УДК 90.03.03, 90.03.37, 389.63:378.14

ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ И ОБЛАСТЕЙ ДЛЯ МОДЕЛИ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ С ИСПОЛЬЗОВАНИЕМ БАЙЕСОВСКОГО ПОДХОДА

Р.З. Хайруллин[1,2]

[1]Национальный исследовательский Московский государственный строительный университет, Москва, Россия,
[2]ФГБУ «ГНМЦ» Минобороны России, Мытищи, Россия,
zrkzrk@list.ru

Аннотация. Представлен алгоритм построения точечных и интервальных статистических оценок для параметров модели множественной линейной регрессии. Представлены результаты сравнения с соответствующими оценками, полученными классическим методом максимального правдоподобия. Предложенный алгоритм может быть эффективно применён при разработке методик СВЧ-измерений на векторных анализаторах цепей, при разработке практических способов выявления систематической погрешности измерений, при коррекции результатов измерений.

Ключевые слова: достоверность измерений, байесовский научный подход, апостериорная информация.

CONSTRUCTING THE CONFIDENCE INTERVALS AND AREAS FOR A MULTIPLE LINEAR REGRESSION MODEL USING BAYESIAN SCIENTIFIC APPROACH

R.Z. Khayrullin[1,2]

[1]Moscow State (National Research) University of Civil Engineering, Moscow, Russia,
[2]FSBI «MSMC» of the Ministry of Defense of the Russian Federation, Moscow, Russia,
zrkzrk@list.ru

Annotation. An algorithm for constructing point and interval statistical estimates for the parameters of the multiple linear regression model is presented. The results of comparison with the corresponding estimates obtained by the classical maximum likelihood method are presented. The proposed algorithm can be effectively applied in the development of microwave measurement techniques on vector network analyzers, in the development of practical methods for detecting systematic measurement errors, in correcting measurement results.

Key words: measurement accuracy, Bayesian scientific approach, a posteriori information.

Введение

В современной практике СВЧ-измерений, проводимых с применением векторных анализаторов цепей (ВАЦ), используется шестиэтапная схема: оценочные измерения; оптимизация параметров измерительной системы в целях получения наилучших результатов для данного устройства; калибровка

в целях определения собственных параметров измерительной системы и исключения некоторых составляющих систематической погрешности; сам процесс измерения; обработка результатов измерения, включающая коррекцию систематической погрешности измерений; сохранение результатов измерений [1]. Встроенные в ВАЦ программируемые вычислители и входящее в ВАЦ программное обеспечение позволяют в автоматическом и полуавтоматическом режимах производить обработку измерительной информации, вносить математическую коррекцию результатов измерений с целью получения наиболее достоверного результата. Южно-Уральский приборостроительный кластер «ПЛАНАР» занимается разработкой и серийным производством анализаторов цепей (С1220, С1420, С2220, С2240), которые могут работать с внешним компьютером и на которых могут быть реализованы специализированные алгоритмы для обработки результатов измерений. Поэтому разработка специализированных алгоритмов, предназначенных для статистической обработки результатов измерений, является важной практической задачей.

Построение эффективных статистических оценок является актуальной задачей для многих областей научно-практической деятельности [2–5], в том числе в метрологии, при разработке методик измерений, при обработке и интерпретации результатов испытаний измерительной техники [6–10]. В [11–13] описано применение байесовского подхода (БП) в задачах построения статистических оценок для случаев, когда исследуемая генеральная совокупность подчиняется нормальному закону распределения, распределению Пуассона, экспоненциальному закону распределения, равномерному закону и закону Парето.

Предлагаемая статья посвящена применению БП к исследованию модели множественной линейной регрессии (ММЛР). БП основан на комплексном использовании статистических данных и априорной информации об исследуемом объекте или процессе. Отметим, что в стандартном программном обеспечении ВАЦ «ПЛАНАР» уже есть реализованные регрессионные методы. Например, функция локально-полиномиальной регрессии LOESS используется при измерении комплексного коэффициента передачи и отражения.

Метод решения задачи

Основные положения байесовского подхода

Пусть в описании ММЛР присутствует s-мерный параметр $\mathbf{\Theta} = (\theta_1, \theta_2, \ldots, \theta_S)^T$. Верхний индекс T здесь и в дальнейшем означает операцию транспонирования вектора или матрицы. Прописными буквами будем обозначать матричные и векторные величины, а строчные буквы будем использовать для обозначения одномерных величин. БП позволяет построить рациональную (наилучшую в некотором смысле) статистическую оценку $\hat{\mathbf{\Theta}}$ этого s-мерного параметра по имеющимся наблюдениям $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.

БП является одним из возможных способов формализации тезиса: чем больше объём выборки n, на основании которой мы строим свою оценку $\hat{\boldsymbol{\Theta}}$, тем большей информацией об этом параметре мы располагаем и тем ближе оценка $\hat{\boldsymbol{\Theta}}$ к истинному значению $\boldsymbol{\Theta}$ в смысле сходимости по вероятности.

Априорные сведения о параметре $\boldsymbol{\Theta}$ обычно формируются на основе анализа предыстории функционирования процесса (если таковая имеется), на профессиональных теоретических соображениях о сущности процесса и его специфики. Вначале эти априорные сведения должны быть представлены в виде функции плотности распределения $\rho(\boldsymbol{\Theta}, \boldsymbol{\Pi})$, зависящей, в общем случае, от векторного параметра $\boldsymbol{\Pi} = (\pi_1, \pi_2, \ldots, \pi_\ell)$, определяющего конкретный вид функции.

Так, например, для нормального закона распределения с неизвестным средним значением и известной дисперсией, неизвестное среднее значение представляет собой случайную величину, имеющую нормальное распределение. Среднее значение и дисперсия апостериорного закона распределения являются средневзвешенными значениями априорных и выборочных значений среднего и дисперсии соответственно [12, 14].

Для нормального распределения с неизвестным средним значением и неизвестной дисперсией указанные характеристики будут представлять собой случайные величины, распределённые по закону двумерного гамма-нормального распределения [12, 14].

Если исследуемая выборка распределена по закону Пуассона, то параметр Пуассона будет иметь гамма-распределение [13, 14].

