ISSN 0278-6419, Moscow University Computational Mathematics and Cybernetics, 2018, Vol. 42, No. 3, pp. 114–118. © Allerton Press, Inc., 2018.
Original Russian Text © T.V. Zakharova, A.A. Fisak, 2018, published in Vestnik Moskovskogo Universiteta, Seriya 15: Vychislitel'naya Matematika i Kibernetika, 2018, No. 3, pp. 19–23.

Modeling a Simultaneous Confidence Band of the Mean Value of Multiple Responses with a Rectangular Domain for Predictors

A. G. Belov*

Department of Computational Mathematics and Cybernetics, Moscow State University, Moscow, 119991 Russia

Received Apri 2, 2018

Abstract—The problem is considered of modeling simultaneous confidence intervals for the mean values of multiple responses in a linear multivariate normal regression model with predictor variables defined in intervals. To solve it, a numerical way of calculating the critical value that determines the simultaneous confidence interval of a given level is used. Simultaneous confidence intervals are numerically modelled and analyzed by comparison for regression, the mean value of multiple responses, and individual observation.

Keywords: simultaneous confidence intervals, normal regression, multiple responses.

DOI: 10.3103/S0278641918030020

1. STATEMENT OF THE PROBLEM

Let us consider the linear multivariate normal regression model of observations

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

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is a column vector of random variables y_i of responses that describe the results of the ith experiment, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is a column vector of random "errors" following the normal distribution law $\mathfrak{L}(\boldsymbol{\varepsilon}) = \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ and not depending on the vector of parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$; and $\mathbf{X} = \left|\left|\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}\right|\right| \in R^{n\times k}$ is the regression matrix of column vectors $\mathbf{x}^{(j)} = (x_{1j}, \ldots, x_{nj})^T$ that affect only mean response Ey_i; we assume that $\mathbf{I}_n = \mathrm{diag}(1, \ldots, 1) \in R^{n\times n}$ and $\mathrm{rank}\mathbf{X} = k$, $k \leqslant n$.

We assume that given m multiple observations $\mathbf{y}_m = (y_1, \ldots, y_m)^T$ corresponding to fixed regressor values $\mathbf{x} = (x_1, \ldots, x_k)^T$ so that $y_j = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon_j$, $1 \leqslant j \leqslant m$, where the column vector of random "errors" $\boldsymbol{\varepsilon}_m = (\varepsilon_1, \ldots, \varepsilon_m)^T$ is independent of $\boldsymbol{\varepsilon}$, $\mathfrak{L}(\boldsymbol{\varepsilon}_m) = \mathcal{N}_m(\mathbf{0}, \sigma^2\mathbf{I}_m)$, and $\mathfrak{L}(\boldsymbol{\varepsilon}_{m0}) = \mathcal{N}_m(\mathbf{0}, \mathbf{I}_m)$ for $\boldsymbol{\varepsilon}_{m0} = \boldsymbol{\varepsilon}_m/\sigma$.

For mean value $\overline{y}_m = \mathbf{e}_m^T\mathbf{y}_m/m = \mathbf{x}^T\boldsymbol{\beta} + \sigma\mathbf{e}_m^T\boldsymbol{\varepsilon}_{m0}/m$ of multiple responses, where $\mathbf{e}_m = (1, \ldots, 1)^T \in R^m$, we use $100(1-\alpha)$-percent confidence pointwise interval [1]

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

where $\mathbf{A} = \mathbf{X}^T\mathbf{X}$, $\hat{y} = \mathbf{x}^T\hat{\boldsymbol{\beta}}$ is the estimator of response y for $\mathbf{x}$, $\hat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{X}^T\mathbf{y} = \boldsymbol{\beta} + \sigma\mathbf{A}^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}_0$ is the estimator of the parameter vector $\boldsymbol{\beta}$ obtained from a sample $\mathbf{y}$ by ordinary least-square (OLS) means, $\hat{\sigma}^2 = S(\hat{\boldsymbol{\beta}})/(n-k)$ is the estimator of σ^2, and $t_{1-\frac{\alpha}{2},n-k}$ is the $100(1-\frac{\alpha}{2})$-percent quantile of the Student distribution $St(n-k)$, so that

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

*E-mail: belov@cs.msu.ru.

$$S(\hat{\boldsymbol{\beta}}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad \boldsymbol{\varepsilon}_0 = \boldsymbol{\varepsilon}/\sigma, \quad \mathfrak{L}(\boldsymbol{\varepsilon}_0) = \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n).$$

From (1), known $100(1-\alpha)$-percent confidence pointwise intervals follow for regression $\mathbf{x}^T\boldsymbol{\beta}$ (as $m \to \infty$) and individual response $y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon_1$ (for $m = 1$), respectively:

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

where $\varepsilon_1 \sim \mathcal{N}_1(0, \sigma^2)$ and does not depend on $\boldsymbol{\varepsilon}$.

We thus have common critical constant $c = t_{1-\frac{\alpha}{2},n-k}$ for all three pointwise confidence intervals such that it determines confidence level $1-\alpha$.

