8. Бодров А.Г., Никитин А.А. Исследование интегрального уравнения плотности биологического вида в пространствах различных размерностей // Вестн. Моск. ун-та. Сер. 15. Вычисл. матем. и киберн. 2015. № 4. С. 7–13. (Bodrov A.G., Nikitin A.A. Examining the biological species steady-state density equation in spaces with different dimensions // Moscow Univ. Comput. Math. and Cybern. 2015. **39**. N 4. P. 157–162.)
9. Baddour N. Operational and convolution properties of three-dimensional Fourier transforms in spherical polar coordinates // J. Opt. Soc. Am. A. 2010. **10**. P. 2144–2155.

Поступила в редакцию
31.01.18

УДК 519.233.24, 519.233.5

А. Г. Белов[1]

МОДЕЛИРОВАНИЕ СОВМЕСТНОЙ ДОВЕРИТЕЛЬНОЙ ПОЛОСЫ СРЕДНЕГО ЗНАЧЕНИЯ ПОВТОРНЫХ ОТКЛИКОВ С ПРЯМОУГОЛЬНОЙ ОБЛАСТЬЮ ДЛЯ ПРЕДИКТОРОВ

В статье рассмотрена задача моделирования совместных доверительных интервалов для среднего значения повторных откликов в линейной множественной нормальной регрессионной модели с предикторными переменными, определенными в интервалах. Для ее решения применен численный метод вычисления критического значения, определяющего совместный доверительный интервал заданного уровня. Проведено численное моделирование и сравнительный анализ совместных доверительных интервалов для регрессии, среднего значения повторных откликов и отдельного наблюдения.

Ключевые слова: совместные доверительные интервалы, нормальная регрессия, повторные отклики.

1. Постановка задачи. Рассмотрим линейную множественную нормальную регрессионную модель наблюдений:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

где $\mathbf{y} = (y_1, \ldots, y_n)^T$ — вектор-столбец случайных величин (с.в.) y_i откликов, описывающих результаты i-го опыта, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ — вектор-столбец случайных "ошибок" с нормальным законом распределения $\mathfrak{L}(\boldsymbol{\varepsilon}) = \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, не зависящий от вектора параметров $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$; $\mathbf{X} = \left\|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}\right\| \in R^{n\times k}$ — регрессионная матрица из вектор-столбцов $\mathbf{x}^{(j)} = (x_{1j}, \ldots, x_{nj})^T$, оказывающих влияние только на среднее значение отклика Ey_i, при этом $\mathbf{I}_n = \operatorname{diag}(1, \ldots, 1) \in \in R^{n\times n}$, $\operatorname{rank}\mathbf{X} = k$, $k \leqslant n$.

Пусть имеется m повторных наблюдений $\mathbf{y}_m = (y_1, \ldots, y_m)^T$, соответствующих фиксированным значениям регрессоров $\mathbf{x} = (x_1, \ldots, x_k)^T$: $y_j = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon_j$, $1 \leqslant j \leqslant m$, где вектор-столбец случайных "ошибок" $\boldsymbol{\varepsilon}_m = (\varepsilon_1, \ldots, \varepsilon_m)^T$ не зависит от $\boldsymbol{\varepsilon}$ и $\mathfrak{L}(\boldsymbol{\varepsilon}_m) = \mathcal{N}_m(\mathbf{0}, \sigma^2\mathbf{I}_m)$, а $\mathfrak{L}(\boldsymbol{\varepsilon}_{m0}) = \mathcal{N}_m(\mathbf{0}, \mathbf{I}_m)$ для $\boldsymbol{\varepsilon}_{m0} = \boldsymbol{\varepsilon}_m/\sigma$.

Для среднего значения повторных откликов $\overline{y}_m = \mathbf{e}_m^T\mathbf{y}_m/m = \mathbf{x}^T\boldsymbol{\beta} + \sigma\mathbf{e}_m^T\boldsymbol{\varepsilon}_{m0}/m$, где $\mathbf{e}_m = \\ = (1, \ldots, 1)^T \in R^m$, используется $100(1-\alpha)\%$-й доверительный поточечный интервал [1]

$$\left(\hat{y} \mp t_{1-\frac{\alpha}{2},\, n-k}\,\hat{\sigma}\sqrt{\frac{1}{m} + \mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}\right), \tag{1}$$