Для ММЛР коэффициенты регрессионного уравнения будут случайными величинами, имеющими многомерное гамма-нормальное распределение [14, 15].

Пусть данные $X_1, X_2, \ldots, X_n$ порождаются в соответствии с законом распределения вероятностей $f(X_1, X_2, \ldots, X_n|\boldsymbol{\Theta})$. Предполагается, что $X_1, X_2, \ldots, X_n$ при фиксированном $\boldsymbol{\Theta}$ являются статистически взаимно независимыми (образуют случайную выборку из анализируемой генеральной совокупности). Условная функция правдоподобия имеет вид [14–17]:

$$L(X_1, X_2, ..., X_n|\boldsymbol{\Theta}) = f(X_1|\boldsymbol{\Theta}) \cdot f(X_2|\boldsymbol{\Theta}) \cdot f(X_n|\boldsymbol{\Theta}). \quad (1)$$

Вычисление апостериорного распределения $\tilde{\rho}(\boldsymbol{\Theta}|X_1, X_2, ..., X_n)$ осуществляется с помощью формулы Байеса:

$$\tilde{\rho}(\boldsymbol{\Theta}|X_1, X_2, ..., X_n) = \frac{\rho(\boldsymbol{\Theta}) \cdot L(X_1, X_2, ..., X_n|\boldsymbol{\Theta})}{\int L(X_1, X_2, ..., X_n|\boldsymbol{\Theta}) \cdot \rho(\boldsymbol{\Theta}) d\boldsymbol{\Theta}}. \quad (2)$$

В качестве Байесовских точечных оценок $\hat{\boldsymbol{\Theta}}^{(Б)}$ обычно используют среднее значение, рассчитанное с использованием функции плотности распределения (2) [14, 15].

Отметим, что для определения общего вида $\tilde{\rho}(\boldsymbol{\Theta}|X_1, X_2, ..., X_n)$ достаточно знать только числитель правой части (2), так как знаменатель этого выражения играет роль нормирующего множителя и от $\boldsymbol{\Theta}$ не зависит. Это существенно упрощает процесс практического построения точечных оценок и доверительных интервалов.

Для построения доверительного интервала для параметра $\boldsymbol{\Theta}$ на основе БП необходимо по заданной доверительной вероятности P_0 определить $50 \cdot (1 + P_0)$%- и $50 \cdot (1 - P_0)$%-е точки закона распределения (2). Эти точки и соответствуют границам доверительного интервала.

Процесс реализации и алгоритмизации пересчёта по формуле (2) значительно упрощается, если использовать распределения, сопряжённые с наблюдаемой генеральной совокупностью (распределения, сопряжённые с функцией правдоподобия (1)).

Понятие об априорных распределениях, сопряжённых с наблюдаемой генеральной совокупностью (функцией правдоподобия)

Семейство априорных распределений $\mathbf{G} = \{\rho(\boldsymbol{\Theta}); \boldsymbol{\Pi}\}$ называется сопряжённым по отношению к наблюдаемой генеральной совокупности $f(X_1, X_2, \ldots, X_n|\boldsymbol{\Theta})$ (к функции правдоподобия $L(X_1, X_2, \ldots, X_n|\boldsymbol{\Theta})$ (1)), если и апостериорное распределение $\tilde{\rho}(\boldsymbol{\Theta}|X_1, X_2, ..., X_n)$, вычисленное по формуле (2), снова принадлежит этому же семейству $\mathbf{G} = \{\rho(\boldsymbol{\Theta}); \boldsymbol{\Pi}\}$. Другими словами, семейство распределений $\mathbf{G}$ сопряжено с $L(X_1, X_2, ..., X_n|\boldsymbol{\Theta})$, если оно замкнуто относительно операции (2) пересчёта априорного распределения в апостериорное.

Поэтому использование в качестве априорных законов распределения сопряжённых по отношению к (1) плотностей вероятностей приводит только к необходимости пересчёта значений его параметров $\boldsymbol{\Pi}$ при переходе от априорного распределения к апостериорному.

Однако сопряжённые распределения существуют не всегда. Сформулируем в соответствии с [11–15] достаточные условия существования сопряжённых априорных распределений.

Если функция правдоподобия $L(X_1, X_2, \ldots, X_n|\boldsymbol{\Theta})$ представима в форме произведения двух функций Ψ и ν:

$$L(X_1, X_2, ..., X_n|\boldsymbol{\Theta}) = \\ = \nu(T_1(X_1, X_2, ..., X_n), ..., T_m(X_1, X_2, ..., X_n); \boldsymbol{\Theta}) \cdot \Psi(X_1, X_2, ..., X_n), \tag{3}$$

причём $\Psi(X_1, X_2, \ldots, X_n)$ — некоторая функция от $X_1, X_2, \ldots, X_n$, не зависящая от параметра $\boldsymbol{\Theta}$, а функция ν зависит как от параметра $\boldsymbol{\Theta}$, так и от m функций $\{T_j(X_1, X_2, \ldots, X_n), (j = 1, 2, \ldots, m)\}$, называемых статистиками [14–15], то существует семейство $\mathbf{G} = \{\rho(\boldsymbol{\Theta}); \boldsymbol{\Pi}\}$ априорных распределений, сопряжённое с функцией правдоподобия $L(X_1, X_2, \ldots, X_n|\boldsymbol{\Theta})$ (1).

Построение и исследование модели множественной линейной регрессии

Построение модели

Рассмотрим модель с нормальными, в среднем нулевыми, взаимно независимыми и гомоскедастичными остатками [14–17]:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \boldsymbol{\varepsilon}, \tag{4}$$

где $\mathbf{Y}$ — наблюдённые значения зависимой переменной,

$$\mathbf{X} = \begin{pmatrix} 1 & X_1^{(1)} & \ldots & X_k^{(1)} \\ 1 & X_1^{(2)} & \ldots & X_k^{(2)} \\ 1 & \ldots & \ldots & \ldots \\ 1 & X_1^{(n)} & \ldots & X_k^{(n)} \end{pmatrix}$$

— наблюдённые значения объясняющих переменных $\mathbf{X}$; $\boldsymbol{\Theta} = (\theta_0, \theta_1, \ldots, \theta_k)^T$ и $h = (D\boldsymbol{\varepsilon})^{-1}$ — неизвестные параметры модели. Напомним, что регрессионные остатки должны подчиняться нормальному закону распределения:

$$\boldsymbol{\varepsilon} \in N_n(\mathbf{0}; \boldsymbol{\Sigma}_\varepsilon),$$

где $\mathbf{I}_n$ — единичная матрица размерности n, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \frac{1}{h} \cdot \mathbf{I}_n$ — ковариационная матрица остатков.