The aim of this work is to construct a simultaneous confidence band for the mean of multiple responses in the form

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

and therefore a simultaneous confidence band for the regression and individual response, respectively:

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

where D is a rectangular domain defined as

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

The main problem is to find critical constant c given by $P\{T < c\}$ such that confidence bands (3), and thus (4), have the level $1-\alpha$, where

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

2. CALCULATING c

Since required critical constant c determines confidence bands (3) and (4), it suffices to know how to calculate it for any of these bands, particularly for regression $\mathbf{x}^T\boldsymbol{\beta}$. Constant c is in this case defined as $P\{T < c\}$, and we find from (5) that

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

There are a number of similar approaches [2–4] to solving the latter optimization problem.

We represent variable T in the form

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

$$(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{P}^T\mathbf{P}, \quad \mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_k) \in R^{k \times k}, \quad \mathbf{Z} = (\mathbf{P}^T)^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k). \tag{7}$$

Since it is difficult to find a formula for the distribution of T, we must simulate it by generating random variables $\mathbf{Z}$ and $\hat{\sigma}/\sigma \sim \sqrt{\chi^2_{n-k}/(n-k)}$ and then substituting them into (7). The main difficulty in calculating T is finding Q. The value of Q can be found by solving the problem

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

where $\Omega = \{\mathbf{s} : \mathbf{s} = \gamma\boldsymbol{\nu} \text{ and } \boldsymbol{\nu} \in L, \gamma > 0\}$, $L = \{\mathbf{P}\mathbf{x} : x_i \in [a_i, b_i],\ i = 1, \ldots, k\}$. It is easy to see that $\mathbf{s}^T\mathbf{Z}/(||\mathbf{s}||||\mathbf{Z}||)$ is the cosine of the angle between $\mathbf{s}$ and $\mathbf{Z}$. If $\widehat{\mathbf{s}} \in \Omega$ solves (8), it is therefore also the solution to

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

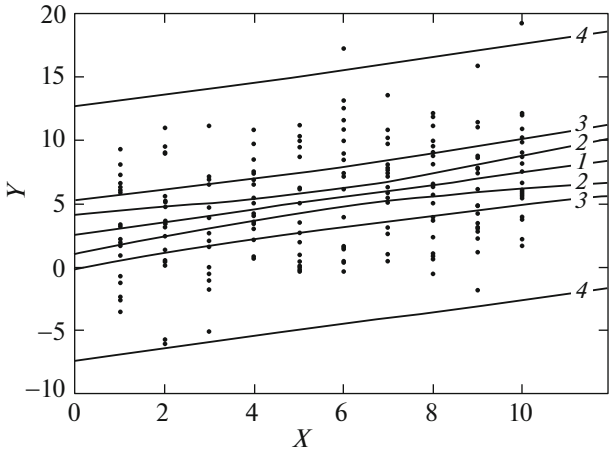

Fig. 1. Confidence bands: regression function (line *1*); 95-percent confidence band for regression (lines *2*); 95-percent simultaneous limits for the mean value of multiple responses ($m = q$, lines *3*); 95-percent band of individual observations ($m = 1$, lines *4*) for $x = 1, \ldots, 10$, $n = 200$, and $q = 20$.

To solve this quadratic programming problem below, we use the "active set" algorithm described in detail in [4], since it is the most effective algorithm that converges in a finite number of steps.

Critical constant c can thus be determined as follows: Sufficiently large number M of values T_i of random variable T is simulated. The $(1-\alpha)M$th highest value $\widehat{c}$ from the generated variation series is then considered an estimate of c. This approach is based on sample $100(1-\alpha)$th percentile $\widehat{c}$ converging almost certainly to theoretical $100(1-\alpha)$th percentile c as $M \to \infty$. Using the asymptotic normality of $\widehat{c}$ with mean c and standard error $s = \sqrt{\dfrac{\alpha(1-\alpha)}{g^2(c)M}}$, we can now calculate the standard error of an estimate for $\widehat{c}$, where h is a smoothing parameter ($h = 0,01$ in the calculations below) and $g(c)$ is the density function of the distribution of random variable T, which can be estimated as

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

Table 1. Simulation data

n	Generation number	$\widehat{c}$	$s(\widehat{c})$
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

