# МАТЕМАТИКА

УДК 519.22-24
MSC 62-07, 62B10, 62H86

## Об аппроксимации прогноза частичными предсказаниями в условиях неполных данных*

*Н. П. Алексеева*[1], *Ф. С. Ш. Ал-Джубури*[2]

[1] Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9
[2] Университет информационных технологий и коммуникаций,
Ирак, Багдад, ул. Ал-Нидал

**Для цитирования:** *Алексеева Н. П., Ал-Джубури Ф. С. Ш.* Об аппроксимации прогноза частичными предсказаниями в условиях неполных данных // Вестник Санкт-Петербургского университета. Математика. Механика. Астрономия. 2022. Т. 9 (67). Вып. 4. С. 575–589. https://doi.org/10.21638/spbu01.2022.401

В статье рассматриваются применение метода случайных подпространств для прогнозирования по неполным данным и построение оценки полного прогноза по набору частичных предсказаний. Не умаляя общности, изучаются центрированные частичные предсказания. Согласно статистической модели, внедиагональные элементы корреляционной матрицы частичных предсказаний рассматриваются случайными с заданными математическим ожиданием и дисперсией. Получены аналитические выражения математического ожидания определителя и алгебраических дополнений этой матрицы. В результате построен класс более точных оценок полного прогноза, которые отличаются от среднего частичного предсказания множителями, зависящими от статистических параметров корреляционной матрицы частичных предсказаний. Приведены результаты моделирования и практического прогнозирования на неполных биогеографических данных.

*Ключевые слова*: метод случайных подпространств, статистическая модель, матрица со случайными элементами, частичные предсказания, множественная регрессия.

**1. Введение.** В большинстве методов многомерного статистического анализа предполагается, что данные полные. Если имеются пропуски, то обычно или удаляется соответствующее наблюдение, или исключается сама переменная, или предла-

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 20-01-00096).

гается недостающее значение заменить чем-то случайным. Известны методы, позволяющие получить искусственным образом аналог полных данных с заданными статистическими характеристиками [1, 2]. Однако если посмотреть на проблему с другой стороны, то можно обойтись без какого бы то ни было заполнения пропусков. Все зависит от того, к какому искажению модели приводит неполнота данных — если к смещению модели, например в дисперсионном анализе для зависимых выборок, то нужна централизация модели с дальнейшим пересчетом ковариационной матрицы ошибок [3]. Для линейных многомерных статистических задач можно воспользоваться идеей частичности, долгое время пропагандируемой в работах А. Г. Барта [4], а в настоящее время известной по методу случайных подпространств [5]. Суть такого подхода заключается в рассмотрении комплекса частичных предсказаний, построенных по подмножествам наблюдений или по части предикторов.

В данной работе рассматривается классическая задача множественной линейной регрессии [6], при решении которой в случае неполноты данных возникает проблема построения наилучшего линейного предсказания переменной $Y$ по независимым переменным $X_1, \ldots, X_n$, представленным в виде произведений $\eta_1 X_1, \ldots, \eta_n X_n$, где $\eta = (\eta_1, \ldots, \eta_n)$ — вектор дихотомических случайных величин. Нужно найти линейную комбинацию независимых переменных

$$\tilde{Y} = \sum_{j=1}^{n} \beta_j(\eta) X_j \tag{1}$$

с коэффициентами $\beta_j(\eta)$, зависящими от значений реализации дихотомического вектора $\eta$. Другими словами, коэффициенты регрессии должны меняться в соответствии со структурой имеющихся зависимых переменных. Вместо того чтобы напрямую искать численное решение этой в общем-то непростой задачи, заметим, что аналогичную структуру имеет прогноз, представленный в виде среднего арифметического частичных предсказаний, который используется при построении метаоценок в регрессионной задаче по методу Random Forest [7].

Среднее частичное предсказание не является оптимальной оценкой полного прогноза, а может использоваться только в качестве начального приближения для коэффициентов из (1). В этом можно убедиться, если рассмотреть классическую задачу построения наилучшего линейного предсказания по частичным регрессиям, выбранным в качестве предикторов. Для того чтобы выяснить, насколько отличается это решение от среднего частичного предсказания, рассмотрим статистическую модель, в которой внедиагональные элементы корреляционной матрицы частичных предсказаний имеют вид случайных одинаково распределенных величин. Если вычислить математические ожидания определителя и алгебраических дополнений этой матрицы, то можно будет выразить математическое ожидание предсказания в виде среднего арифметического частичных регрессий с точностью до некоторого корректирующего множителя, равного единице только в вырожденном случае. После этого могут быть получены коэффициенты $\beta_j(\eta)$.

Частичные предсказания используются для любого вида прогнозирования, не только в регрессионной задаче. Но, например, в задаче классификации можно как раз обойтись обычным усреднением частичных классификаторов, поскольку важно не столько само значение классифицирующей функции, сколько его положение относительно граничного значения, хотя для оценки апостериорных вероятностей все остается актуальным. Этот метод был применен для изучения возможности прогно-

зирования числа госпитализаций кардиологических больных, наблюдаемых в течение длительного времени [8], и для изучения механизма формирования маркеров разрушения внеклеточного матрикса при туберкулезе легких [9].

**2. Статистическая модель структуры зависимости частичных предсказаний.** Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — вектор случайных величин с ковариационной матрицей $\mathbf{\Sigma} = \{\sigma_{ij}\}_{i,j=1}^n$, по которым строится прогноз для переменной $Y$ с дисперсией $\sigma^2$. Прогнозируемую переменную $Y$ принято называть зависимой, а компоненты вектора $\mathbf{X}$ — независимыми переменными. Не умаляя общности, будем считать все величины центрированными, т. е. $\mathbb{E}Y = 0$ и $\mathbb{E}X_i = 0$, $i = 1, \ldots, n$. Обозначим через $\mathbf{\Lambda} = \{\lambda_{ij}\}_{i,j=1}^n$ корреляционную матрицу независимых переменных, через $\mathbf{l} = (l_1, \ldots, l_n)$ — вектор смешанных вторых моментов с зависимой переменной, $l_j = \mathbb{E}\,YX_j = \varrho_j\sigma\sqrt{\sigma_{jj}}$, $j = 1, \ldots, n$, с соответствующим вектором коэффициентов корреляции $\varrho = (\varrho_1, \ldots, \varrho_n)$. Наилучшее линейное предсказание переменной $Y$ в смысле минимума среднеквадратичного отклонения $\mathbb{E}(Y - \sum\limits_{j=1}^{n} \beta_j X_j)^2$ выражается [10] в виде линейной комбинации

$$f_0 = F(\mathbf{X}) = \sum_{j=1}^{n} \beta_j X_j, \quad \text{где } \beta_j = \frac{1}{\Sigma}\sum_{i=1}^{n} l_i \Sigma_{ij}, \tag{2}$$

где $\Sigma$ и $\Sigma_{ij}$ — обозначены соответственно определитель и алгебраические дополнения по $i$-строке и $j$-столбцу матрицы $\mathbf{\Sigma}$. Для дальнейшего исследования необходимы некоторые факты из теории множественной регрессии. Приведем их применительно к выбранным обозначениям с доказательствами, вынесенными в приложение.