Проверка условия существования сопряжённого семейства априорных распределений для регрессионной модели

Для модели (4) функция правдоподобия наблюдений ($\mathbf{X}$, $\mathbf{Y}$) может быть представлена в форме [14–17]:

$$L(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Theta}; h) = \frac{h^{n/2}}{(2\pi)^{n/2}} \exp\left(-\frac{h}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta})\right). \tag{5}$$

Преобразуя аргумент экспоненты, добавив и вычтя $\mathbf{X}\hat{\boldsymbol{\Theta}}$, где $\hat{\boldsymbol{\Theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ — оценка МНК параметров регрессии $\boldsymbol{\Theta}$, получим [14–15]:

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\Theta}) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Theta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\Theta}}) + (\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}).$$

Выражая сумму квадратов оценённых остатков через оценку остаточной дисперсии по МНК $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Theta}})^T (\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Theta}})/(n-k-1)$, представим функцию правдоподобия в виде:

$$L(\mathbf{X},\mathbf{Y}|\mathbf{\Theta};h) = \frac{h^{n/2}}{(2\pi)^{n/2}} \exp\left(-\left(\frac{n-k-1}{2}\hat{\sigma}^2\right)\right) \cdot h \cdot \exp\left(-\frac{h}{2}(\hat{\mathbf{\Theta}} - \mathbf{\Theta})^T \mathbf{X}^T \mathbf{X}(\hat{\mathbf{\Theta}} - \mathbf{\Theta})\right). \quad (6)$$

Отметим, что $\hat{\sigma}^2$ и $\hat{\mathbf{\Theta}}$ в конечном счёте определяются по $\mathbf{Y^T Y}$, $\mathbf{X}^T\mathbf{Y}$ и $\mathbf{X}^T\mathbf{X}$. Так что функция правдоподобия (6) представима в форме (3). Следовательно, существует априорное распределение параметров $\mathbf{\Theta}$ и h, сопряжённое с (1).

Определение общего вида сопряжённого априорного распределения

В [14, 15] показано, что если изначально вид функции плотности распределения неизвестен, то его можно достаточно просто определить. Нужно в качестве первого приближения задать или $\rho(\theta_i)$ = const, если параметр θ_i может принимать как положительные, так и отрицательные значения; или $\rho(\ln\theta_i)$ = const, если параметр θ_i может принимать только положительные значения. После совершения одного шага пересчёта функции правдоподобия с использованием статистических данных X_1, X_2, …, X_n мы сразу получим общий вид сопряжённого распределения (гамма-нормального распределения для ММЛР). Отметим, что нарушение условия нормировки на функцию плотности распределения при задании первого приближения не приводит к каким-либо вычислительным трудностям.

Используя выражение (6) для функции правдоподобия $L(\mathbf{X};\mathbf{Y}|\mathbf{\Theta};h)$, запишем функцию плотности распределения в виде:

$$\begin{aligned} \rho(\mathbf{\Theta};h|\mathbf{X};\mathbf{Y}) &= \rho(\mathbf{\Theta};h)\cdot L(\mathbf{X};\mathbf{Y}|\mathbf{\Theta};h) \approx \\ \approx \frac{1}{h} h^{n/2} \exp\left(-\frac{n-k-1}{2}\sigma^2 h\right) &\exp\left(-\frac{h}{2}(\hat{\mathbf{\Theta}} - \mathbf{\Theta})^T (\mathbf{X}^T\mathbf{X})(\hat{\mathbf{\Theta}} - \mathbf{\Theta})\right) = \\ = h^{\frac{n-k-1}{2}} \exp\left(-\frac{n-k-1}{2}\sigma^2 h\right) &\cdot h^{\frac{k+1}{2}} \exp\left(-\frac{h}{2}(\hat{\mathbf{\Theta}} - \mathbf{\Theta})^T (\mathbf{X}^T\mathbf{X})(\hat{\mathbf{\Theta}} - \mathbf{\Theta})\right). \end{aligned} \quad (7)$$

Правая часть (7) определяет (с точностью до нормирующего множителя, не зависящего от $\mathbf{\Theta}$ и h) так называемое многомерное гамма-нормальное распределение [14, 15] с параметром сдвига $\hat{\mathbf{\Theta}}$, матрицей точности $(\mathbf{X}^T\mathbf{X})$ и параметрами $\alpha = (n-k-1)/2$ и $\beta = \hat{\sigma}^2 \cdot (n-k-1)/2$. Здесь и в дальнейшем знак приближенного равенства (≈) означает равенство с точностью до нормирующего множителя.

Сформулируем свойства многомерного гамма-нормального распределения [14, 15], которое будет использовано при построении статистических оценок.

Свойство 1

Частное распределение векторного параметра $\mathbf{\Theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T$ есть многомерное обобщённое распределение Стьюдента с 2α степенями свободы, параметром сдвига $\mathbf{\Theta}_0 = (\theta_1^0, \theta_2^0, ..., \theta_k^0)^T$ и матрицей точности $\mathbf{B} = \frac{\alpha}{\beta} \cdot \mathbf{\Lambda_0}$.

Свойство 2

Частное распределение скалярного параметра h есть гамма-распределение с параметрами α и β.

Свойство 3

Условное распределение векторного параметра $\mathbf{\Theta}$ при условии, что параметр h задан: $h = h_0$ является k-мерным нормальным распределением $N(\mathbf{\Theta}_0; (h_0 \mathbf{\Lambda_0})^{-1})$.

Таким образом, для ММЛР компонентами векторного параметра $\mathbf{\Pi}$ являются: α, β, $\mathbf{\Lambda_0}$, $\mathbf{\Theta_0}$.

