S©MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru

Том 10, стр. 727–732 (2013) УДК 519.21
MSC 62F12

УСТОЙЧИВОСТЬ ПРОЦЕССА ЧАСТИЧНЫХ СУММ ОСТАТКОВ МНОГОМЕРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

И.С. БОРИСОВ

Abstract. We discuss a refinement of the MacNeill's result (1978) on limit behavior of the so-called residual process of a linear regression model. We study stability of the process with respect to L_2-variations of the regressor. As an example, we consider the case when the regressor is a smooth function of the variational series based on n identically distributed observations not necessarily independent.

Keywords: linear regression, random regressor, residual process, least-square estimator, variational series.

Рассматривается модель линейной регрессии, введенная в 1978 году в статье Макнейла [1], в редакции более поздней работы Бишофа [2]:

$$Y_i = \sum_{k=1}^{m} \theta_k f_k(i/n) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

где $\{\varepsilon_i\}$ – независимые одинаково распределенные центрированные случайные величины с дисперсией $\sigma^2 < \infty$, $\theta = (\theta_1, \ldots, \theta_m)$ – неизвестный векторный параметр, $f(x) = (f_1(x), \ldots, f_m(x))$ – известный векторный регрессор с непрерывными линейно независимыми координатами, имеющими на $[0, 1]$ ограниченную вариацию. Нам понадобятся два скалярных произведения в пространствах квадратично-интегрируемых функций:

$$\langle g_1, g_2 \rangle := \int_0^1 g_1(x) g_2(x) dx,$$

Borisov I.S., Stability of the partial sum process of residuals in a multiple linear regression model

Работа поддержана РФФИ (гранты 13-01-12415 офи-м, 13-01-00511).
Поступила 27 ноября 2013 г., опубликована 30 декабря 2013 г.

$$\langle g_1, g_2\rangle_n := \int_0^1 g_1(x)g_2(x)\lambda_n(dx) = \frac{1}{n}\sum_{i=1}^{n} g_1(i/n)g_2(i/n),$$

где $\lambda_n(x)$ – распределение с массами $1/n$ в точках $\{i/n;\, i = 1,\dots,n\}$. Соответствующую каждому из введенных скалярных произведений норму будем обозначим как $\|\cdot\|$ и $\|\cdot\|_n$ соответственно.

В отличие от [1] и [2] мы ослабляем ограничения на регрессор, а именно будем предполагать, что

(i) координаты $f_k(x)$ регрессора $f(x)$ непрерывны почти всюду на $[0,1]$ и ограничены на отрезках $[\varepsilon, 1-\varepsilon]$ при всех $\varepsilon > 0$;

(ii) функции $f_k^2(x)$ интегрируемы на $[0,1]$ в смысле Лебега, и при $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n-1} f_k^2(i/n) \to \int_0^1 f_k^2(x)dx > 0.$$

Ясно, что приведенные условия допускают особенности (интегрируемые) рассматриваемых функций в точках 0 и 1 и покрывают классическую интегрируемость по Риману на отрезке $[0,1]$. Например, если неограниченные функции f_k монотонны и квадратично интегрируемы по Лебегу на $[0,1]$, то условия (i) и (ii) будут выполнены. То же самое можно утверждать для конечных линейных комбинаций таких функций или для кусочно-монотонных функций $f_k(x)$ с конечным числом интервалов монотонности, скажем, для полиномиальных преобразований монотонных функций при соответствующих моментных ограничениях. Липшицевы преобразования монотонных функций с условиями (i) и (ii) также удовлетворяют этим условиям.

В дальнейшем нам иногда будет удобно использовать "полные" интегральные суммы Римана вида $\frac{1}{n}\sum_{i=1}^{n} g(f_k(i/n)) = \int_0^1 g(f_k(x))\lambda_n(dx)$, где без ограничения общности будем полагать $f_k(1) = 0$, если $f_k(x)$ неограничена в окрестности точки 1. При этом нижеследующие интегралы Лебега не будут реагировать на подобные изменения интегрирующих функций в отдельных точках. Отметим, что при таком соглашении и выполнении условий (i) и (ii) имеет место предельное соотношение

$$\lim_{n\to\infty}\int_a^b f_k^2(x)\lambda_n(dx) = \int_a^b f_k^2(x)dx \tag{2}$$

при любых $a, b \in [0,1]$.