**Предложение 1.** *Пусть* $\mathbf{Z}$ *и* $\varrho$ *— векторы с компонентами* $Z_i = \frac{X_i}{\sqrt{\sigma_{ii}}}$, $\varrho_i = \frac{l_i}{\sigma\sqrt{\sigma_{ii}}}$ *соответственно,* $i = 1, 2, \ldots, n$, *и* $\mathbf{\Sigma} = \mathbb{E}\mathbf{X}\mathbf{X}^T$, $\mathbf{l} = \mathbb{E}\,\mathbf{X}Y$. *Тогда: 1) наилучшее линейное предсказание (2) для* $Y$ *по* $\mathbf{X}$ *имеет вид* $F(\mathbf{X}) = \mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{l} = \sigma\mathbf{Z}^T\mathbf{\Lambda}^{-1}\varrho$; *2) дисперсия предсказания* $F(\mathbf{X})$ *равна* $\mathbb{D}F(\mathbf{X}) = \mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbf{l}$; *3) коэффициент детерминации имеет вид* $R^2 = \sigma^{-2}\mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbf{l} = \varrho^T\mathbf{\Lambda}^{-1}\varrho$; *4) если* $\mathbf{A}$ *— квадратная матрица полного ранга порядка* $n$, *то* $F(\mathbf{A}X) = F(\mathbf{X})$.

По любому вектору $\mathbf{X}_\tau = (X_{t_1}, \ldots, X_{t_m})$, составленному из подмножества независимых переменных, где $\tau = (t_1, \ldots, t_m) \subseteq (1, 2, \ldots, n)$, можно построить частичное предсказание $F(\mathbf{X}_\tau) = \mathbf{X}_\tau^T\mathbf{\Sigma}_m^{-1}\mathbf{l}_\tau$, где $\mathbf{\Sigma}_m = \mathbb{E}\mathbf{X}_\tau\mathbf{X}_\tau^T$ — частичная ковариационная матрица, а $\mathbf{l}_\tau = (l_{t_1}, \ldots, l_{t_m})$. Если какие-то $n$ линейно независимых частичных предсказаний $\mathbf{f} = (f_1, \ldots, f_n)$, $f_i = F(\mathbf{X}_{\tau_i})$, рассматривать в качестве предикторов вместо переменных $X_1, \ldots, X_n$, то, согласно последнему положению в предложении 1, мы получим то же самое предсказание: $F(\mathbf{f}) = F(\mathbf{X})$. Поскольку частичные предсказания очевидно положительно коррелированы между собой, а коэффициенты корреляции частичных предсказаний с зависимой переменной $\mathrm{cor}(Y, f_i)$ менее вариабельны, чем компоненты вектора $\varrho = \mathrm{cor}(Y, \mathbf{X})$, рассмотрим статистическую модель, в которой корреляционная матрица $\mathbf{\Lambda}_n$ частичных предсказаний $f_1, \ldots, f_n$ имеет вид случайной матрицы:

$$\mathbf{\Lambda}_n = \begin{bmatrix} 1 & r + x_{12} & \ldots & r + x_{1n} \\ r + x_{21} & 1 & \ldots & r + x_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ r + x_{n1} & r + x_{n2} & \ldots & 1 \end{bmatrix}, \tag{3}$$

где $\mathbb{E}x_{jk} = 0$, $\mathbb{D}x_{jk} = \sigma_0^2$, $x_{jk} = x_{kj}$, $0 < r < 1$, и вектор $\mathrm{cor}(Y, \mathbf{f})$ имеет вид $\varrho = \varrho_0 \mathbf{1}^n + \varepsilon$, где $0 < \varrho_0 < 1$, $\mathbf{1}^n = (1, 1, \ldots, 1)$ — вектор из $n$ единиц, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$, $\mathbb{E}\varepsilon_i = 0$, $\mathbb{D}\varepsilon_i = \sigma_1^2$. Случайные величины $\varepsilon_i$ и $x_{jk}$ будем считать независимыми в вероятностном смысле.

В соответствии с моделью (3) математическое ожидание оценки (2) можно рассматривать в качестве оценки полного прогноза $f_0$, зависящей от параметров $r$, $\varrho$, $\sigma_0$. Первые две характеристики отвечают за точность прогнозирования и могут использоваться в качестве управляющих параметров при выборе числа предикторов в частичных предсказаниях. При небольшой дисперсии $\sigma_0^2$ возможно выражение полного прогноза через функцию от среднего частичного предсказания. Дисперсия $\sigma_1^2$, указывающая на паритетные свойства частичных предсказаний, напрямую не входит в оценку $f_0$ и отвечает за ее эффективность.

Вычисление математического ожидания оценок (2) в указанных предположениях оказалось непростой задачей, поэтому рассмотрим по отдельности математические ожидания определителей и алгебраических дополнений корреляционной матрицы $\mathbf{\Lambda}_n$ и используем для оценивания их отношение. Свойства полученных оценок можно будет изучить при помощи моделирования.

**3. Математическое ожидание специального определителя со случайными компонентами.** Обозначим через $\mathbf{U}_n$ частный случай матрицы $\mathbf{\Lambda}_n$ при $\sigma_0 = 0$. Насколько значимо сказывается отклонение $\mathbf{\Lambda}_n$ от $\mathbf{U}_n$ или разнообразие коэффициентов $\varrho_j$ на точности полного предсказания по частичным, можно далее узнать, если построить оценку полного прогноза с учетом статистических свойств $\mathbf{\Lambda}_n$ и $\varrho_j$. Представим матрицу $\mathbf{\Lambda_n}$ в виде суммы двух матриц $\mathbf{\Lambda_n} = \mathbf{R}_n + \mathbf{V}_n$, где $a = 1 - r$,

$$\underbrace{\begin{bmatrix} 1 & r + x_{12} & \ldots & r + x_{1n} \\ r + x_{12} & r & \ldots & r + x_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ r + x_{1n} & r + x_{2n} & \ldots & 1 \end{bmatrix}}_{\mathbf{\Lambda}_n} = \underbrace{\begin{bmatrix} r & r & \ldots & r \\ r & r & \ldots & r \\ \vdots & \vdots & \ldots & \vdots \\ r & r & \ldots & r \end{bmatrix}}_{\mathbf{R}_n} + \underbrace{\begin{bmatrix} a & x_{12} & \ldots & x_{1n} \\ x_{12} & a & \ldots & x_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ x_{1n} & x_{2n} & \ldots & a \end{bmatrix}}_{\mathbf{V}_n},$$

для которых справедливо несколько утверждений, доказательства которых вынесены в приложение.

