СОВМЕСТНЫЕ ДОВЕРИТЕЛЬНЫЕ ПОЛОСЫ ДЛЯ СРЕДНЕГО ЗНАЧЕНИЯ ПОВТОРНЫХ НАБЛЮДЕНИЙ

А.Г.Белов[1]

[1]Лаборатория обратных задач ВМК МГУ имени М.В. Ломоносова

Конференция “Ломоносовские чтения 2019”, ВМК МГУ
Секция “Кафедра математической статистики и лаборатория обратных задач”
15.04.2019 — 25.04.2019

Постановка задачи

Линейная модель наблюдений

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

$\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$, $\mathbf{X} = \left\|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}\right\| \in R^{n\times k}$,
$\mathrm{rank}(\mathbf{X}) = k$, $k \le n$, $\mathbf{x}^{(j)} = (x_{1j}, \ldots, x_{nj})^T$,
$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, $\mathbf{I}_n = \mathrm{diag}(1, \ldots, 1) \in R^{n\times n}$.
$\mathbf{y}_r = (y_{r1}, \ldots, y_{rm})^T$ — m повторных наблюдений для
$\mathbf{x}_r = (x_{r1}, \ldots, x_{rk})^T$: $y_{rj} = \mathbf{x}_r^T\boldsymbol{\beta} + \varepsilon_{rj}$, $1 \le j \le m$, $1 \le r \le N$, где
$\boldsymbol{\varepsilon}_r = (\varepsilon_{r1}, \ldots, \varepsilon_{rm})^T$ не зависит от $\boldsymbol{\varepsilon}$ и $\mathfrak{L}(\boldsymbol{\varepsilon}_r) = \mathcal{N}_m(\mathbf{0}, \sigma^2\mathbf{I}_m)$
Для среднего значения повторных откликов $\overline{y}_r = \frac{1}{m}\,\mathbf{e}_m^T\mathbf{y}_r$,

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

Постановка задачи

где $\mathbf{e}_m = (1, \ldots, 1)^T \in R^m$, $\mathbf{A} = \mathbf{X}^T\mathbf{X}$, $\hat{y}_r = \mathbf{x}_r^T\hat{\boldsymbol{\beta}}$ — оценка отклика y_r для $\mathbf{x}_r$, $\hat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{y}$ — оценка вектора параметров $\boldsymbol{\beta}$, найденная по выборке $\mathbf{y}$ с помощью метода наименьших квадратов (МНК), $\hat{\sigma}^2 = S(\hat{\boldsymbol{\beta}})/(n-k)$ — 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

$$1-\alpha = \mathrm{P}\{|t_{n-k}| < t_{1-\frac{\alpha}{2},\,n-k}\}, 0 < \alpha < 1, S(\hat{\boldsymbol{\beta}}) = (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}).$$

Из (1) следуют известные $100(1-\alpha)\%$-е доверительные поточечные интервалы для среднего $\mathbf{x}_r^T\boldsymbol{\beta}$ (при $m \to \infty$) и индивидуального $y_r = \mathbf{x}_r^T\boldsymbol{\beta} + \varepsilon_r$ (при $m = 1$) значения отклика для вектора заданных значений регрессоров $\mathbf{x}_r$, соответственно:

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

Постановка задачи

На практике наиболее часто требуется на уровне доверия $1-\alpha$ оценить доверительную полосу для среднего не для каждого отдельного вектора регрессоров $\mathbf{x}_r$, (поточечные доверительные границы — ПДГ), а одновременно по всей наблюдаемой области регрессоров $\mathbf{x}_r, r = 1, \ldots, N$ (совместные доверительные границы — СДГ). То есть необходимо, чтобы выполнялось соотношение:

$$P\{\overline{y}_r \in D_r, \forall r = 1, \ldots, N\} = P\left\{\bigcap_{r=1}^{N}(\overline{y}_r \in D_r)\right\} = 1-\alpha, \qquad (2)$$

где $L_r = \hat{y}_r - t_{1-\frac{\alpha}{2}, n-k}\,\hat{\sigma}\sqrt{\frac{1}{m} + \mathbf{x}_r^T \mathbf{A}^{-1}\mathbf{x}_r}$,

$U_r = \hat{y}_r + t_{1-\frac{\alpha}{2}, n-k}\,\hat{\sigma}\sqrt{\frac{1}{m} + \mathbf{x}_r^T \mathbf{A}^{-1}\mathbf{x}_r}$, $D_r = [L_r, U_r]$.