Таким образом, сопряжённые априорные распределения параметров ($\mathbf{\Theta}$, h) для ММЛР имеют общий вид:

$$\rho(\mathbf{\Theta}; h) \approx h^{(k+1)/2} \left|\Lambda_0\right|^{\frac{1}{2}} \exp\left(-\frac{h}{2}(\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0)^T \mathbf{\Lambda}_0 (\hat{\mathbf{\Theta}} - \mathbf{\Theta}_0)\right) h^{\alpha-1} \exp(-\beta h), \qquad (8)$$

в котором конкретное задание векторного параметра сдвига Θ_0, матрицы точности $\mathbf{\Lambda}_0$, имеющей размерность $(k + 1) \cdot (k + 1)$, и скалярных параметров α и β однозначно определяют априорный закон распределения векторного параметра $\mathbf{\Theta}$ и скалярного параметра h.

Алгоритм расчёта параметров многомерного гамма-нормального распределения

С учётом свойств многомерного гамма-нормального распределения [14, 15]:

$$E\mathbf{\Theta} = Et\left(2\alpha \middle| \theta_0; \Lambda_0 \frac{\alpha}{\beta}\right) = \mathbf{\Theta}_0;$$

$$\Sigma_{\mathbf{\Theta}} = \Sigma_{t\left(2\alpha|\theta_0;\frac{\alpha}{\beta}\Lambda_0\right)\mathbf{\Theta}} = \frac{\alpha}{\alpha - 1}\left(\frac{\alpha}{\beta}\Lambda_0\right)^{-1} = \begin{pmatrix} \Delta_0^2 & & & 0 \\ & \Delta_1^2 & & \\ & & \ldots & \\ 0 & & & \Delta_k^2 \end{pmatrix}, \qquad (9)$$

где Δ_i^2 — заданные значения априорных дисперсий компонент вектора

$$\mathbf{\Theta} = (\theta_0, \theta_1, \theta_2, \ldots, \theta_k); \quad j = 0, 1, 2, \ldots, k.$$

Таким образом, векторный параметр сдвига определяется заданным вектором априорных средних значений $\boldsymbol{\Theta_0}$, а диагональные элементы $\lambda_0^{(j)}$ (j = = 1, 2, …, k) матрицы $\boldsymbol{\Lambda}_0$ определяются по формулам:

$$\lambda_0^{(j)} = \frac{1}{\Delta_j^2} \cdot \frac{\beta}{\alpha - 1}; \qquad \alpha = \frac{h_0^2}{\Delta_h^2}; \qquad \beta = \frac{h_0}{\Delta_h^2};\ h_0 = Eh; \qquad \Delta_h^2 = Dh. \quad (10)$$

Пересчёт значений параметров при переходе от априорного сопряжённого распределения к апостериорному

Байесовское оценивание коэффициентов регрессии $\boldsymbol{\Theta} = (\theta_0, \theta_1, \theta_2, \ldots, \theta_k)$ и параметра h в ММЛР (4) предполагает использование апостериорного распределения $\tilde{\rho}(\boldsymbol{\Theta}; h | \mathbf{X}, \mathbf{Y})$ этих параметров, определяемого по схеме (2). Подставляя в правую часть соотношения (2) в качестве априорного многомерное гамма-распределение (8), а также функцию правдоподобия $L(\mathbf{X}, \mathbf{Y} | \boldsymbol{\Theta}; h)$ (7), преобразованную к виду:

$$L(\mathbf{X}, \mathbf{Y} | \boldsymbol{\Theta}; h) \approx h^{\frac{n-k-1}{2}} e^{-\left(\frac{n-k-1}{2}\tilde{\sigma}^2\right)h} h^{\frac{n+1}{2}} \exp\left(-\frac{h}{2}(\hat{\Theta} - \Theta)^T (X^T X)(\hat{\Theta} - \Theta)\right),$$

получим после ряда тождественных преобразований [18–19] апостериорную плотность $\rho(\boldsymbol{\Theta}; h | \mathbf{X}, \mathbf{Y})$ в форме многомерного гамма-нормального распределения (8), параметры которого определяются по параметрам $\boldsymbol{\Theta}_0$, $\boldsymbol{\Lambda}_0$, α, β априорного распределения и наблюдениям ($\mathbf{X}$, $\mathbf{Y}$) следующими соотношениями:

$$\tilde{\boldsymbol{\Theta}}_0 = (\boldsymbol{\Lambda}_0 + \mathbf{X}^T\mathbf{X})^{-1}(\boldsymbol{\Lambda}_0 + \mathbf{X}^T\mathbf{X}); \qquad \tilde{\boldsymbol{\Lambda}}_{\mathbf{0}} = \boldsymbol{\Lambda}_0 + \mathbf{X}^T\mathbf{X}; \qquad \tilde{\alpha} = \alpha + \frac{n}{2};$$

$$\tilde{\beta} = \beta + \frac{1}{2}[(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\Theta}}_0)^T \mathbf{Y} + (\boldsymbol{\Theta}_0 - \tilde{\boldsymbol{\Theta}}_0)^T \boldsymbol{\Lambda}_0 \boldsymbol{\Theta}_0]. \quad (11)$$

Таким образом, полученные формулы (11) позволяют осуществить пересчёт параметров при переходе от априорного распределения к апостериорному.

Метод построения и исследования прогнозной модели

В ММЛР (4) введём в рассмотрение, наряду с наблюдёнными значениями $\mathbf{X}$ и $\mathbf{Y}$ анализируемых переменных $\mathbf{Y}$, их прогнозные значения на q шагов вперёд:

$$\tilde{\mathbf{X}} = \begin{pmatrix} 1 & X_{n+1}^{(1)} & X_{n+1}^{(2)} & \ldots & X_{n+1}^{(k)} \\ 1 & X_{n+2}^{(1)} & X_{n+2}^{(2)} & \ldots & X_{n+2}^{(k)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n+q}^{(1)} & X_{n+q}^{(2)} & \ldots & X_{n+q}^{(k)} \end{pmatrix}; \quad \tilde{\mathbf{Y}} = \begin{pmatrix} Y_{n+1} \\ Y_{n+2} \\ \ldots \\ Y_{n+q} \end{pmatrix}, \quad (12)$$

а также соответствующие остатки $\tilde{\boldsymbol{\varepsilon}} = (\varepsilon_{n+1}, \varepsilon_{n+2}, \ldots, \varepsilon_{n+q})^T$.