где $\mathbf{A} = \mathbf{X}^T\mathbf{X}$, $\hat{y} = \mathbf{x}^T\hat{\boldsymbol{\beta}}$ — оценка отклика y для $\mathbf{x}$, $\hat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{y} = \boldsymbol{\beta} + \sigma\mathbf{A}^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}_0$ — оценка вектора параметров $\boldsymbol{\beta}$, найденная по выборке $\mathbf{y}$ с помощью метода наименьших квадратов (МНК),

[1] Факультет ВМК МГУ, ст. науч. сотр., к.ф.-м.н., e-mail: belov@cs.msu.ru

$\hat\sigma^2 = S(\hat{\boldsymbol\beta})/(n-k)$ — оценка σ^2, $t_{1-\frac{\alpha}{2},\,n-k}$ есть $100(1-\frac{\alpha}{2})\%$-й квантиль распределения Стьюдента $St(n-k)$, так что

$$1-\alpha = \mathrm{P}\{|t_{n-k}| < t_{1-\frac{\alpha}{2},\,n-k}\}, \quad 0<\alpha<1,$$

$$S(\hat{\boldsymbol\beta}) = (\mathbf{y}-\mathbf{X}\hat{\boldsymbol\beta})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol\beta}), \quad \boldsymbol\varepsilon_0 = \boldsymbol\varepsilon/\sigma, \quad \mathfrak{L}(\boldsymbol\varepsilon_0) = \mathcal{N}_n(\mathbf{0},\,\mathbf{I}_n).$$

Из (1) следуют известные $100(1-\alpha)\%$-е доверительные поточечные интервалы для регрессии $\mathbf{x}^T\boldsymbol\beta$ (при $m\to\infty$) и индивидуального $y=\mathbf{x}^T\boldsymbol\beta+\varepsilon_1$ (при $m=1$) значения отклика соответственно:

$$\left(\hat y \mp t_{1-\frac{\alpha}{2},\,n-k}\,\hat\sigma\sqrt{\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}\right), \quad \left(\hat y \mp t_{1-\frac{\alpha}{2},\,n-k}\,\hat\sigma\sqrt{1+\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}\right), \tag{2}$$

где $\varepsilon_1 \sim \mathcal{N}_1(0,\,\sigma^2)$ и не зависит от $\boldsymbol\varepsilon$.

Таким образом, для всех трех поточечных доверительных интервалов имеет место единая критическая константа $c = t_{1-\frac{\alpha}{2},\,n-k}$, которая определяет доверительный уровень $1-\alpha$.

Цель статьи заключается в построении совместной доверительной полосы для среднего повторных откликов вида

$$\left(\hat y \mp c\,\hat\sigma\sqrt{\frac{1}{m}+\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}\right)\ \forall\mathbf{x}\in D, \tag{3}$$

а следовательно, и для совместной доверительной полосы регрессии и индивидуального значения отклика соответственно:

$$\left(\hat y \mp c\,\hat\sigma\sqrt{\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}\right), \quad \left(\hat y \mp c\,\hat\sigma\sqrt{1+\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}\right) \quad \forall\mathbf{x}\in D, \tag{4}$$

где D есть прямоугольная область, определяемая как

$$D = \{(x_1,\ldots,x_k)^T : -\infty \leqslant a_i \leqslant x_i \leqslant b_i \leqslant \infty,\ i=1,\ldots,k\}.$$

Основная задача состоит в нахождении критической константы c, определяемой как $P\{T<c\}$, такой, чтобы доверительная полоса (3), а значит и (4), имели уровень $1-\alpha$, где

$$T = \sup_{x_i\in[a_i,b_i],\,1\leqslant i\leqslant k} \frac{|\hat y-\overline{y}_m|}{\hat\sigma\sqrt{\frac{1}{m}+\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}} = \sup_{x_i\in[a_i,b_i],\,1\leqslant i\leqslant k} \frac{|\mathbf{x}^T(\hat{\boldsymbol\beta}-\boldsymbol\beta)-\frac{1}{m}\,\sigma\mathbf{e}_m^T\boldsymbol\varepsilon_{m0}|}{\hat\sigma\sqrt{\frac{1}{m}+\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}}. \tag{5}$$

2. Метод вычисления c**.** Поскольку искомая критическая константа c определяет доверительные полосы (3), (4), то достаточно уметь рассчитывать ее для какой-нибудь из этих полос, в частности, для регрессии $\mathbf{x}^T\boldsymbol\beta$. В этом случае константа c определяется как $P\{T<c\}$, где из (5) имеем

$$T = \sup_{x_i\in[a_i,b_i],\,1\leqslant i\leqslant k} \frac{|\mathbf{x}^T(\hat{\boldsymbol\beta}-\boldsymbol\beta)|}{\hat\sigma\sqrt{\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}}. \tag{6}$$

Для решения последней оптимизационной задачи существует множество похожих подходов [2–4].