Как хорошо известно (см., например, [3]), оценка наименьших квадратов неизвестного параметра в (1) в наших обозначениях имеет вид

$$\theta_n^* := (\theta_{n,1}^*, \dots, \theta_{n,m}^*) = \frac{1}{n}\left(\sum_{i\le n} f_1(i/n)Y_i, \dots, \sum_{i\le n} f_m(i/n)Y_i\right) C_n^{-1},$$

где $C_n^{-1} = \left(a_{j,k}^{(n)}\right)_{m\times m}$ – обратная матрица к матрице Грама $C_n := \left(\langle f_i, f_j\rangle_n\right)_{m\times m}$. Отметим, что матрица Грама любой системы линейно независимых элементов в любом гильбертовом пространстве положительно определена. Стало быть, ее обратная матрица корректно задана.

Обозначим

$$Y_i^* := \sum_{k=1}^{m} \theta_{n,k}^* f_k(i/n), \quad S_n(t) := \frac{1}{\sigma\sqrt{n}}\sum_{k\le nt} \varepsilon_k,$$

$$l_g(S_n) := \int_0^1 g(x)dS_n(x) = \frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n} g(k/n)\varepsilon_k.$$

Следуя Макнейлу [1], введем в рассмотрение нормированный процесс частичных сумм так называемых остатков регрессии, который после элементарных преобразований принимает вид

$$Z_n(t) := (\sigma n)^{-1/2}\sum_{i\le nt}(Y_i - Y_i^*) = S_n(t) - \sum_{k=1}^{m}\int_0^t f_k(x)\lambda_n(dx)\sum_{j=1}^{m} a_{j,k}^{(n)} l_{f_j}(S_n). \quad (3)$$

Несколько более сильная версия предельной теоремы Макнейла–Бишофа с удобным для понимания существа дела представлением допредельных и предельных процессов выглядит следующим образом.

Теорема 1. *При $n\to\infty$ и ограничениях (i) и (ii) имеет место C-сходимость в $D[0,1]$ распределений случайных процессов $Z_n(t)$ к распределению п.н. непрерывного гауссовского процесса*

$$Z(t) := W(t) - \sum_{k=1}^{m}\int_0^t f_k(x)dx\sum_{j=1}^{m} a_{j,k} l_{f_j}(W), \quad (4)$$

где $W(t)$ – стандартный винеровский процесс, $\Big(a_{j,k}\Big)_{m\times m}$ – обратная матрица к матрице Грама $C := \Big(\langle f_i, f_j\rangle\Big)_{m\times m}$.

Доказательство сходимости конечномерных распределений рассматриваемых случайных процессов является следствием многомерной центральной предельной теоремы для вектор-сумм вида $(S_n(t_1),\dots,S_n(t_r), l_{f_1}(S_n),\dots,l_{f_m}(S_n))$ при всех $t_i\in[0,1]$, а также независимости приращений процесса $S_n(t)$ и того факта, что для любых квадратично-интегрируемых на $[0,1]$ функций g_1 и g_2, удовлетворяющих (i) и (ii), при $n\to\infty$

$$\mathbf{E} l_{g_1}(S_n) l_{g_2}(S_n) = \langle g_1, g_2\rangle_n \to \langle g_1, g_2\rangle = \mathbf{E} l_{g_1}(W) l_{g_2}(W).$$

Отметим, что гауссовские случайные величины $l_g(W)$ с вышеприведенной ковариацией корректно определяются для любых $g\in L_2[0,1]$. В качестве функций g_i можно брать индикаторы любых отрезков вида $[0,t]$. Так что на самом деле речь идет только о нормальной аппроксимации конечного набора случайных величин вида $l_{g_k}(S_n)$ для функций g_k в условиях (i) и (ii), а с помощью известного приема Крамера–Уолда эта задача сводится к проблеме нормальной аппроксимации одной такой случайной величины. Выполнимость же условия Линдеберга в этом случае следует из оценки

$$\frac{1}{nB_n^2}\sum_{i=1}^{n} f^2(i/n)\mathbf{E}\varepsilon_i^2 I\left(|f(i/n)\varepsilon_i| \ge \delta B_n\sqrt{n}\right)$$

$$\le \frac{N^2}{B_n^2}\mathbf{E}\varepsilon_1^2 I\left(|\varepsilon_1| \ge \delta N^{-1} B_n\sqrt{n}\right) + \frac{\sigma^2}{nB_n^2}\sum_{i\le n:\,|f(i/n)|\ge N} f^2(i/n),$$

где $N,\delta>0$ произвольны, $B_n^2 := \sigma^2 n^{-1}\sum_{i\le n} f^2(i/n)$ имеет конечный предел в силу (ii) и соглашения касательно величины $f(1)$. Остается только отметить, что предел по n второго слагаемого правой части этого неравенства в силу (2) может быть сделан сколь угодно малым выбором достаточно большого N. Подобная схема проверки условия Линдеберга содержится, например, в [4].