Методы коррекции

Метод Бонферрони (Bonferroni). Пусть имеют место ПДГ (1) с уровнем доверия $1-\alpha_r$, т.е. $1-\alpha_r = P\{(\overline{y}_r \in D_r)\}, r = 1, \ldots, N$. Тогда справедливы следующие Булевые неравенства:

$$P\left\{\bigcap_{r=1}^{N}(\overline{y}_r \in D_r)\right\} \geq 1 - \sum_{r=1}^{N} P\{(\overline{y}_r \in D_r)^c\} = 1 - \sum_{r=1}^{N} \alpha_r,$$

где $(\cdot)^c$ — операция дополнения.

Таким образом, если выбрать $\alpha_r = \alpha/N, r = 1, \ldots, N$ и построить N ПДГ с индивидуальным уровнем покрытия $1-\alpha_r$, то получим СДГ с уровнем не менее $1-\alpha$, т. е.

$$P\left\{\bigcap_{r=1}^{N}(\overline{y}_r \in D_r)\right\} \geq 1 - \alpha.$$

Альтернативной процедурой для метода Бонферрони является коррекция достигаемых p-значений p_r, $r = 1, \ldots, N$ t-статистик соответствующих N ПДГ по формуле $\widetilde{p}_{(r)} = \min(1, Np_r)$.