Представим величину T в виде

$$T = Q\frac{\|\mathbf{Z}\|}{(\hat\sigma/\sigma)}, \quad Q = \sup_{x_i\in[a_i,b_i],\,1\leqslant i\leqslant k} \frac{|(\mathbf{P}\mathbf{x})^T\mathbf{Z}|}{\|\mathbf{P}\mathbf{x}\|\|\mathbf{Z}\|}, \tag{7}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{P}^T\mathbf{P}, \quad \mathbf{P} = (\mathbf{p}_1,\ldots,\mathbf{p}_k)\in R^{k\times k}, \quad \mathbf{Z} = (\mathbf{P}^T)^{-1}(\hat{\boldsymbol\beta}-\boldsymbol\beta)/\sigma \sim \mathcal{N}_k(\mathbf{0},\,\mathbf{I}_k).$$

Поскольку получить формулу для распределения T сложно, то необходимо проводить ее моделирование посредством генерации случайных величин (с.в.) $\mathbf{Z}$ и с.в. $\hat\sigma/\sigma \sim \sqrt{\chi^2_{n-k}/(n-k)}$ с их дальнейшей подстановкой в (7). Основная трудность расчета T заключается в вычислении Q. Величина Q может быть получена посредством решения задачи

$$Q = \sup_{\mathbf{s}\in\Omega} \frac{|\mathbf{s}^T\mathbf{Z}|}{\|\mathbf{s}\|\|\mathbf{Z}\|}, \tag{8}$$

где $\Omega = \{\mathbf{s} : \mathbf{s} = \gamma\boldsymbol{\nu},\ \boldsymbol{\nu} \in L,\ \gamma > 0\}$, $L = \{\mathbf{P}\mathbf{x} : x_i \in [a_i,\ b_i],\ i = 1, \ldots, k\}$. Нетрудно заметить, что $\mathbf{s}^T\mathbf{Z}/(\|\mathbf{s}\|\|\mathbf{Z}\|)$ есть косинус угла между $\mathbf{s}$ и $\mathbf{Z}$. Поэтому, если $\widehat{\mathbf{s}} \in \Omega$ есть решение (8), то оно также является решением

$$\inf_{\mathbf{s}\in\Omega} \|\mathbf{s} - \mathbf{Z}\|^2.$$

Для решения этой задачи квадратичного программирования ниже будет использован "active set"-алгоритм, подробно описанный в [4], как наиболее эффективный и сходящийся за конечное число шагов.

Таким образом, критическая константа c может определяться следующим путем. Моделируется достаточно большое число M значений T_i с.в. T. Тогда $(1-\alpha)M$-е наибольшее значение $\widehat{c}$ из сгенерированного вариационного ряда считается оценкой c. Такой подход основан на том факте, что выборочная $100(1-\alpha)$-я перцентиль $\widehat{c}$ сходится почти наверное к теоретической $100(1-\alpha)$-й перцентили c при $M \to \infty$. При этом, с учетом асимптотической нормальности $\widehat{c}$ со средней c и стандартной ошибкой $s = \sqrt{\dfrac{\alpha(1-\alpha)}{g^2(c)M}}$ может быть рассчитана стандартная ошибка оценки $\widehat{c}$, где h — параметр сглаживания (в вычислениях ниже $h = 0,01$), $g(c)$ — функция плотности распределения с.в. T, которая может быть оценена как

$$g(\widehat{c}) \approx \frac{1}{Mh\sqrt{2\pi}} \sum_{i=1}^{M} \exp\left\{-\left(\frac{\widehat{c} - T_i}{4h}\right)^2\right\}.$$

3. Численное моделирование. Вычислим доверительные полосы для простой регрессии на модельных данных. Для этого выберем $l = 10$ натуральных значений регрессора $x = 1, \ldots, l$ линейной $f(x) = 0.5x+2$ зависимости. Затем для каждого из $f(x_i)$, $i = 1, \ldots, l$, независимо моделируем q случайных значений y_{ij} путем аддитивного внесения в $f(x_i)$ случайной нормально распределенной ошибки $\mathfrak{L}(\epsilon) = \mathcal{N}_1(0, 4)$ с дисперсией $\sigma^2 = 4$. В результате получим облако из $n = lq$ значений $y_{ij} = f(x_i) + \varepsilon_{ij}$, $1 \leqslant i \leqslant l$, $1 \leqslant j \leqslant q$, $l = 10$, $q = 20$, изображенных в виде кружков на рис. 1. При этом каждому x_i соответствует q повторяющихся наблюдений.