Доказательство плотности семейства распределений процессов $Z_n(t)$ в (3) или, другими словами, асимптотической при $n \to \infty$ малости по вероятности их модулей непрерывности $\omega_{Z_n}(\delta)$, тоже достаточно простое. В самом деле, в представлении (3) случайный процесс $S_n(t)$ – хорошо изученный классический объект, удовлетворяющий указанному условию плотности. Кроме того, для модуля непрерывности $\omega_{\varphi_{n,k}}(\delta)$ неслучайных функций $\varphi_{n,k}(t) := \int_0^t f_k(x)\lambda_n(dx)$ имеет место очевидная оценка

$$\omega_{\varphi_{n,k}}(\delta) \le \sup_{\varepsilon<t,s<1-\varepsilon:|t-s|\le\delta} |\varphi_{n,k}(t) - \varphi_{n,k}(s)|$$

$$+2\int_0^\varepsilon |f_k(x)|\lambda_n(dx) + 2\int_{1-\varepsilon}^1 |f_k(x)|\lambda_n(dx),$$

где $\varepsilon, \delta \in (0, 1/2)$ – произвольные, нам остается лишь заметить, что оба интеграла в правой части этого неравенства с ростом n могут быть сделаны сколь угодно малыми выбором ε в силу (2) и неравенства Коши–Буняковского, а первое слагаемое – сколь угодно малым, выбором соответствующего $\delta \equiv \delta(\varepsilon)$ при всех достаточно больших n, поскольку для любых $t, s \in [\varepsilon, 1-\varepsilon]$

$$|\varphi_{n,k}(t) - \varphi_{n,k}(s)| \le \sup_{\varepsilon\le x\le 1-\varepsilon} |f_k(x)|(|t-s| + 1/n).$$

А так как последовательность случайных величин $\sum_{j=1}^m a_{j,k}^{(n)} l_{f_j}(S_n)$ ограничена по вероятности равномерно по n, то отсюда и следует асимптотическая малость по вероятности указанного выше модуля непрерывности. Теорема доказана.

Обсудим вопрос устойчивости предельного поведения случайных процессов $Z_n(t)$ при том или ином возмущении векторного регрессора $\{f_j(x);\, j \le m\}$. Будем предполагать, что при каждом n измерение компоненты $f_j(x)$ регрессора производится с некоторой случайной ошибкой $\delta_{n,j}(x)$; при этом для всех n случайный вектор-процесс $(\delta_{n,1}(x), \dots, \delta_{n,m}(x))$ не зависит от последовательности шумов $\{\varepsilon_j\}$. Вместо функций f_j в (1) подставляются функции $\tilde f_j := f_j + \delta_{n,j}$. Соответствующий процесс остатков регрессии в (3) обозначим $\tilde Z_n(t)$. Из теоремы 1 вытекает

Следствие. *Пусть* $\mathbf{E}\|\delta_{n,j}\|_n^2 \to 0$ *при* $n \to \infty$ *для всех* $j \le m$. *Тогда имеет место* C*-сходимость в* $D[0,1]$ *распределений случайных процессов* $\tilde Z_n(t)$ *к распределению процесса* $Z(t)$, *определенного в* (4).

Доказательство следует из (3) и элементарных соотношений (после применения неравенства Коши–Буняковского)

$$|\langle \tilde f_i, \tilde f_j\rangle_n - \langle f_i, f_j\rangle_n| \le \|f_i\|\|\delta_{n,j}\|_n + \|f_j\|\|\delta_{n,i}\|_n + \|\delta_{n,i}\|_n\|\delta_{n,j}\|_n,$$

$$\sup_{0\le t\le 1} |\varphi_{n,k}(t) - \tilde\varphi_{n,k}(t)| \le \|\delta_{n,k}\|_n, \quad \text{где} \quad \tilde\varphi_{n,k}(t) := \int_0^t \tilde f_k(x)\lambda_n(dx),$$

$$\mathbf{E}_{\{\delta_{n,j}\}}\left(l_{\tilde f_j}(S_n) - l_{f_j}(S_n)\right)^2 = \|\delta_{n,j}\|_n^2,$$

где символ $\mathbf{E}_{\{\delta_{n,j}\}}$ обозначает усреднение по распределению $\{\varepsilon_j\}$ при фиксации случайных ошибок $\{\delta_{n,j}\}$. Отсюда получаем, что $\sup_{0\le t\le 1} |\tilde Z_n(t) - Z_n(t)| \xrightarrow{p} 0$ при $n \to \infty$, что и доказывает наше утверждение.