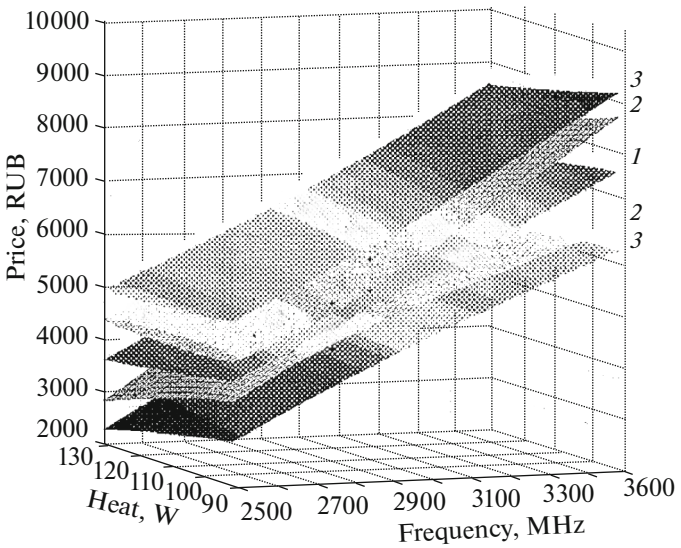

Fig. 2. Simultaneous confidence limits: estimate of regression (plane *1*); 95-percent confidence band for the mean value of multiple responses ($m = 5$, planes *2*) and for an individual observation ($m = 1$, planes *3*).

3. NUMERICAL SIMULATION

We now compute confidence bands for simple regression on model data. To do so, we choose $l = 10$ natural values of regressor $x = 1, \ldots, l$ of linear dependence $f(x) = 0.5x + 2$. For each $f(x_i)$, $i = 1, \ldots, l$, we then independently simulate q random variables y_{ij} by additively inserting into $f(x_i)$ normally distributed random error $\mathfrak{L}(\epsilon) = \mathcal{N}_1(0, 4)$ with dispersion $\sigma^2 = 4$. As a result, we obtain a cloud of $n = lq$ values $y_{ij} = f(x_i) + \varepsilon_{ij}$ for $1 \leqslant i \leqslant l$, $1 \leqslant j \leqslant q$, $l = 10$, and $q = 20$, which are shown in the form of circles in Fig. 1. Here, each x_i corresponds to q multiple observations.

Figure 1 shows the regression function (line *1*), the 95-percent confidence band for regression (lines *2*), the 95-percent simultaneous limits for the mean value of multiple responses for $m = q$ (lines *3*) and for individual observations of $m = 1$ (lines *4*). To simulate T up to 30000, we use generations and critical value $\widehat{c}$, along with its standard error $s(\widehat{c})$. Table 1 presents the intermediate calculation results.

Comparing $c = t_{1-\frac{\alpha}{2},n-k} = t_{0.975,198} = 1.972$ with value $\widehat{c} = 2.4155$ for $M = 30000$ (see Table 1), we may conclude that the width of pointwise confidence limits (1) is less than the corresponding simulated simultaneous bands. However, the latter bands are narrower than the simultaneous confidence limits obtained using the Bonferroni correction [5], which is a less accurate way of doing so (for the given example, $c = t_{1-\frac{\alpha}{2l},n-k} = t_{0.9975,198} = 2.839$).

4. EXAMPLE

For a two-factor model $k = 2$, we consider a sample of $n = 35$ prices for six-core processors of the Phenom 2 series from AMD that differ by operation frequency (MHz) and heat dissipation (W), see Table 2 (the data were obtained from Internet resource http://market.yandex.ru/).

Based on these data, we performed calculations of confidence bands for the main value of muptiple responses ($m = 5$) and observation ($m = 1$), which are presented in Fig. 2. For generation number $M = 30000$, we found $\widehat{c} = 2.9093$, $s(\widehat{c}) = 0.0137$.

The above calculations were performed using the author's SSB (Simulation Simultaneous Bands) program written in the MatLab environment, version 7.0.5. The program includes an interface for importing data and setting the desired simulation parameters. The results from calculations are written to a separate file and can be represented graphically for models with one or two predictors.

Table 2. Data on AMD processors

n	Frequency, MHz	Heat, W	Price, RUB	n	Frequency, MHz	Heat, W	Price, RUB
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				

5. CONCLUSION

We have considered a numerical way of calculating a confidence band for the mean value of multiple responses in a linear multivariate normal regression model with a rectangular domain for predictors. We performed a numerical simulation of the critical value with the related calculation of the confidence band for the mean value of multiple responses, regression, and response. Finally, we performed a comparative analysis of the calculated bands.

REFERENCES

1. A. G. Belov, "Confidence prediction of the mean values of multiple observations," Moscow Univ. Comput. Math. Cybern. **36**, 65–70 (2016).
2. D. Q. Naiman, "Simultaneous confidence-bounds in multiple-regression using predictor variable constraints," J. Am. Stat. Assoc. **82**, 214–219 (1987).
3. W. Liu, M. Jamshidian, and Y. Zhang, "Multiple comparison of several linear regression lines," J. Am. Stat. Assoc. **99**, 395–403 (2004).
4. W. Liu, M. Jamshidian, Y. Zhang, and J. Donnelly, "Simulation-based simultaneous confidence bands in multiple linear regression with predictor variables constrained in intervals," J. Comput. Graph. Stat. **14**, 459–484 (2005).
5. C. E. Bonferroni, "Il calcolo delle assi curazioni su gruppi di test," in *Studi Onore del Professore Salvatore Ortu Carboni* (Rome, Italy, 1935), pp. 13–60.

Translated by I. Tselishcheva