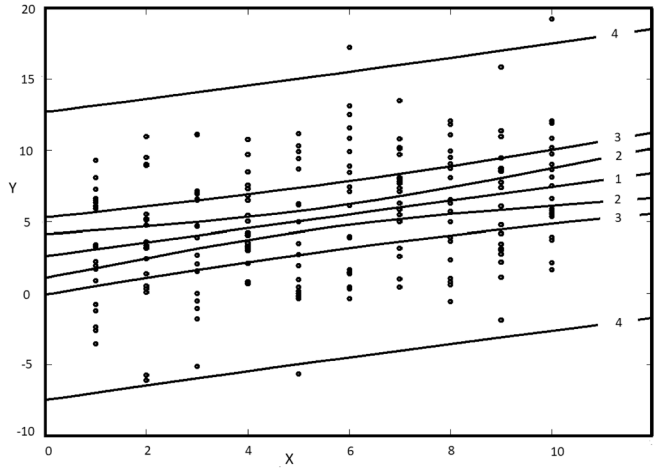

Рис. 1. Доверительные полосы: функция регрессии (линия 1); 95%-я доверительная полоса регрессии (линии 2); 95%-е совместные границы среднего значения повторных откликов ($m = q$, линии 3); 95%-я доверительная полоса отдельных наблюдений ($m = 1$, линии 4) для $x = 1, \ldots, 10$, $n = 200$, $q = 20$

На рис. 1 изображены функция регрессии (линия 1), 95%-я доверительная полоса регрессии (линии 2), 95%-е совместные границы среднего значения повторных откликов для случая $m = q$ (линии 3), а также отдельных наблюдений для случая $m = 1$ (линии 4). При моделировании T было использовано до 30000 генераций, при этом вычислялась критическая величина $\widehat{c}$ и ее стандартная ошибка $s(\widehat{c})$. В табл. 1 представлены промежуточные результаты расчетов.

Таблица 1

Данные моделирования

n	Число генераций	$\widehat{c}$	$s(\widehat{c})$
1	6840	2.3943	0.0276
2	9120	2.4031	0.0222
3	11400	2.4031	0.0205
4	13680	2.4042	0.0177
5	15960	2.4073	0.0163
6	18240	2.4072	0.0140
7	20520	2.4031	0.0127
8	22800	2.4073	0.0127
9	25080	2.4144	0.0120
10	27360	2.4172	0.0116
11	29640	2.4144	0.0115
12	30000	2.4155	0.0114

Сравнивая $c = t_{1-\frac{\alpha}{2},\,n-k} = t_{0.975,\,198} = 1.972$ с вычисленной при $M = 30000$ величиной $\widehat{c} = 2.4155$ (см. табл. 1), можно сделать вывод, что ширина поточечных доверительных границ (1) будет меньше соответствующих смоделированных совместных полос. Однако последние полосы более узкие, чем совместные доверительные границы, полученные менее точным методом коррекции Бонферрони [5] (для данного примера $c = t_{1-\frac{\alpha}{2l},\,n-k} = t_{0.9975,\,198} = 2.839$).

4. Пример. Для двухфакторной модели $k = 2$ рассмотрим выборку $n = 35$ цен на шестиядерные процессоры серии Phenom 2 фирмы AMD, различающиеся рабочей частотой (МГц) и тепловыделением (Вт) (см. табл. 2) (данные получены из интернет-ресурса http://market.yandex.ru/).