В качестве примера описанного в следствии случайного возмущения регрессора рассмотрим следующую модель.

Пусть $\{X_j^{(k)};\, j \geq 1\}$, $k = 1, \ldots, m$, – конечный набор последовательностей одинаково распределенных в каждой последовательности случайных величин (но не обязательно независимых или стационарно связанных) с соответствующими функциями распределения F_k и конечными вторыми моментами. При этом сами эти последовательности могут быть произвольно связанными. По выборке объема n из каждой этой последовательности построим вариационные ряды $X_{n:1}^{(k)} \leq \ldots \leq X_{n:n}^{(k)}$, $k = 1, \ldots, m$. Мы изучаем многомерную регрессионную модель со случайным регрессором (сохранив обозначения из (1))

$$Y_i = \sum_{k=1}^{m} \theta_k g_k(X_{n:i}^{(k)}) + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{5}$$

Предполагается, что функции $g_1, \ldots, g_m$ удовлетворяют условию Липшица и

$$\mathbf{E} g_k^2(X_1^{(k)}) = \int_0^1 g_k^2(F_k^{-1}(x))dx < \infty, \quad k = 1, \ldots, m, \tag{6}$$

где $F_k^{-1}(x) := \inf\{t : F_k(t) \geq x\}$ – квантильное преобразование функции F_k. Теперь положим $f_k(x) := g_k(F_k^{-1}(x)),\ k = 1, \ldots, m$. Будем считать, что функции $\{f_k;\, k \leq m\}$ линейно независимые, причем в силу (6) они как суперпозиции липшицуемой и монотонных функций будут удовлетворять (i) и (ii) (принимая в расчет соглашение о $f_k(1)$). Наконец, обозначим через $\tilde{F}_{n,k}(t)$ эмпирическую функцию распределения, построенную по выборке $X_1^{(k)}, \ldots, X_n^{(k)}$. Легко видеть, что $\tilde{F}_{n,k}^{-1}(i/n) = X_{n:i}^{(k)}$. Положим $\tilde{f}_k(x) := g_k(\tilde{F}_{n,k}^{-1}(x))$. Тогда регрессионная модель (5) может быть представлена в виде (1) со случайным регрессором $\tilde{f}$.

Теорема 2. *Пусть регрессионная модель* (5) *удовлетворяет вышеприведенным условиям и, кроме того, для всех* $k = 1, \ldots, m$ *и* t *выполнено*

$$\frac{1}{n^2} \sum_{i,j=1}^{n} \mathbf{P}(\max\{X_i^{(k)}, X_j^{(k)}\} < t) \to F_k^2(t) \quad \textit{при} \quad n \to \infty. \tag{7}$$

Тогда имеет место C*-сходимость в* $D[0,1]$ *распределений случайных процессов* $\tilde{Z}_n(t)$*, построенных в* (3) *для введенных выше параметров рассматриваемой модели, к распределению процесса* $Z(t)$ *в* (4).

Доказательство сводится к проверке условий вышеприведенного следствия. В самом деле, в силу липшицуемости компонент $\{g_k\}$ имеем

$$\mathbf{E} \int_0^1 \left(g_k(\tilde{F}_{n,k}^{-1}(x)) - g_k(F_k^{-1}(x)) \right)^2 \lambda_n(dx)$$

$$\leq C\mathbf{E} \int_0^1 \left(\tilde{F}_{n,k}^{-1}(x) - F_k^{-1}\left(\frac{\lceil xn \rceil \wedge (n-1)}{n} \right) \right)^2 dx \leq 4C \int_{\mathbb{R}} \mathbf{E}|\tilde{F}_{n,k}(t) - F_k^{(n)}(t)||t|dt$$

$$\leq 4C \int_{\mathbb{R}} \mathbf{E}|\tilde{F}_{n,k}(t) - F_k(t)||t|dt + 4C \int_{\mathbb{R}} |F_k(t) - F_k^{(n)}(t)||t|dt, \tag{8}$$

где C – квадрат константы в условии Липшица для функции $g_k(x)$, $F_k^{(n)}(t)$ – функция распределения, соответствующая квантильному преобразованию $F_k^{-1}((\lceil xn \rceil \wedge (n-1))/n)$; здесь $\lceil a \rceil$ – ближайшее к a натуральное число, не меньшее a. Второе неравенство в (8) непосредственно следует из [5], где получены неравенства, связывающие интегральные расстояния соответственно в пространствах функций распределений и их квантильных преобразований.