**Лемма 1.** *Пусть матрица $\mathbf{\Lambda}_n$ имеет вид (3), и матрица $\mathbf{W}_n(k)$ получена из матрицы $\mathbf{V}_n$ через замену $k$-го столбца на столбик из констант $r$. Через $\Lambda_n$, $W_n(k)$ и $V_n$ обозначены соответствующие определители, $v_n = \mathbb{E}V_n$. Тогда для любых $k = 1, 2, \ldots, n$ справедливы равенства:*

$$1)\ \mathbb{E}W_n(k) = rv_{n-1}, \quad 2)\ \mathbb{E}\Lambda_n = rnv_{n-1} + v_n. \tag{4}$$

**Лемма 2.** *Обозначим нечетный факториал через $\phi_k = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2k-1) = \frac{(2k)!}{2^k k!}$ и рассмотрим функцию вида*

$$J_n(\sigma_0, r) = \sum_{k=0}^{\lfloor n/2 \rfloor} C_n^{2k} (-1)^k \phi_k \sigma_0^{2k} (1 - r)^{n-2k}, \ n = 1, 2, 3, \ldots. \tag{5}$$

*Для математического ожидания $v_n = \mathbb{E}V_n$ определителя матрицы $\mathbf{V}_n$ с диагональными элементами в виде $a = 1 - r$ и с независимыми случайными внедиагональными элементами $x_{ij} = x_{ji}$ с $\mathbb{E}x_{ij} = 0$ и дисперсией $\mathbb{D}x_{ij} = \sigma_0^2$ справедливы рекуррентное выражение*

$$v_n = av_{n-1} - (n-1)\sigma_0^2 v_{n-2}, \quad \text{где} \quad v_1 = a, \quad v_2 = a^2 - \sigma_0^2, \tag{6}$$

*и равенство*

$$v_n = J_n(\sigma_0, r)\,. \tag{7}$$

**Теорема 1.** *Пусть имеется матрица $\mathbf{\Lambda}_n$ со случайными недиагональными элементами $r + x_{ij}$ из (3), в которой $\mathbb{E}x_{ij} = 0$, $\mathbb{D}x_{ij} = \sigma_0^2$, $x_{ij} = x_{ji}$, функция $J_n(\sigma_0, r)$ имеет вид (5). Тогда для математического ожидания определителя и алгебраических дополнений $\tilde{\Lambda}_{n,kj} = (-1)^{k+j}\Lambda_{n,kj}$ матрицы $\mathbf{\Lambda}_n$, где $k \neq j$, справедливы соответственно равенства $\mathbb{E}\Lambda_n = rnJ_{n-1}(\sigma_0, r) + J_n(\sigma_0, r)$ и $\mathbb{E}\tilde{\Lambda}_{n,kj} = -rJ_{n-2}(\sigma_0, r)$.*

Полученные результаты можно использовать для построения ожидаемого наилучшего предсказания в случае, когда корреляционная матрица предикторов, в качестве которых рассматриваются частичные предсказания, имеет вид (3).

**4. Оценка полного прогноза по частичным предсказаниям.** Поскольку в дальнейшем речь пойдет только об алгебраических дополнениях, будем использовать для них обозначение $\Lambda_{n,kj} = \tilde{\Lambda}_{n,kj}$. Рассмотрим предсказание $F(\mathbf{f}) = \sigma\mathbf{Z}^T\mathbf{\Lambda}_n^{-1}\varrho = \sigma\sum\limits_{j=1}^{n} z_j\gamma_j$, где $\gamma_j = \frac{1}{\Lambda_n}\sum\limits_{i=1}^{n}\varrho_i\Lambda_{n,ij}$ — компоненты вектора $\gamma = \Lambda^{-1}\varrho$. Из теоремы 1 имеем

$$\mathbb{E}\Lambda_{n,ii} = r(n-1)J_{n-2} + J_{n-1}\,, \quad \mathbb{E}\Lambda_{n,ij} = -rJ_{n-2}\,,\ i \neq j, \tag{8}$$

где $J_n = J_n(\sigma_0, r)$ из (5). Применяя (8), вычислим математическое ожидание $\sum\limits_{i=1}^{n}\varrho_i\Lambda_{n,ij}$:

$$\mathbb{E}\sum_{i=1}^{n}\varrho_i\Lambda_{n,ij} = \mathbb{E}\left(\sum_{i=1,i\neq j}^{n}\varrho_i\Lambda_{n,ij} + \varrho_j\Lambda_{n,jj}\right) =$$

$$= -\sum_{i\neq j}^{n}\varrho_i rJ_{n-2} + \varrho_j(r(n-1)J_{n-2} + J_{n-1}) = rnJ_{n-2}(\varrho_j - \bar{\varrho}) + \varrho_j J_{n-1}\,.$$

Обозначим через $\delta_j = 1 - \frac{\bar{\varrho}}{\varrho_j}$, тогда $\mathbb{E}\left(\sum\limits_{i=1}^{n}\varrho_i\Lambda_{n,ij}\right) = \varrho_j(J_{n-1} + rnJ_{n-2}\delta_j)$. Используем $f_j = \sigma\varrho_j z_j$, а также полученные равенства для оценки полного прогноза $F(\mathbf{X})$ в виде общего частичного предсказания

$$f_b = \sigma\sum_{j=1}^{n}\frac{z_j}{\mathbb{E}\Lambda_n}\mathbb{E}\left(\sum_{i=1}^{n}\varrho_i\Lambda_{n,ij}\right) = \sum_{j=1}^{n}c_j f_j, \tag{9}$$

$$\text{где} \quad c_j = \frac{J_{n-1}(\sigma_0, r) + rnJ_{n-2}(\sigma_0, r)\delta_j}{J_n(\sigma_0, r) + rnJ_{n-1}(\sigma_0, r)}\,.$$

Итак, видим, что общее частичное предсказание $f_b = f_b(r, n, \sigma_0, \varrho)$ зависит от нескольких параметров. При отсутствии вариабельности ($\sigma_0 = 0$) в корреляциях между частичными предсказаниями и одинаковости множественных коэффициентов корреляции отдельных частичных предсказаний, т. е. $\varrho_i = \varrho_0$, имеем оценку $f_\alpha = f_b(r, n, 0, \varrho_0 \mathbf{1}^n)$ вида

$$f_\alpha = C_\alpha \bar{f}\,, \quad \text{где} \;\; C_\alpha = \frac{n}{1 + r(n-1)}\,, \tag{10}$$

которую будем называть сбалансированным средним частичным предсказанием. Помимо этого можно рассматривать оценки

$$f_\beta = C_\beta \bar{f}, \;\; \text{где} \quad C_\beta = \sum_{j=1}^{n} c_j, \tag{11}$$

и

$$f_\gamma = f_b(r, n, \sigma_0, \varrho_0 \mathbf{1}^n) = C_\gamma \bar{f}, \;\; \text{где} \;\; C_\gamma^{-1} = r + \frac{J_n(\sigma_0, r)}{n J_{n-1}(\sigma_0, r)}\,. \tag{12}$$

На рис. 1 представлены точки упорядоченного полного предсказания, построенного по моделированным данным при числе предикторов $n = 7$ и объеме выборки $N = 200$. Видно, насколько лучше, чем среднее частичное предсказание $\bar{f}$, полный прогноз аппроксимируется общими частичными предсказаниями вида $f_\alpha$ из (10) и $f_\beta$ из (11). На рис. 2 представлены результаты моделирования, в частности построены бокс-плоты средних квадратов отклонений полного прогноза от его оценок: $\bar{f}$, $f_\alpha$, $f_\beta$, $f_\gamma$. При помощи статистических критериев однородности получены значимые отличия $\bar{f}$ и $f_b$ между собой и от остальных оценок, в то время как отклонения от полного прогноза предсказаний $f_\beta$, $f_\gamma$, $f_\alpha$ значимо не отличаются.