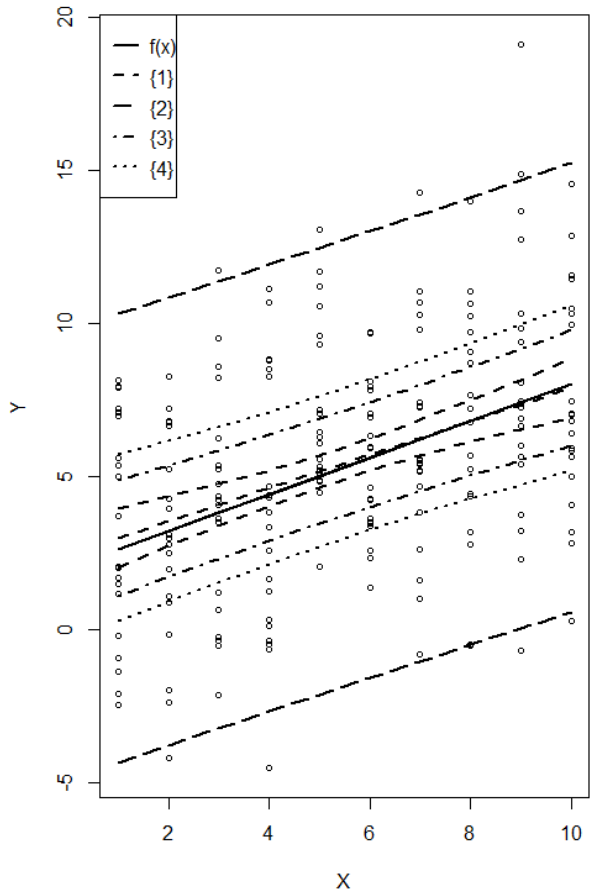
Методы коррекции

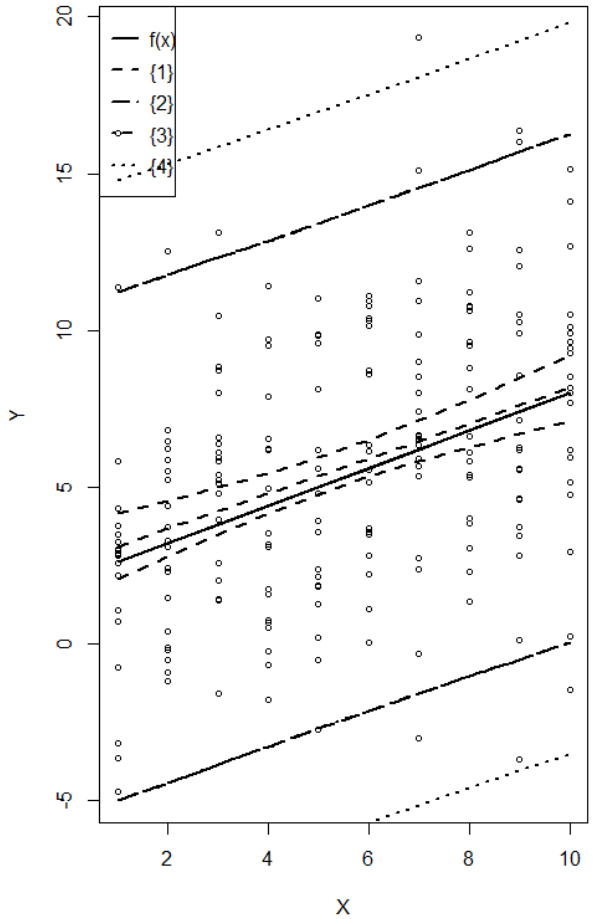
Метод Шидака (*Šidàk*), основанный на справедливости неравенства $\alpha/N \leq 1-(1-\alpha)^{1/N}$ при $N \geq 1$. Для этого метода полагают $\alpha_r = 1-(1-\alpha)^{1/N}, r = 1, \ldots, N$.
Метод Холма (Holm) — $\alpha_r = \alpha/(N-r+1), r = 1, \ldots, N$ или $\widetilde{p}_{(r)} = \min(1, \max((N-r+1)p_{(r)}, \widetilde{p}_{(r-1)}))$, где $p_{(1)} \leq \cdots \leq p_{(N)}$ — упорядоченные достигаемые p-значения t-статистик соответствующих ПДГ.
Метод Бенджамина-Хохберга (Benjamini- Hochberg) — $\alpha_r = \alpha r/N, r = 1, \ldots, N$ или $\widetilde{p}_{(r)} = \min(1, \max(Np_{(r)}/r, \widetilde{p}_{(r-1)}))$;
Метод Бенджамина-Иекутиели (Benjamini-Yekateuli) — $\alpha_r = \alpha r/(Nc), r = 1, \ldots, N$ или $\widetilde{p}_{(r)} = \min(1, \max(Np_{(r)}c/r, \widetilde{p}_{(r-1)}))$, где $c = \sum_{i=1}^{N} \frac{1}{i}$.

Численное моделирование

$f(x) = 0.5x + 2,\ x = 1, \ldots, l,\ f(x_i),\ i = 1, \ldots, l,\ l = 10, q = 20$

$y_{ij} = f(x_i) + \varepsilon_{ij},\ i = 1, \ldots, l,\ j = 1, \ldots, q,\ \mathfrak{L}(\epsilon) = \mathcal{N}_1(0, 4)$





Постановка задачи (общая)

Наша задача заключается в построении совместной доверительной полосы для среднего повторных откликов вида

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

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 проблема состоит в нахождении критической константы c, определяемой как $P\{T < c\}$, такой, чтобы доверительная полоса (3) имела уровень $1 - \alpha$, где

$$T = \sup_{x_i \in [a_i, b_i], i=1,\ldots,k} \frac{|\hat{y} - \overline{y}_m|}{\hat{\sigma}\sqrt{\frac{1}{m} + \mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}}. \tag{4}$$

Метод вычисления c

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

$$T = \sup_{x_i \in [a_i, b_i], i=1,\ldots,k} \frac{|\mathbf{x}^T(\hat{\beta} - \beta)|}{\hat{\sigma}\sqrt{\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}}}. \tag{5}$$

Для решения последней оптимизационной задачи существует множество похожих подходов $[3-5]$

Метод вычисления c

Представим величину T в виде

$$T = Q\frac{\|\mathbf{Z}\|}{(\hat{\sigma}/\sigma)}, Q = \sup_{x_i\in[a_i,b_i],i=1,\ldots,k} \frac{|(\mathbf{Px})^T\mathbf{Z}|}{\|\mathbf{Px}\|\|\mathbf{Z}\|}, \quad (6)$$

где $(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{P}^T\mathbf{P}$, $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_k) \in R^{k\times k}$, $\mathbf{Z} = (\mathbf{P}^T)^{-1}(\hat{\beta} - \beta)/\sigma \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k)$. 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 (6). Основная трудность расчета T заключается в вычислении Q. Величина Q может быть получена посредством решения задачи

$$Q = \sup_{\mathbf{s}\in\Omega} \frac{|\mathbf{s}^T\mathbf{Z}|}{\|\mathbf{s}\|\|\mathbf{Z}\|}, \quad (7)$$

где $\Omega = \{\mathbf{s} : \mathbf{s} = \gamma\boldsymbol{\nu}, \boldsymbol{\nu} \in L, \gamma > 0\}$, $L = \{\mathbf{Px} : x_i \in [a_i, b_i], i = 1, \ldots, k\}$. 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 (7), то оно также является решением