Отметим также, что по построению $F_k^{(n)}(t)$ – кусочно-постоянная функция и $\sup_t |F_k(t) - F_k^{(n)}(t)| \le 2/n$.

Для оценки первого интеграла правой части (8) разобьем его на сумму двух по областям $\{t : |t| \le N\}$ и $\{t : |t| > N\}$, где $N > 0$ произвольное. Интеграл по первой области сходится к нулю в силу (7) и очевидной оценки

$$\Big(\mathbf{E}|\tilde{F}_{n,k}(t) - F_k(t)|\Big)^2 \le \mathbf{E}(\tilde{F}_{n,k}(t) - F_k(t))^2$$
$$= \frac{1}{n^2}\sum_{i,j=1}^{n}[\mathbf{P}(\max\{X_i^{(k)}, X_j^{(k)}\} < t) - F_k^2(t)] \to 0.$$

Интеграл по области $\{t : |t| > N\}$ очевидным образом мажорируется с точностью до постоянного множителя суммой $\int_{-\infty}^{-N} F_k(t)|t|dt + \int_N^{\infty}(1 - F_k(t))tdt$, которая может быть сделана сколь угодно малой выбором достаточно большого N в силу конечности второго момента $X_1^{(k)}$. Значит, можно так подобрать $N \equiv N(n) \to \infty$, чтобы оба эти интеграла стремились к нулю с ростом n.

В силу отмеченной выше близости функции $F_k^{(n)}(t)$ и $F_k(t)$ для оценки второго интеграла правой части (8) нам также нужно доказать малость его суммарного интегрального хвоста. Ограничимся оценкой отрицательного (т. е. по области $(-\infty, -N)$) хвоста. Поскольку по построению $F_k^{(n)}(-t) \equiv 0$ при $t \ge |F_k^{-1}(1/n)|$, то нам достаточно оценить соответствующий интеграл по области интегрирования $[F_k^{-1}(1/n), -N]$, предполагая, естественно, что $F_k^{-1}(1/n) < 0$ и $N < |F_k^{-1}(1/n)|$. В силу отмеченной выше оценки близости рассматриваемых функций имеем при $n \to \infty$

$$\int_{F_k^{-1}(1/n)}^{-N} |F_k(t) - F_k^{(n)}(t)||t|dt \le \frac{1}{n}\left(F_k^{-1}(1/n)\right)^2 \le \int_0^{1/n}\left(F_k^{-1}(x)\right)^2 dx \to 0.$$

Отсюда немедленно следует нужная равномерная по n малость указанного отрицательного интегрального хвоста для функции $F_k^{(n)}(t)$. Оценка положительного хвоста проводится совершенно аналогично.

Таким образом, условия приведенного выше следствия выполнены, т. е. теорема 2 доказана.

Автор благодарит рецензента за ряд полезных замечаний.

Список литературы

[1] I.B. MacNeill, *Limit processes for sequences of partial sums of regression residuals*, Ann. Probab., **6**:4 (1953), 695–698.

[2] W. Bischoff, *A functional central limit theorem for regression models*, Ann. Statist., **26**:4 (1998), 1348–1410. MR1647677

[3] Дж. Себер, *Линейный регрессионный анализ*, Мир, Москва, 1980. MR0580805

[4] Ю.Ю. Линке, А.И. Саханенко, *Асимптотически нормальное оценивание параметра в задаче дробно-линейной регрессии*, Сиб. матем. журн., **41**:1 (2000), 150–163. MR1756483

[5] И.С. Борисов, А.В. Шадрин, *Некоторые замечания к неравенству Ш.С.Эбралидзе*, Теория вероятн. и ее примен., **41**:1 (1996), 177–181. MR1404902

Игорь Семенович Борисов
Институт математики им. С. Л. Соболева СО РАН,
пр-т академика Коптюга 4,
630090, Новосибирск, Россия; Новосибирский государственный университет
E-mail address: sibam@math.nsc.ru