-p)}{f^2(0)} \mathbf{V}_X^{-1}. \tag{17}$$

Для модели с двумя квантилями дополнительно будем предполагать, что $f(\mu) = f(-\mu)$. Тогда индикаторные оценки параметров $\boldsymbol{\theta}$ и μ асимптотически независимы и

$$\mathbf{D}_{\boldsymbol{\theta}} = \frac{p}{2f^2(\mu)} \mathbf{V}_X^{-1}, \quad D_\mu = \frac{p(1-2p)}{2f^2(\mu)}. \tag{18}$$

В тех же условиях для модели с тремя квантилями получаем асимптотически независимые оценки с ковариациями

$$\mathbf{D}_{\boldsymbol{\theta}} = \frac{Q(\alpha_A^2+Q)}{(\alpha_A\alpha+Q)^2}\frac{p}{2f^2(\mu)}\mathbf{V}_X^{-1}, \quad D_\mu = \frac{p(1-2p)}{2f^2(\mu)}. \tag{19}$$

Примечательно, что при аналогичных условиях (17) и (18) совпадают с асимптотическими ковариациями оценок, полученных по методу взвешенных модулей остатков (см. [3] и [7] соответственно).

Сравнивая (17) и (18), приходим к выводу, что если к знанию об одной общей квантили добавить знание о симметричной ей квантили, то качество оценивания параметров $\boldsymbol{\theta}$ возрастает (асимптотическая относительная эффективность — АОЭ — равна $2(1-p)>1$).

Для модели с тремя квантилями ситуация осложняется использованием априорной догадки α_A вместо неизвестного $\alpha = [f(0)-f(\mu)]/f(\mu)$. Если рассматривать только унимодальные распределения ($\alpha>0$) и выбирать $\alpha_A = AQ$ при некотором $A>0$, то (18) уступает (19) по асимптотической относительной эффективности при оценивании $\boldsymbol{\theta}$ только когда $\alpha>[\sqrt{1+A^2Q}-1]/A$. Вычисления показывают, что если взять $A = 1/2$, то это ограничение выполняется при $p < 1/4$ для всех распределений семейства Стьюдента (включая распределение Коши), таких как нормальное, логистическое и Лапласа. При этом для всех этих распределений асимптотическая относительная эффективность неограниченно возрастает при приближении p к нулю (т. е. на крайних квантилях). Более того, с увеличением затянутости хвостов распределения рассматриваемая АОЭ увеличивается, а потери в величине АОЭ от использования α_A вместо α не превышают 10 %.

Список литературы

[1] Дмитриев Ю.Г., Тарасенко П.Ф. Автоматизированная система “Октава” для геологического прогнозирования // Вычисл. технологии. 2003. Т. 8. Спецвыпуск. С. 74–91.

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

[3] Koenker R., Bassett G. Regression quantiles // Econometrica. 1978. Vol. 46, N 1. P. 33–50.

[4] Koenker R., Bassett G. Robust tests for heteroscedasticity based on regression quantiles // Econometrika. 1982. Vol. 50. P. 43–61.

[5] Portnoy S.L. Asymptotic behavior of the number of regression quantile breakpoints // SIAM J. Sci. Statist. Comp. 1991. Vol. 12. P. 867–33.

[6] Chen L.A., Chiang Y.C. Symmetric type quantile and trimmed means for location and linear regression model // J. Nonparametric Statist. 1996. Vol. 7. P. 171–185.

[7] 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. Vol. 126, N 2. P. 423–440.

Поступила в редакцию 23 августа 2007 г.