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

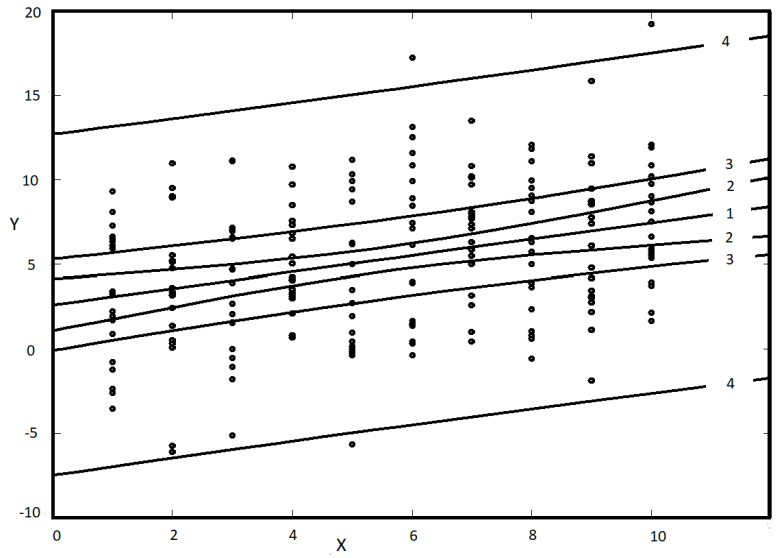
Метод вычисления c

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

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

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 $y_{ij} = f(x_i) + \varepsilon_{ij}$, $i = 1, \ldots, l$, $j = 1, \ldots, q$, $l = 10$, $q = 20$, изображенных в виде кружков на рис. 12. При этом каждому x_i соответствует q повторяющихся наблюдений.



Численное моделирование

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: **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
.	....	....	...
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

Численное моделирование

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 узки, чем совместные доверительные границы, полученные менее точным методом коррекции Бонферрони [2] (для данного примера $c = t_{1-\frac{\alpha}{2l},\,n-k} = t_{0.9975,\,198} = 2,839$).

Двухфакторный пример

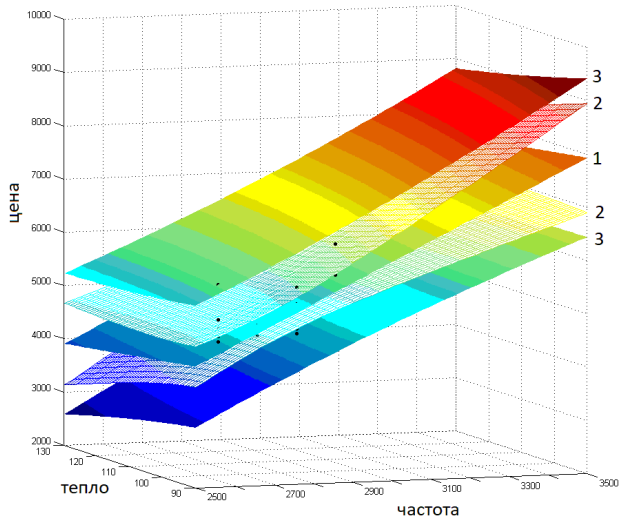
Для двухфакторной модели $k = 2$ рассмотрим выборку $n = 35$ цен на шестиядерные процессоры серии Phenom 2 фирмы AMD, различающиеся рабочей частотой (МГц) и тепловыделением (Вт) (см. табл. 2) (данные получены из интернет-ресурса http://market.yandex.ru/).

Таблица: **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
.	...	...	...	...	...	...	...
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				

Двухфакторный пример

Для этих данных проведены расчеты доверительных полос среднего значения повторных откликов ($m = 5$) и наблюдения ($m = 1$), которые представлены на рис. 16. 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 вычислений записываются в отдельный файл и могут быть представлены графически для моделей до двух предикторов.

Заключение

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

Литература

- Белов А. Г. Доверительное прогнозирование среднего значения повторных наблюдений // Вестник Моск. ун-та, Сер.15, Вычислительная математика и кибернетика. 2016. **2**. С.14-19.
- Bonferroni C. E. Il calcolo delle assi curazioni su gruppi di test // In Studi Onore del Professore Salvatore Ortu Carboni. Rome: Italy, 1935, P.13-60.
- Naiman D. Q. Simultaneous confidence-bounds in multiple-regression using predictor variable constraints // J. the Amer. Stat. Assoc. 1987. **82**. P.214–219.
- Liu W., Jamshidian M., Zhang Y. Multiple comparison of several linear regression lines // J. the Amer. Stat. Assoc. 2004. **99**. P.395–403.
- 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**. N2. P.459–484.

Конец

Спасибо за внимание!