Тогда в соответствии (4):

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\Theta} + \tilde{\boldsymbol{\varepsilon}}; \qquad \tilde{\boldsymbol{\varepsilon}} \in N_q(\mathbf{O}; \frac{1}{h} \cdot \mathbf{I}_q).$$

Для того чтобы строить точечные и интервальные оценки для $\tilde{\mathbf{Y}}$ по заданным значениям $\mathbf{X}$, $\tilde{\mathbf{X}}$, $\mathbf{Y}$, очевидно, надо располагать плотностью условного распределения $\rho(\mathbf{X}\big|\tilde{\mathbf{X}}, \mathbf{Y})$, которую обычно называют «прогнозной функцией плотности вероятности».

Но поскольку из прогнозной модели следует, что распределение вектора $\tilde{\mathbf{Y}}$ зависит также от параметров $\boldsymbol{\Theta}$ и h, а они в БП интерпретируются как случайные величины, имеющие соответствующее апостериорное распределение, то прогнозная функция плотности $\rho(\tilde{\mathbf{Y}}\big|\mathbf{X}; \tilde{\mathbf{X}}, \mathbf{Y})$ может быть определена по формуле [18]:

$$\rho(\tilde{\mathbf{Y}}\big|\mathbf{X}; \tilde{\mathbf{X}}; \mathbf{Y}) \approx \left[1 + \frac{1}{\nu}(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\Theta}}_0)^T \mathbf{B}'(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\Theta}}_0)\right]^{-\frac{\nu+q}{2}}, \quad \nu = n-k-1, \tag{13}$$

где

$$\mathbf{B}' = \frac{\tilde{\alpha}}{\tilde{\beta}}\left[\mathbf{I}_q - \tilde{\mathbf{X}}(\Lambda_0 + \mathbf{X}^{\mathbf{T}}\mathbf{X} + \tilde{\mathbf{X}}^{\mathbf{T}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathbf{T}}\right]. \tag{14}$$

Таким образом, условное распределение q-мерного вектора $\tilde{\mathbf{Y}}$ при заданных значениях $\mathbf{X}$, $\mathbf{Y}$, $\tilde{\mathbf{X}}$ описывается обобщённым многомерным t-распределением с $\nu = n - k - 1$ степенями свободы, параметром сдвига $\tilde{\mathbf{X}}\tilde{\boldsymbol{\Theta}}_0$ и матрицей точности $\mathbf{B}'$, определённой соотношением (14), то есть $\rho(\tilde{\mathbf{Y}}\big|\mathbf{X}; \tilde{\mathbf{X}}; \mathbf{Y}) = t(n-k-1\big|\tilde{\mathbf{X}}\tilde{\boldsymbol{\Theta}}_0; \mathbf{B}')$.

Используя свойства обобщённого t-распределения Стьюдента [14, 15], получаем следующие байесовские прогнозы для $\tilde{\mathbf{Y}}$:

а) точечный байесовский прогноз для компонент вектора $\tilde{\mathbf{Y}}$ определяется соотношением $\hat{y}_{n+m} = (\hat{\boldsymbol{\Theta}}^{(\text{Б})})^T x_{n+m}$, $m = 1, 2, ..., q$;

б) интервальный байесовский прогноз для компонент вектора $\tilde{\mathbf{Y}}$ с вероятностью P_0 определяется соотношением

$$y_{n+m} \in \left[\hat{y}_{n+m} \pm t_{\frac{1-P_0}{2}}(n-k-1) \cdot \frac{1}{\sqrt{C'_m}}\right]; \qquad m = 1, 2, ..., q, \tag{15}$$

где $t_\varepsilon(\nu)$ — 100ε%-я точка стандартного $t(\nu)$ распределения Стьюдента, а величины C'_m вычисляются путём замены $(k \times k)$-матрицы $\tilde{\mathbf{B}}$ на $(q \times q)$-матрицу $\mathbf{B}'$, определённую соотношением (14);

в) байесовская прогнозная доверительная область $\Delta\tilde{Y}$ для вектора $\tilde{\mathbf{Y}} = (y_{n+1},...,y_{n+q})^T$ состоит, с заданной вероятностью P_0, из всех тех $\tilde{\mathbf{Y}} = (y_{n+1},...,y_{n+q})^T$, которые удовлетворяют неравенству

$$\frac{1}{q}(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}\hat{\boldsymbol{\Theta}}^{(Б)})^T\Sigma_{\tilde{Y}}^{-1}(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}\hat{\boldsymbol{\Theta}}^{(Б)})<F_{1-P_0}(q;n-k-1), \tag{16}$$

где $F_\varepsilon(\nu_1,\nu_2)$ — 100ε%-я точка распределения Фишера $F(\nu_1,\nu_2)$; $\hat{\boldsymbol{\Theta}}^{(Б)}$ — байесовская точечная оценка параметров регрессии $\boldsymbol{\Theta}$, $\boldsymbol{\Sigma}_{\tilde{\mathbf{Y}}}=\frac{n-k-1}{n-k-3}\left(\mathbf{B}'\right)^{-1}$ — ковариационная матрица вектора $\tilde{\mathbf{Y}}$.

Отметим также, что при применении МНК для одномерной регрессионной модели интервальная оценка для прогнозного значения y в точке $x=x_{n+m}$ имеет вид, аналогичный (15) [14–15]:

$$y_{n+m}\in\left(b_0+b_1x_{n+m}\pm t_\gamma\hat{s}\sqrt{\frac{1}{n}+\frac{(x_{n+m}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+1}\right). \tag{17}$$

Общий вид доверительного интервала, найденного с надёжностью γ, изображён на рис. 1.