Таблица 2

Данные по процессорам AMD

n	Частота, МГц	Тепло, Вт	Цена, руб	n	Частота, МГц	Тепло, Вт	Цена, руб
1	2900	95	5164	19	2600	95	5022
2	2900	95	5198	20	2600	95	5687
3	2900	95	5523	21	3250	125	6311
4	2900	95	5785	22	3250	125	6668
5	2900	95	6370	23	3250	125	6886
6	2900	95	4710	24	3250	125	6992
7	2800	95	4800	25	3250	125	7242
8	2800	95	5275	26	3000	125	5732
9	2800	95	5501	27	3000	125	5786
10	2800	95	5580	28	3000	125	5809
11	2700	95	4663	29	3000	125	5870
12	2700	95	4690	30	3000	125	5920
13	2700	95	4804	31	2800	125	4636
14	2700	95	4857	32	2800	125	4740
15	2700	95	4890	33	2800	125	4772
16	2600	95	4611	34	2800	125	4969
17	2600	95	4719	35	2800	125	5200
18	2600	95	4860				

Для этих данных проведены расчеты доверительных полос среднего значения повторных откликов ($m = 5$) и наблюдения ($m = 1$), которые представлены на рис. 2. Для числа генераций $M = 30000$ имеем $\widehat{c} = 2.9093$, $s(\widehat{c}) = 0.0137$.

Приведенные выше расчеты выполнены с помощью авторской программы SSB (Simulation Simultaneous Bands), написанной в среде MatLab версии 7.0.5. Программа включает в себя интерфейс для импорта данных и задания желаемых параметров моделирования. Результаты вы-

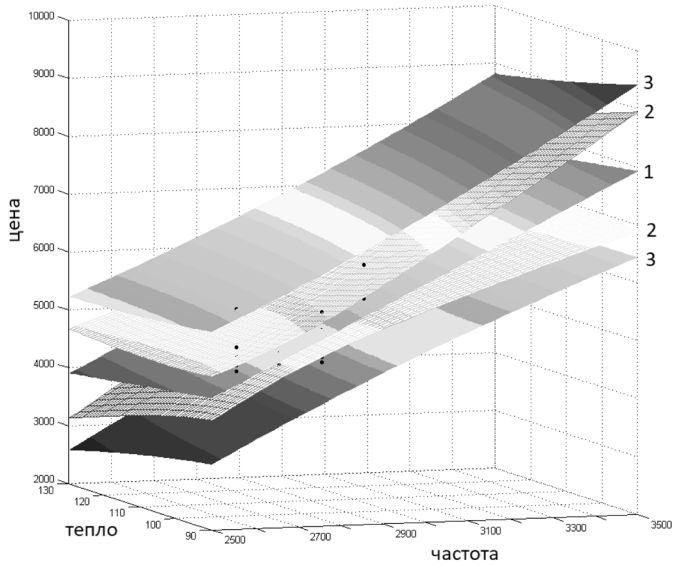

Рис. 2. Совместные доверительные границы: оценка регрессии (плоскость 1); 95%-я доверительная полоса для среднего значения повторных откликов ($m = 5$, плоскости 2) и отдельного наблюдения ($m = 1$, плоскости 3)

числений записываются в отдельный файл и могут быть представлены графически для моделей с одним или двумя предикторами.

5. Заключение. Таким образом, в работе описан численный метод расчета доверительной полосы для среднего значения повторных откликов в линейной множественной нормальной регрессии с прямоугольной областью для предикторов. Проведено численное моделирование критической величины с соответствующим вычислением доверительной полосы для среднего значения повторных откликов, регрессии и отклика. Выполнен сравнительный анализ рассчитанных полос.

СПИСОК ЛИТЕРАТУРЫ

1. Белов А. Г. Доверительное прогнозирование среднего значения повторных наблюдений // Вестн. Моск. ун-та. Сер. 15. Вычисл. матем. и киберн. 2016. № 2. С. 14–19. (Belov A. G. Confidence prediction of the mean values of multiple observations // Moscow Univ. Comput. Math. and Cybern. 2016. **36**. N 2. P. 65–70.)
2. Naiman D. Q. Simultaneous confidence-bounds in multiple-regression using predictor variable constraints // J. the Amer. Stat. Assoc. 1987. **82**. P. 214–219.
3. Liu W., Jamshidian M., Zhang Y. Multiple comparison of several linear regression lines // J. the Amer. Stat. Assoc. 2004. **99**. P. 395–403.
4. Liu W., Jamshidian M., Zhang Y., Donnelly J. Simulation-based simultaneous confidence bands in multiple linear regression with predictor variables constrained in intervals // J. Comput. and Graph. Stat. 2005. **14**. N 2. P. 459–484.
5. Bonferroni C. E. Il calcolo delle assi curazioni su gruppi di test // Studi Onore del Professore Salvatore Ortu Carboni. Rome, Italy, 1935. P. 13–60.

Поступила в редакцию
02.04.18