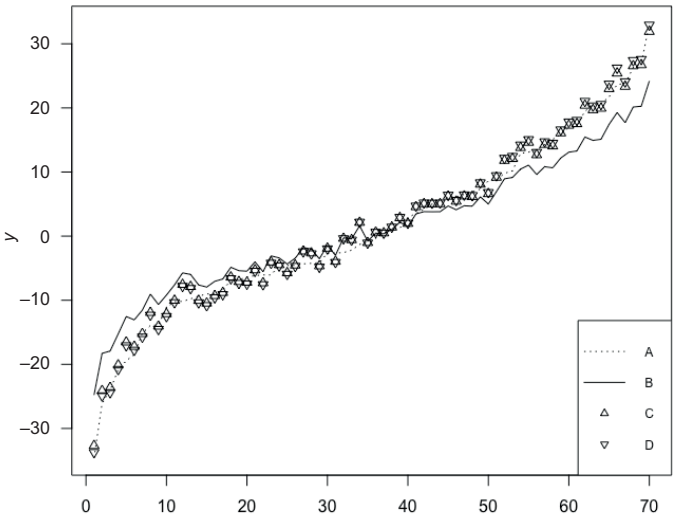


*Рис. 1.* Оценивание полного прогноза по модельной выборке. Предсказания: $A$ — полное $f_0$; $B$ — среднее частичное $\bar{f}$; $C$, $D$ — общие частичные $f_\beta$ и $f_\alpha$ соответственно.

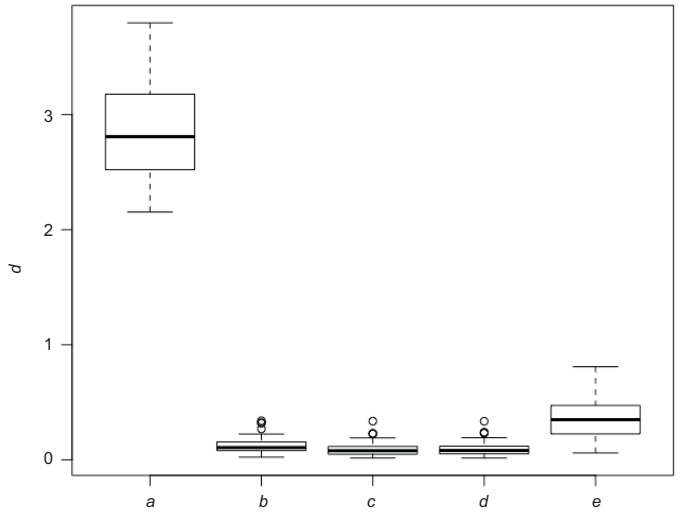


*Рис. 2.* Точность оценивания полного прогноза $f_0$ по квадрату $d$ его отклонения от предсказаний: $a$ — среднего $\bar{f}$; $b, c, d, e$ — от общих частичных $f_\beta$, $f_\gamma$, $f_\alpha$, $f_b$ соответственно.

**5. Оценки параметров регрессионной модели с переменными коэффициентами.** Опираясь на полученную оценку (9), можно уточнить полиномиальную структуру коэффициентов $\beta(\eta)$ из (1), где $\eta = (\eta_1, \ldots, \eta_n)$ — дихотомический вектор. Пусть $\mathcal{B}(\Omega_n)$ есть множество всех подмножеств из множества $\Omega_n = \{1, 2, \ldots, n\}$. Из $\mathcal{B}(\Omega_n)$, согласно некоторому правилу[1], во множестве $\mathcal{B}_0 \subseteq \mathcal{B}(\Omega_n)$ собираются соответствующие элементы $\tau$. Обозначим через $\mathbb{0}_\eta$ номера нулевых элементов вектора $\eta$ и через $N(\eta, \mathcal{B}_0)$ — число элементов множества $\{\tau : \tau \in \mathcal{B}_0\}$ таких, что $\tau \cup \mathbb{0}_\eta = \emptyset$. Определим подмножество $\mathcal{B}_j \subseteq \mathcal{B}_0$ как множество элементов из $\mathcal{B}_0$, содержащих элемент $j$, т. е. $\mathcal{B}_j = \{\tau : \tau \in \mathcal{B}_0, j \in \tau\}$. Обозначим через $a_{\tau,j}$ коэффициент частичной $\tau$-регрессии по независимой переменной $X_j$, $j \in \tau$. Тогда, применяя технологию усреднения частичных регрессий, коэффициенты из (1) можно представить в виде

$$\beta_j(\eta) = \frac{C}{N(\eta, \mathcal{B}_0)} \sum_{\tau \in \mathcal{B}_j} a_{\tau,j} \prod_{i \in \tau} \eta_i, \quad j = 1, 2, \ldots, n\,, \tag{13}$$

где $C \in \{1, C_\alpha, C_\beta, C_\gamma\}$. При $C = 1$ коэффициенты (13) приводят к оценке предсказания в виде $\bar{f}$. Остальным коэффициентам $C$ соответствуют оценки $f_\alpha, f_\beta, f_\gamma$. Очевидно, что прогноз осуществим только в случае $N(\eta, \mathcal{B}_0) \neq 0$.

**Пример 1.** *Пусть число переменных $n = 5$ и отобраны частичные предсказания по сочетаниям переменных из множества $\mathcal{B}_0 = \{\{1,2\}, \{1,2,3\}, \{1,2,4\}, \{3,5\}\}$. Тогда $\mathcal{B}_1 = \{\{1,2\}, \{1,2,3\}, \{1,2,4\}\}$ есть множество подмножеств, содержащих элемент 1. Аналогично строятся $\mathcal{B}_2 = \{\{1,2\}, \{1,2,3\}, \{1,2,4\}\}$, $\mathcal{B}_3 = \{\{1,2,3\}, \{3,5\}\}$, $\mathcal{B}_4 = \{\{1,2,4\}\}$, $\mathcal{B}_5 = \{\{3,5\}\}$. Используя формулу (13), получаем*

[1] Например, рассматриваются только такие сочетания независимых переменных, при которых множественные коэффициенты корреляции не ниже заданного уровня.

*коэффициенты регрессии в зависимости от реализации вектора* $\eta = (\eta_1, \dots, \eta_5)$*:*

$$\begin{aligned}
\beta_1(\eta) &= C(a_{12,1}\eta_1\eta_2 + a_{123,1}\eta_1\eta_2\eta_3 + a_{124,1}\eta_1\eta_2\eta_4)/N(\eta, \mathcal{B}_0),\\
\beta_2(\eta) &= C(a_{12,2}\eta_1\eta_2 + a_{123,2}\eta_1\eta_2\eta_3 + a_{124,2}\eta_1\eta_2\eta_4)/N(\eta, \mathcal{B}_0),\\
\beta_3(\eta) &= C(a_{123,3}\eta_1\eta_2\eta_3 + a_{35,3}\eta_3\eta_5)/N(\eta, \mathcal{B}_0),\\
\beta_4(\eta) &= Ca_{124,4}\eta_1\eta_2\eta_4/N(\eta, \mathcal{B}_0),\\
\beta_5(\eta) &= Ca_{35,5}\eta_3\eta_5/N(\eta, \mathcal{B}_0)\,.
\end{aligned}$$