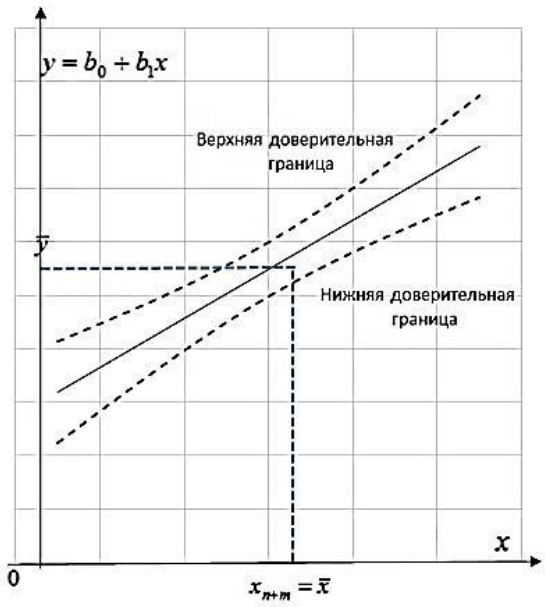

Рис. 1. Доверительная область для одномерной модели

Доверительный интервал имеет наименьшую величину при $x_{n+m} = \overline{x}$, то есть когда прогнозное значение признака равно $\overline{x}$ (при этом среднее слагаемое под знаком квадратного корня обращается в ноль). По мере удаления x_{n+m} от среднего значения $\overline{x}$ ширина доверительного интервала увеличивается, а точность оценки ухудшается. Многомерный случай не имеет такой наглядной интерпретации.

Результаты моделирования

Результаты построения и исследования уравнения регрессии

Пусть в результате двадцати измерений некоторого параметра были получены значения: **X** = (17, 22, 27, 33, 33, 39, 39, 41, 42, 45, 46, 47, 48, 51, 51, 52, 54, 60, 60, 100); **Y** = (381, 372, 419, 449, 477, 433, 478, 440, 479, 511, 517, 494, 511, 498, 520, 487, 548, 483, 534, 629).

Построим и исследуем регрессионную модель:

$$y = \theta_0 + \theta_1 x + \varepsilon. \qquad (18)$$

Пусть при анализе предыстории функционирования объекта получена следующая априорная информация: $h_0 = Eh = 0{,}002$; $\Delta_0^2 = D\theta_0 = 2{,}25$; $\Delta_1^2 = D\theta_1 = 0{,}01$; $\Delta_h^2 = Dh = 25 \cdot 10^{-8}$.

Для данной модели существует сопряжённое с наблюдаемой генеральной совокупностью распределение параметров θ_0, θ_1, h, которое описывается трёхмерным гамма-нормальным распределением с параметрами $\boldsymbol{\Theta}^0 = (\theta_0^0, \theta_1^0)^T$, $\boldsymbol{\Lambda}_0$, α, β , определяемыми в соответствии с (10):

$$\boldsymbol{\Theta}^0 = (330; 2{,}85)^T; \quad \alpha = \frac{h_0^2}{\Delta_h^2} = 16; \quad \beta = \frac{h_0}{\Delta_h^2} = 8000; \quad \boldsymbol{\Lambda}_0 = \begin{pmatrix} 2{,}37 & , & 0 \\ 0 & , & 53333{,}3 \end{pmatrix}.$$

Параметры апостериорного гамма-нормального распределения вычисляются в соответствии с формулами пересчёта (11):

$$\tilde{\boldsymbol{\Theta}}_0 = \begin{pmatrix} 349{,}0 \\ 2{,}9 \end{pmatrix}; \qquad \tilde{\alpha} = 26; \qquad \tilde{\beta} = 14578; \qquad \tilde{\boldsymbol{\Lambda}}_0 = \begin{pmatrix} 22{,}37 & , & 907 \\ 907 & , & 100176 \end{pmatrix}.$$

Точечные байесовские оценки параметров θ_0, θ_1, h определяются средними значениями соответствующих частных апостериорных распределений. С учётом свойств многомерного гамма-нормального распределения [14, 15] имеем:

$$\hat{\boldsymbol{\Theta}}^{(\text{Б})} = E(\boldsymbol{\Theta} | \mathbf{X}, \mathbf{Y}) = \tilde{\boldsymbol{\Theta}}_0 = (349{,}0; 2{,}90)^T; \quad \hat{h}^{(\text{Б})} = E(h | \mathbf{X}, \mathbf{Y}) = \frac{\tilde{\alpha}}{\tilde{\beta}} = 0{,}00178.$$

При выводе интервальных байесовских оценок используется величина $(\hat{\theta}_j^{(Б)} - \theta_j)\sqrt{\tilde{C}_j}$, которая имеет распределение Стьюдента, $\tilde{C}_j$ — параметр точности. В нашем случае: $\tilde{\mathbf{B}} = \frac{\tilde{\alpha}}{\tilde{\beta}}\tilde{\mathbf{\Lambda}}_0 = \begin{pmatrix} 0{,}040 & 1{,}618 \\ 1{,}618 & 178{,}665 \end{pmatrix}$; $\tilde{C}_0 = 0{,}0254$; $\tilde{C}_1 = 113{,}217$.

Следовательно, с вероятностью P_0 = 0,90 должны выполняться неравенства: $\left|(\hat{\theta}_0^{(Б)} - \theta_0)\right|\sqrt{0{,}0254} < t_{0{,}05}(52)$; $\left|(\hat{\theta}_1^{(Б)} - \theta_1)\right|\sqrt{113{,}217} < t_{0{,}05}(52)$. Поэтому $\theta_0 \in [338{,}5; 359{,}5]$ и $\theta_1 \in [2{,}743; 3{,}057]$.

Поскольку параметр h подчиняется апостериорному гамма-распределению с параметрами $\tilde{\alpha}$ и $\tilde{\beta}$, то $h \in [\gamma_{0{,}95}(\tilde{\alpha}; \tilde{\beta}); \gamma_{0{,}05}(\tilde{\alpha}; \tilde{\beta})]$. С учётом того, что $\gamma_q(\tilde{\alpha}; \tilde{\beta}) = \frac{1}{2\tilde{\beta}}\chi_q^2(2\tilde{\alpha})$, имеем $h \in [0{,}00125;\ 0{,}00239]$.

Результаты сравнения оценивания с помощью БП и ММП представлены в верхней части таблицы.