*Например, для тех индивидов, у которых есть все пять наблюдений кроме третьего, т. е.* $\eta = (1,1,0,1,1)$*, получаем* $\mathbb{O}(\eta) = \{3\}$*. Из четырех элементов множества* $\mathcal{B}_0$ *не содержат* 3 *два элемента* $\{\{1,2\},\{1,2,4\}\}$*, поэтому* $N(\eta = (1,1,0,1,1), \mathcal{B}_0) = 2$*, и индивидуальный прогноз имеет вид*

$$\tilde{Y} = C\left((a_{12,1} + a_{124,1})X_1 + (a_{12,2} + a_{124,2})X_2 + a_{124,4}X_4\right)/2\,,$$

*а в случае* $\eta = (1,1,1,0,1)$ *получаем* $\mathbb{O}(\eta) = \{4\}$*. Из четырех элементов множества* $\mathcal{B}_0$ *не содержат* 4 *три элемента* $\{\{1,2\},\{1,2,3\},\{3,5\}\}$*, поэтому* $N(\eta = (1,1,1,0,1), \mathcal{B}_0) = 3$*, и прогноз определяется четырьмя переменными:*

$$\tilde{Y} = C\left(X_1(a_{12,1} + a_{123,1}) + X_2(a_{12,2} + a_{123,2}) + X_3(a_{123,3} + a_{35,3}) + X_5a_{35,5}\right)/3\,.$$

*В случае* $\eta = (0,1,1,1,1)$ *получаем прогноз в виде* $\tilde{Y} = C\left(X_3a_{35,3} + X_5a_{35,5}\right)$*. Таким образом, уравнения регрессии автоматически подстраиваются под структуру имеющихся данных. Возможны варианты реализаций вектора* $\eta$*, при которых прогнозирование невозможно осуществить, в частности при* $\eta = (0,1,0,1,1)$ *или* $\eta = (0,1,1,1,0)$*. В таком случае есть смысл пересмотреть правило отбора частичных предсказаний на предмет расширения множества* $\mathcal{B}_0$*.*

**6. Пример прогнозирования по неполным данным.** Для иллюстрации метода рассмотрим возможности прогнозирования по данным наблюдений «Многолетний мониторинг гидрологии и зоопланктона в Белом море: Картеш Д1» (https://www.st.nmfs.noaa.gov/copepod/time-series/ru-10101/). На рис. 3 представлены данные о численности зоопланктона *Calanus* в течение 58 лет наблюдений и линии регрессии $f_0$ и $f_\alpha$ в зависимости от девяти климатических показателей, таких как средняя температура гидрологических весны и лета, время окончания лета предыдущего года, сроки полного очищения ото льда, климатические индексы и т. д. Частичные регресии строились по $m = 8$ предикторам при исключении какого-то одного из девяти. Получены 9 частичных предсказаний для полных наблюдений и от 1 до 8 — для неполных. Вычислены значения $\bar{f}$, оценки $\hat{r} = 0.892$, $\hat{\sigma}_0 = 0.065$, $\hat{\varrho}_0 = 0.563, \hat{\sigma}_1 = 0.025$ и предсказание $f_\alpha$. Множественные коэффициенты корреляции для $f_0$ и $f_\alpha$ равны соответственно 0.6 ($n = 51$) и 0.58 ($n = 58$). На графике точки, для которых прогноз может быть вычислен только по частичным предсказаниям, выделены черным цветом.

При уменьшении параметра $m$, например при переходе к $m = 3$ по сравнению с $m = 8$, снижаются оценки $\hat{r} = 0.73$ и $\hat{\varrho}_0 = 0.42$ и увеличивается $\hat{\sigma}_0 = 0.24$. Множественный коэффициент корреляции снижается до 0.5, что кажется вполне закономерным явлением, поскольку часто уменьшение числа предикторов в частичных регессиях приводит к их большей разнородности и, как следствие, к ухудшению качества прогнозирования через сбалансированное среднее частичное предсказание.

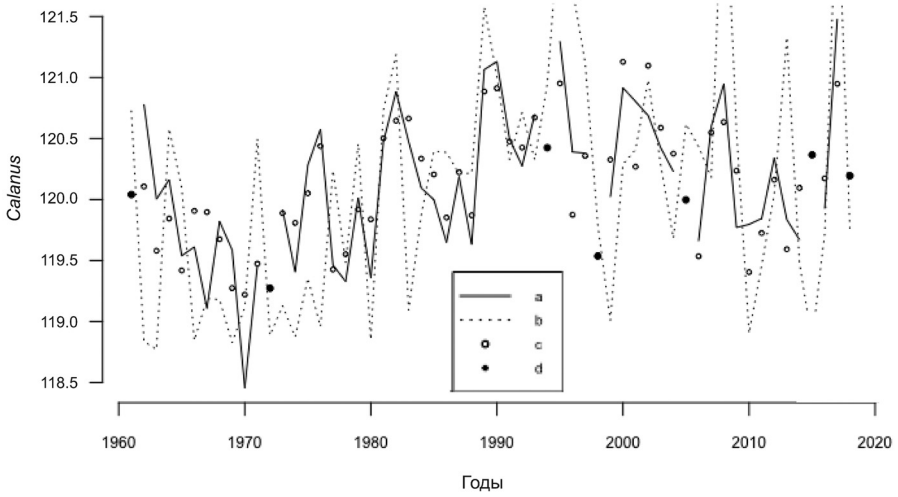


*Рис. 3.* Результаты прогнозирования численности планктона *Calanus* по климатическим показателям: $a$ — прогноз $f_0$ по полным данным; $b$ — зависимая переменная; $c, d$ — сбалансированное среднее частичное предсказание $f_\alpha$ с учетом пропущенных наблюдений.

Заметим, что не увеличивается только оценка $\hat{\sigma}_1 = 0.011$ (для сравнения: $\hat{\sigma}_1 = 0.025$ при $m = 8$), что свидетельствует об устойчивых паритетных свойствах частичных регрессий вне зависимости от качества прогнозирования. Это объясняется тем, что отбираются наиболее значимые частичные предсказания, в целом с одинаковой точностью прогнозирующие численность популяции.

**7. Заключение.** Таким образом, получены рекуррентные и явные выражения для алгебраических дополнений матрицы со случайными недиагональными элементами. Это позволило ввести мультипликативные поправки, улучшающие прогнозирование при помощи усреднения частичных предсказаний. Наиболее простое выражение имеет оценка $f_\alpha$, представляющая собой произведение среднего частичного предсказания на коэффициент, зависящий от степени связности частичных предсказаний и их количества. Более сложные оценки в виде линейных комбинаций частичных предсказаний с неодинаковыми весами объективно отличаются большей вариабельностью. Такой подход позволяет решить регрессионную задачу с переменным числом предикторов и может использоваться для корректировки любых метаоценок, в которых применяется усреднение частичных решений. В дальнейшем предполагается провести более подробное исследование эффективности полученных оценок. Помимо этого, предполагается адаптировать метод для оценки апостериорных вероятностей по неполным данным в задаче линейной классификации.

В практическом плане преимущество метода заключается в том, что удается охватить большее число наблюдений без искусственного заполнения пропусков. Это особенно важно при работе с медико-биологическими данными так называмого обсервационного типа, в которых по объективным причинам невозможно достичь полноты данных, и очень сложно убедить доктора, анализирующего истории болезни

своих пациентов, что кому-то можно приписать недостающие данные, какими бы они хорошими статистическими свойствами ни обладали.

**8. Приложение.**

Доказательство предложения 1.

1. Поскольку $\mathbf{Z}$ и $\varrho$ — векторы с компонентами $Z_i = \frac{X_i}{\sqrt{\sigma_{ii}}}$, $\varrho_i = \frac{l_i}{\sigma\sqrt{\sigma_{ii}}}$ соответственно, $i = 1, 2, \ldots, n$, и $\mathbf{\Sigma} = \mathbb{E}\mathbf{X}\mathbf{X}^T$, $\mathbf{l} = \mathbb{E}\,\mathbf{X}Y$, равенство $F(\mathbf{X}) = \mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{l}$ следует из (2) и из определения обратной матрицы. Для получения равенства $\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{l} = \sigma\mathbf{Z}^T\mathbf{\Lambda}^{-1}\varrho$ рассмотрим квадратную матрицу $S$ с элементами $\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{nn}}$ на главной диагонали и с нулями вне ее, при помощи которой можно выразить $\mathbf{X} = S\mathbf{Z}$, $\mathbf{l} = \sigma S\varrho$, $\mathbf{\Sigma} = S\mathbf{\Lambda}S$ и получить

$$\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{l} = \mathbf{Z}^T S S^{-1}\mathbf{\Lambda}^{-1}S^{-1}(\sigma S\varrho) = \sigma\mathbf{Z}^T\mathbf{\Lambda}^{-1}\varrho.$$

2. Учитывая центрированность $\mathbb{E}Y = 0$ и $\mathbb{E}X_i = 0$, получаем дисперсию предсказания $F(\mathbf{X})$:

$$\mathbb{D}F(\mathbf{X}) = \mathbb{E}F(\mathbf{X})^2 = \mathbb{E}F(\mathbf{X})F(\mathbf{X})^T = \mathbb{E}\mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{l} = \\ = \mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbb{E}X\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{l} = \mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbf{l}.$$

3. Вычисляем ковариацию $\mathbb{E}YF(\mathbf{X}) = \mathbb{E}Y\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{l} = \mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbf{l}$, а затем множественный коэффициент корреляции

$$R = \frac{\mathbb{E}YF(\mathbf{X})}{\sigma\sqrt{\mathbb{D}F(\mathbf{X})}} = \sigma^{-1}\sqrt{\mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbf{l}}, \text{ откуда } R^2 = \sigma^{-2}\mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbf{l}, \quad \mathbb{D}F(\mathbf{X}) = \sigma^2R^2\,.$$

4. Для вектора $\mathbf{W} = \mathbf{A}\mathbf{X}$ ковариационная матрица имееет вид $L_w = \mathbb{E}(\mathbf{A}\mathbf{X})(\mathbf{A}\mathbf{X})^T = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T$, а вектор вторых смешанных моментов $\mathbf{l}_w = \mathbf{A}\mathbf{l}$, откуда получаем

$$F(\mathbf{A}\mathbf{X}) = \mathbf{W}^T{\mathbf{L}_\mathbf{w}}^{-1}\mathbf{l}_w = \mathbf{X}^T\mathbf{A}^T(\mathbf{A}^T)^{-1}\mathbf{\Sigma}^{-1}\mathbf{A}^{-1}\mathbf{A}\mathbf{l} = \mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{l} = F(\mathbf{X})\,.$$

Доказательство леммы 1.

1. Сначала покажем, что $\mathbb{E}W_n(k) = rv_{n-1}$. Обозначая через $a = 1-r$, рассмотрим определитель $W_n(1)$, который получен из определителя матрицы $\mathbf{V}_n$ через замену первого столбца на константы $r$:

$$W_n(1) = \begin{vmatrix} r & x_{12} & x_{13} & \ldots & x_{1n} \\ r & a & x_{23} & \ldots & x_{2n} \\ r & x_{23} & a & \ldots & x_{3n} \\ \vdots & \vdots & & \ldots & \vdots \\ r & x_{2n} & x_{3n} & \ldots & a \end{vmatrix}. \tag{14}$$

Покажем, что $\mathbb{E}W_n(1) = r\mathbb{E}V_{n-1}$. Действительно, при разложении определителя по первому столбцу первое слагаемое вида $r\mathbb{E}V_{n-1} \neq 0$, в то время как остальные математические ожидания равны нулю. Например, минор $W_{n,21}(1)$ имеет вид

$$W_{n,21}(1) = \begin{vmatrix} x_{12} & x_{13} & \ldots & x_{1n} \\ x_{23} & a & \ldots & x_{3n} \\ \vdots & & & \vdots \\ x_{2n} & x_{3n} & \ldots & a \end{vmatrix} \tag{15}$$

и математическое ожидание минора $\mathbb{E}W_{n,21}(1) = 0$ за счет того, что имеются одна строка и один столбец, компоненты которых имеют нулевые математические ожидания. Для $k > 1$ все аналогично. При вычислении $\mathbb{E}W_n(k)$ аналогично ненулевым будет $k$-й элемент $k$-го столбца. Таким образом, утверждение $\mathbb{E}W_n(k) = rv_{n-1}$ справедливо вне зависимости от того, какое место занимает столбец из одинаковых значений $r$.

2. Для доказательства $\mathbb{E}\Lambda_n = rnv_{n-1} + v_n$ представим определитель $\Lambda_n$ в виде суммы по двум компонентам первого столбца:

$$\Lambda_n = \begin{vmatrix} r+a & r+x_{12} & \dots & r+x_{1n} \\ r+x_{12} & r+a & \dots & r+x_{23} \\ \vdots & \vdots & \dots & \vdots \\ r+x_{1n} & r+x_{2n} & \dots & r+a \end{vmatrix} =$$

$$= \begin{vmatrix} r & r+x_{12} & \dots & r+x_{1n} \\ r & r+a & \dots & r+x_{23} \\ \vdots & \vdots & \dots & \vdots \\ r & r+x_{2n} & \dots & r+a \end{vmatrix} + \begin{vmatrix} a & r+x_{12} & \dots & r+x_{1n} \\ x_{12} & r+a & \dots & r+x_{23} \\ \vdots & \vdots & \dots & \vdots \\ x_{1n} & r+x_{2n} & \dots & r+a \end{vmatrix}.$$

Очевидно первый определитель можно привести к виду $W_n(1)$ из (14), а второй — раскладываем аналогично по второму столбцу и т. д. Отсюда

$$\Lambda_n = W_n(1) + W_n(2) + \ldots + W_n(n) + V_n. \tag{16}$$

Переходя к математическому ожиданию, получаем необходимое выражение на основании предыдущего утверждения данной леммы. □

Доказательство леммы 2.