Таблица

Сравнение байесовского подхода и метода максимального правдоподобия

Модель множественной линейной регрессии				
	ММП		**БП**	
Параметр	Среднее значение	Доверительный интервал	Среднее значение	Доверительный интервал
θ_0	344,7	[316,11; 373,3]	349,0	[338,5; 359,5]
θ_1	3,05	[2,45; 3,64]	2,90	[2,743; 3,057]
h	0,005	[0,0009; 0,00285]	0,00178	[0,00125; 0,00239]
Прогнозная модель				
	ММП		**БП**	
y_{21}	710,7	[647,1; 774,3]	697,4	[653,6; 741,2]
y_{22}	771,7	[699,2; 844,2]	755,5	[710,7; 800,3]

Видно, что БП позволяет сузить доверительный интервал для θ_0 в 2,6 раза, для θ_1 — в 3,7 раза и для h — в 1,4 раза по сравнению с подходом, основанным на ММП.

Отметим, что доверительная область для оценки $\tilde{y}$ в случае одномерной регрессионной модели с надёжностью γ для заданного значения x описывается соотношением [14–15]:

$$\tilde{y}_x \in \left(b_0 + b_1 x \pm t_\gamma \hat{s} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right),$$

где t_γ определяется по таблице распределения Стьюдента для уровня значимости γ и числа степеней свободы $\alpha = 1 - \gamma$ и $\nu = n - 2$. Эта область изображена на рис. 2.

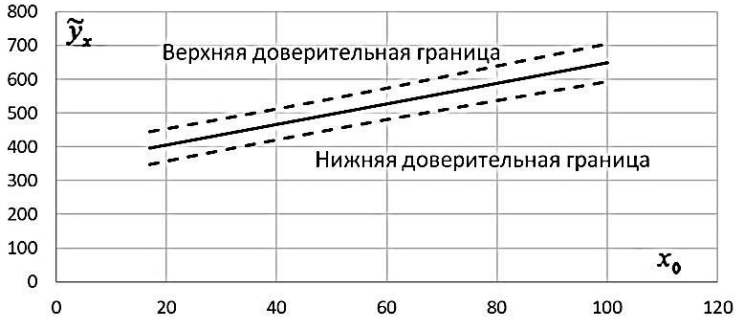

Рис. 2. График уравнения линейной регрессии и доверительная область

Отметим, что границы области — криволинейные (качественно границы имеют вид, аналогичный изображённому на рис. 1), однако в масштабе рисунка указанная криволинейность слабо просматривается. Это связано со спецификой статистической информации рассматриваемого примера. Отметим, что у прогнозного значения y_{n+m}, задаваемого соотношением (17), дисперсия больше, чем у оценки $\tilde{y}_x$, поскольку подкоренное выражение на единицу больше, чем в приведённой выше формуле. Поэтому соответствующая доверительная область оказывается шире.

Результаты прогнозирования с использованием уравнения регрессии

Пусть $q = 2$, $x_{21} = 120$, $x_{22} = 140$. Тогда $\tilde{\mathbf{X}} = \begin{pmatrix} 1 & 120 \\ 1 & 140 \end{pmatrix}$. Плотность условного распределения вектора $\tilde{Y}$ при заданных **X,** $\tilde{\mathbf{X}}$ и **Y** описывается обобщённым многомерным *t*-распределением с числом степеней свободы $\nu = 20 - 1 - 1 = 18$, параметром сдвига $\begin{pmatrix} 1 & 120 \\ 1 & 140 \end{pmatrix}\begin{pmatrix} 349{,}0 \\ 2{,}9 \end{pmatrix}$ и матрицей точности $\mathbf{B}'$, определённой соотношением (14).

Произведя вычисления по (15)–(16) и используя свойства обобщённого многомерного *t*-распределения для $P_0 = 0{,}90$, имеем:

$$\hat{y}_{21} = 349 + 2{,}9 \cdot 120 = 697{,}4; \quad \hat{y}_{22} = 349 + 2{,}9 \cdot 140 = 755{,}5; \quad \mathbf{B}' = \begin{pmatrix} 0{,}00159 & -0{,}00022 \\ -0{,}00022 & 0{,}00152 \end{pmatrix};$$

$$\mathbf{\Sigma}_{\tilde{\mathbf{Y}}} = \begin{pmatrix} 721{,}8 & 106{,}8 \\ 106{,}8 & 757{,}4 \end{pmatrix}; \qquad y_{21} \in [653{,}6; 741{,}2]; \qquad y_{22} \in [710{,}7; 800{,}3];$$

$$\Delta_{\tilde{Y}} = \left\{ \begin{pmatrix} y_{21} \\ y_{22} \end{pmatrix} : \frac{1}{2} \cdot \begin{pmatrix} y_{21} - 697{,}4 \\ y_{22} - 755{,}5 \end{pmatrix}^T \begin{pmatrix} 0{,}00159 & -0{,}00022 \\ -0{,}00022 & 0{,}00152 \end{pmatrix} \begin{pmatrix} y_{21} - 697{,}4 \\ y_{22} - 755{,}5 \end{pmatrix} < 2{,}62 \right\}. \quad (19)$$

Результаты сравнения оценивания с помощью БП и ММП для прогнозной модели представлены в нижней части таблицы.

Из таблицы видно, что в задаче прогнозирования БП позволяет сузить ширину прогнозной интервальной оценки для y_{21} в 1,45 раза, а для y_{22} — в 1,62 раза.

Доверительная область, построенная на основе (19) при совместном прогнозировании y_{21} и y_{22}, представляет собой эллипс. Для трёх разных значений доверительной вероятности P_0 доверительные области изображены на рис. 3.