Справедливость (6) следует из разложения определителя $V_n$ по элементам первого столбца. Далее убеждаемся, что для (7) справедливо (6). Пусть $n = 2m$. Тогда

$$v_n = J_n(\sigma_0, r) = \sum_{k=0}^{\lfloor n/2 \rfloor} C_n^{2k} (-1)^k \phi_k \sigma_0^{2k} a^{n-2k},$$

$$av_{n-1} - (n-1)\sigma_0^2 v_{n-2} = \sum_{k=0}^{m-1} C_{2m-1}^{2k} (-1)^k \phi_k \sigma_0^{2k} a^{2m-2k} +$$

$$+(2m-1) \sum_{k=0}^{m-1} C_{2m-2}^{2k} (-1)^{k+1} \phi_k \sigma_0^{2k+2} a^{2m-2k-2}.$$

Во втором выражении поменяем индекс суммирования $t = k + 1$ и отделим крайние слагаемые:

$$\sum_{k=0}^{m-1} C_{2m-1}^{2k} (-1)^k \phi_k \sigma_0^{2k} a^{2m-2k} + \sum_{t=1}^{m} (2m-1) C_{2m-2}^{2t-2} (-1)^t \phi_{t-1} \sigma_0^{2t} a^{2m-2t} =$$

$$= a^{2m} + \sum_{k=1}^{m-1} C_{2m-1}^{2k} (-1)^k \phi_k \sigma_0^{2k} a^{2m-2k} +$$

$$+\sum_{t=1}^{m-1}(2m-1)C_{2m-2}^{2t-2}(-1)^t\phi_{t-1}\sigma_0^{2t}a^{2m-2t}+(-1)^m\phi_m\sigma_0^{2m}=$$
$$=a^{2m}+B+(-1)^m\phi_m\sigma_0^{2m}\,.$$

Заменим $\phi_{t-1}=\frac{\phi_t}{2t-1}$ и обозначим через $C_k=(-1)^k\sigma_0^{2k}a^{2m-2k}\phi_k$,

$$B=\sum_{k=1}^{m-1}C_k\left(\frac{(2m-1)!}{(2k)!(2m-1-2k)!}+\frac{(2m-2)!}{(2k-2)!(2m-2k)!}\cdot\frac{2m-1}{2k-1}\right).$$

Вычислим отдельно выражение в скобках:

$$\frac{(2m-1)!}{(2k-1)!(2m-2k-1)!}\left(\frac{1}{2k}+\frac{1}{2m-2k}\right)=C_{2m}^{2k}\,.$$

Таким образом, $av_{n-1}-(n-1)\sigma_0^2v_{n-2}=v_n$, и (7) верно при $n=2m$. Аналогично при $n=2m+1$ покажем, что $av_{2m}-2m\sigma^2v_{2m-1}=v_{2m+1}$:

$$av_{2m}-2m\sigma^2v_{2m-1}=\sum_{k=0}^{m}C_{2m}^{2k}(-1)^k\phi_k\sigma_0^{2k}a^{2m-2k+1}+$$
$$+2m\sum_{k=0}^{m-1}C_{2m-1}^{2k}(-1)^{k+1}\phi_k\sigma_0^{2k+2}a^{2m-2k-1}=$$
(применим обозначения $k=t-1,\quad \phi_{t-1}=\phi_t(2t-1), C_k=\phi_k(-1)^k\sigma_0^{2k}a^{2m-2k+1}$)
$$=a^{2m+1}+\sum_{k=1}^{m}C_{2m}^{2k}(-1)^k\phi_k\sigma_0^{2k}a^{2m-2k+1}+\sum_{t=1}^{m}2mC_{2m-1}^{2t-2}(-1)^t\phi_{t-1}\sigma_0^{2t}a^{2m-2t+1}=$$
$$=a^{2m+1}+\sum_{k=1}^{m}C_k\left(\frac{(2m)!}{(2k)!(2m-2k)!}+\frac{(2m-1)!2m}{(2k-2)!(2k-1)(2m-2k+1)!}\right).$$

Выражение в скобках имеет вид

$$\frac{(2m)!}{(2k-1)!(2m-2k)!}\left(\frac{1}{2k}+\frac{1}{2m-2k+1}\right)=C_{2m+1}^{2k}\,.\qquad\square$$

Доказательство теоремы 1.

Из лемм 1 и 2 следует утверждение $\mathbb{E}\Lambda_n=rn\mathbb{E}V_{n-1}+\mathbb{E}V_n=rnJ_{n-1}(\sigma_0,r)+J_n(\sigma_0,r)$, где $\Lambda_n$ — определитель матрицы $\mathbb{\Lambda}_n$. Для того чтобы получить второе утверждение, рассмотрим определитель $W_n(1)$, который получен из определителя матрицы $\mathbf{V}_n$ через замену первого столбца на константы $r$, и соответствующий минор. Итак,

$$W_n(1)=\begin{vmatrix} r & x_{12} & x_{13} & \dots & x_{1n}\\ r & a & x_{23} & \dots & x_{2n}\\ r & x_{23} & a & \dots & x_{3n}\\ \vdots & \vdots & & & \vdots\\ r & x_{2n} & x_{3n} & \dots & a \end{vmatrix},\qquad W_{n,21}(1)=\begin{vmatrix} x_{12} & x_{13} & \dots & x_{1n}\\ x_{23} & a & \dots & x_{3n}\\ \vdots & \dots & \dots & \vdots\\ x_{2n} & x_{3n} & \dots & a \end{vmatrix}. \quad (17)$$

Обозначим через $A_{n-1} = W_{n,21}(1)$ и $A_{n-1}(i)$ определители, которые получены в результате замены $i$-го столбца в $A_{n-1}$ столбцом из $r$. Например,

$$A_{n-1}(1) = \begin{vmatrix} r & x_{13} & \dots & x_{1n} \\ r & a & \dots & x_{3n} \\ \vdots & & & \vdots \\ \vdots & \dots & \dots & \vdots \\ r & x_{3n} & \dots & a \end{vmatrix}, \quad A_{n-1}(2) = \begin{vmatrix} x_{12} & r & x_{14} & \dots & x_{1n} \\ x_{23} & r & x_{34} & \dots & x_{3n} \\ x_{24} & r & a & \dots & x_{4n} \\ \vdots & \dots & & \dots & \vdots \\ x_{2n} & r & x_{4n} & \dots & a \end{vmatrix}.$$

По аналогии с (16) представим минор $\Lambda_{n,21}$ в виде суммы

$$\Lambda_{n,21} = \begin{vmatrix} r + x_{12} & r + x_{13} & \dots & r + x_{1n} \\ r + x_{23} & r + a & \dots & r + x_{3n} \\ \vdots & & & \vdots \\ \vdots & \dots & \dots & \vdots \\ r + x_{2n} & r + x_{3n} & \dots & r + a \end{vmatrix} = \sum_{i=1}^{n-1} A_{n-1}(i) + A_{n-1}\,.$$