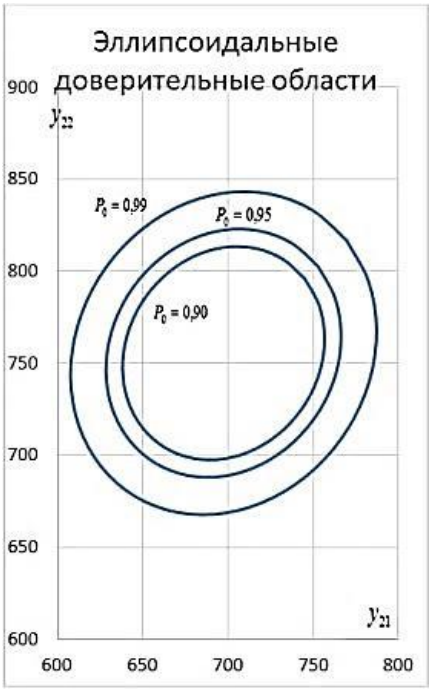

Рис. 3. Доверительные области при совместном прогнозировании двух значений y_{21} и y_{22}

Отметим, что центральные оси эллипса развёрнуты относительно исходной системы координат примерно на угол 41°. Это означает, что величины y_{21} и y_{22} имеют примерно одинаковые степени роста при соответствующем изменении x_{21} и x_{22} (одинаковая степень роста реализуется при угле 45°).

Заключение

В статье изложен метод построения статистических оценок параметров ММЛР. Представленные результаты в совокупности с результатами [11–13] охватывают достаточно широкое многообразие законов распределения и статистических моделей, встречающихся в практике обработки результатов измерений и исследования измерительных задач.

При построении статистических оценок на основе БП существенную роль играют распределения, сопряжённые с наблюдаемой генеральной совокупностью. В статье сформулированы необходимые условия существования сопряжённых распределений для коэффициентов ММЛР, представлен алгоритм определения общего вида сопряжённого распределения. Показано, что коэффициенты ММЛР являются случайными величинами, имеющими гамма-нормальное распределение. Описан алгоритм расчёта неизвестных параметров гамма-нормального закона распределения.

БП особенно эффективен в задачах оценки метрологических характеристик измерительных комплексов и средств измерений в случае, когда многократное повторение испытаний представляется нецелесообразным или невозможным. Если же имеется практическая возможность увеличения объёма и количества выборок, то БП и классические методы будут давать всё более близкие результаты.

Представленные в статье формулы пересчёта параметров ММЛР, а также формулы пересчёта для различных законов распределения [11–13] не требуют для реализации больших вычислительных ресурсов и могут быть реализованы с помощью программного обеспечения ВАЦ или при работе ВАЦ в комплексе с внешним компьютером. Полученные в статье результаты могут быть использованы для построения самообучающихся и самонастраивающихся систем, поскольку позволяют учитывать появляющуюся в процессе измерений дополнительную информацию и использовать её для построения более достоверных статистических оценок.

Список литературы

1. Джоэль П. Дансмор Измерения параметров СВЧ-устройств с использованием передовых методик векторного анализа цепей. — М.: Техносфера, 2019. — 736 с.
2. Duyguİçen D. A new approach for probability calculation of fuzzy events in Bayesian Networks // International Journal of Approximate Reasoning. — 2019. — V. 108. — P. 76–88.
3. Yang H., Jintao Ke J., Jieping Ye J. A universal distribution law of network detour ratios Transportation Research Part C // Emerging Technologies. — 2018. — V. 96. — P. 22–37.
4. Higgins V., Asgari S., Adeli K. Choosing the best statistical method for reference interval estimation // Clinical Biochemistry. — 2019. — V. 71. — P. 14–16.
5. Touzani S., Ravache B., Crowe E., Granderson J. Statistical change detection of building energy consumption. Applications to savings estimation // Energy and Building Journal. — 2019. — V. 18 (515). — P. 123–136.
6. Lavrik E., Frankenfeld U., Mehta S., Panasenko I., Schmidt H. High-precision contactless optical 3D-metrology of silicon sensors Nuclear Instruments and Methods in Physics Research. Section A // Accelerators, Spectrometers, Detectors and Associated Equipment. — 2019. — V. 935. — P. 167–172.

7. Francisco S., Guzmán J., Rosa B., Rodríguez C., Doimeadios M., Ángel R. Analytical metrology for nanomaterials. Present achievements and future challenges // Analytica Chimica Acta. — 2019. — V. 1059. — P. 1–15.
8. Gao W., Haitjema H., Fang F., Leach R., Cheung C., Savio E., Linares J. On-machine and in-process surface metrology for precision manufacturing // CIRP Annals. — 2019. — V. 68. — I. 2.
9. Кузнецов В.А., Исаев Л.К., Шайко И.А. Метрология. — М.: Стандартинформ, 2005. — 298 с.
10. ГОСТ РВ 0015-002-2012. Система разработки и постановки на производство военной техники. Система менеджмента качества. Общие требования. — М.: Стандартинформ, 2012. — 67 с.
11. Волчков А.А., Исаев Ю.А., Леонова К.С., Фуфаева О.А., Хайруллин Р.З. Метод построения оценок точности измерений на основе использования апостериорной информации // Вестник метролога. — 2019. — № 4. — С. 18–21.
12. Хайруллин Р.З. Применение байесовского подхода в задачах построения статистических оценок при обработке результатов испытаний измерительной техники // Вестник метролога. — 2020. — № 1. — С. 9–15.
13. Хайруллин Р.З., Закутин А.А. Применение байесовского подхода к построению статистических оценок параметров законов распределения случайных величин // Измерительная техника. — 2020. — № 11. — С. 14–21.
14. Айвазян С.А. Байесовский подход в эконометрическом анализе // Прикладная эконометрика. — 2008. — № 1 (9). — С. 93–130.
15. Айвазян С.А., Мхитарян В.С. Прикладная статистика в задачах и упражнениях. — М.: Юнити-Дана, 2001. — 270 с.
16. Вентцель Е.С. Исследование операций. — М.: Наука, 1972. — 552 с.
17. Вентцель Е.С., Овчаров Л.А. Теория случайных процессов и её инженерные приложения. — М.: Наука, 2014. — 383 с.
18. Зельнер А. Байесовские методы в эконометрии / пер. с англ. Г.Г. Пирогова и Ю.П. Федоровского. — М.: Статистика, 1980. — 438 с.
19. Де Гроот М. Оптимальные статистические решения: пер. с англ. — М.: Мир, 1974. — 492 с.

Статья поступила в редакцию: 03.11.2021 г.
Статья прошла рецензирование: 17.11.2021 г.
Статья принята в работу: 19.11.2021 г.