Согласно лемме 1, $\mathbb{E}A_{n-1} = \mathbb{E}W_{n,21}(1) = 0$. Далее при разложении определителя $A_{n-1}(1)$ по первому столбцу получаем $\mathbb{E}A_{n-1}(1) = r\mathbb{E}V_{n-2} = rv_{n-2}$, так как в случае $k \neq 1$ имеет место $\mathbb{E}[A_{n-1}(1)]_{k1} = 0$. При $i > 1$, раскладывая определитель $A_{n-1}(i)$ по столбцу с константами, получим определитель матрицы, у которой есть хотя бы одна строка и один столбец с элементами $x_{ij}$, поэтому, переходя к математическому ожиданию, получаем ноль. Следовательно, $\mathbb{E}A_{n-1}(i) = 0$ для $i > 1$. Таким образом, для минора $\Lambda_{n,21}$ справедливо $\mathbb{E}\Lambda_{n,21} = rv_{n-2} = rJ_{n-2}(\sigma_0, r)$, а среднее алгебраическое дополнение имеет вид $-rJ_{n-2}(\sigma_0, r)$. Остальные алгебраические дополнения путем соответствующего числа перестановок столбцов или строк сводятся к $\Lambda_{n,21}$. Например, несложно показать, что $\mathbb{E}\Lambda_{n,31} = -\mathbb{E}\Lambda_{n,21} = -rJ_{n-2}(\sigma_0, r)$. Остальные аналогично. Таким образом, можно утверждать, что для любых $i \neq j$ имеет место $\mathbb{E}(-1)^{i+j}\Lambda_{ij} = -rJ_{n-2}(\sigma_0, r)$. □

**Литература**


1. Vink G., Frank L. E., Pannekoek J., Buuren S. Predictive mean matching imputation of semi-continuous variables. *Statistica Neerlandica* **68** (1), 61–90 (2014). https://doi.org/10.1111/stan.12023

2. Van Buuren S., Groothuis-Oudshoorn C. mice: Multivariate imputation by chained equations in R. *Journal of Statistimathbf Software* **45** (3), 1–67 (2011).

3. Alexeyeva N. Dual balance correction in Repeated Measures ANOVA with missing data. *Electronic Journal of Applied Statistical Analysis* **10** (1), 146–159 (2017). https://doi.org/10.1285/i20705948v10n1p146

4. Барт А. Г. *Анализ медико-биологических систем. Метод частично обратных функций.* Санкт-Петербург, Изд-во С.-Петерб. ун-та (2003).

5. Ho Tin Kam. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (8), 832–844 (1998). https://doi.org/10.1109/34.709601

6. Крамер Г. *Математические методы статистики,* пер. с англ. Москва, Мир (1975).

7. Ho Tin Kam. Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition.* Montreal, QC, 14–16 August 1995, 278–282 (1995).

8. Алексеева Н. П., Горлова И. А., Бондаренко Б. Б. Возможности прогнозирования артериальной гипертензии на основе метода проективной классификации. *Артериальная гипертензия* **23** (5), 472–480 (2017). https://doi.org/10.18705/1607-419X-2017-23-5-472-480

9. Эсмедляева Д. С., Алексеева Н. П., Новицкая Т. А., Дьякова М. Е., Ариэль Б. М., Соколович Е. Г. Активность воспалительного процесса и маркеры деструкции внеклеточного матрикса при туберкулёме легких. *Бюллетень сибирской медицины* **19** (2), 112–119 (2020). https://doi.org/10.20538/1682-0363-2020-2-112-119

10. Afifi A. A., Azen S. P. *Statistical Analysis. A Computer Oriented Approach.* 2nd ed. New York; San Francisco; London, Academic Press (1979).

Статья поступила в редакцию 5 января 2022 г.;
доработана 20 апреля 2022 г.;
рекомендована к печати 9 июня 2022 г.

Контактная информация:

*Алексеева Нина Петровна* — канд. физ.-мат. наук, доц.; nina.alekseeva@spbu.ru
*Ал-Джубури Фатима Садик Шуккир* — аспирант; fatema_sadik79@yahoo.com

## About the full prediction approximation by a lot of partial predictions in case of incomplete data*

*N. P. Alexeyeva*[1], *F. S. Sh. Al-Juboori*[2]

[1] St Petersburg State University, 7–9, Universitetskaya nab., St Petersburg, 199034, Russian Federation
[2] University of Information Technology and Communications,
Iraq, Baghdad, St Al-Nidal

**For citation:** Alexeyeva N. P., Al-Juboori F. S. Sh. About the full prediction approximation by a lot of partial predictions in case of incomplete data. *Vestnik of Saint Petersburg University. Mathematics. Mechanics. Astronomy*, 2022, vol. 9 (67), issue 4, pp. 575–589.
https://doi.org/10.21638/spbu01.2022.401 (In Russian)

In this article, we are talking about the random subspaces method in forecasting under the condition of incomplete data and about estimation of a full forecast based on a set of partial predictions. Centered partial predictions are considered without loss of generality. According to the statistical model, off-diagonal elements in the correlation matrix of partial predictions are considered random with known mathematical expectation and variance. In case of this random matrix, analytical expressions are obtained for the mathematical expectation of the determinant and minors. Based on these results, a class of more accurate estimates of the full prediction is constructed, which differ from the mean partial prediction by a multipliers that depend on the statistical parameters of the correlation matrix of partial predictions. The results of modeling and practical forecasting based on incomplete biogeographic data are presented.

*Keywords*: the random subspace method, statistical model, matrix with random elements, partial predictions, multiple regression.

### References

1. Vink G., Frank L. E., Pannekoek J., Buuren S. Predictive mean matching imputation of semi-continuous variables. *Statistica Neerlandica* **68** (1), 61–90 (2014). https://doi.org/10.1111/stan.12023

2. Van Buuren S., Groothuis-Oudshoorn C. mice: Multivariate imputation by chained equations in R. *Journal of Statistimathbf Software* **45** (3), 1–67 (2011).

3. Alexeyeva N. Dual balance correction in Repeated Measures ANOVA with missing data. *Electronic Journal of Applied Statistical Analysis* **10** (1), 146–159 (2017). https://doi.org/10.1285/i20705948v10n1p146

4. Bart A. G. Analysis of biomedical systems. Method partially inverse functions. St Petersburg, St Petersburg University Press (2003). (In Russian)

5. Ho Tin Kam. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (8), 832–844 (1998). https://doi.org/10.1109/34.709601

*This work was supported by the Russian Foundation for Basic Research (project no. 20-01-00096).

6. Cramer H. *Mathematical Methods Of Statistics.* Asia Publishing House (1975) [Rus. ed.: Cramer H. Matematicheskie metodu statistiki, Mir Publ. (1975)].

7. Ho Tin Kam. Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14–16 August 1995, 278–282 (1995).

8. Alexeyeva N. P., Gorlova I. A., Bondarenko B. B. Possibilities of predicting arterial hypertension based on the method of projective classification. *Arterial hypertension.* **23** (5), 472–480 (2017). https://doi.org/10.18705/1607-419X-2017-23-5-472-480 (In Russian)

9. Esmedlyaeva D. S., Alexeyeva N. P., Novitskaya T. A., Dyakova M. E., Ariel B. M., Sokolovich E. G. Inflammatory process activity and markers of extracellular matrix destruction in pulmonary tuberculoma. *Bulletin of Siberian Medicine* **19** (2), 112–119 (2020). https://doi.org/10.20538/1682-0363-2020-2-112-119 (In Russian)

10. Afifi A. A., Azen S. P. *Statistical Analysis. A Computer Oriented Approach.* 2nd ed. New York; San Francisco; London, Academic Press (1979).

Received: January 5, 2022
Revised: April 20, 2022
Accepted: June 27, 2022

Authors' information:

*Nina P. Alexeyeva* — nina.alekseeva@spbu.ru
*Fatema S. Sh. Al-Juboori* — fatema_sadik79@yahoo.com