DOI: 10.26896/1028-6861-2019-85-1-I-78-86

О ПРОБЛЕМЕ ЗНАЧИМОСТИ ФАКТОРОВ МОДЕЛЕЙ ДИСПЕРСИОННОГО АНАЛИЗА

© **Нина Даниловна Бекарева, Владислав Юрьевич Щеколдин**

Новосибирский государственный технический университет, г. Новосибирск, Россия; e-mail: raix@mail.ru

Статья поступила 20 февраля 2018 г. Поступила после доработки 20 февраля 2018 г. Принята к публикации 25 ноября 2018 г.

В моделях дисперсионного анализа гипотезы о значимости факторов образуют сравнения уровней одного и того же фактора. Если по критерию проверки таких гипотез нет оснований отвергнуть гипотезу о равенстве некоторых уровней факторов, то эти уровни признаются значимо отличающимися друг от друга. Цель данной работы — попытка выяснить, влияют ли значимо уровни факторов по отдельности на исследуемую величину. Исследованы линейные модели дисперсионного анализа в условиях полного и неполного факторного эксперимента. В качестве вычислительной схемы обработки моделей дисперсионного анализа разработаны процедура преобразования моделей неполного ранга в модели полного ранга и представление пространства линейных оцениваемых форм в виде прямой суммы ортогональных подпространств, соответствующих каждому из качественных факторов модели. Выбирая в исходной матрице наблюдений различные группы линейно независимых столбцов и ортогонализируя эту систему векторов — столбцов матрицы, можно получить различные ортогональные базисы пространства линейных оцениваемых форм. Проекции вектора отклика на векторы ортогонального базиса, соответствующие одному и тому же базису, определяют вклад данного фактора в общую сумму квадратов, получаемую в результате проекции отклика на все пространство линейных оцениваемых форм в целом. При различных ортогональных базисах пространства линейных оцениваемых форм эти вклады факторов меняются. В условиях неполного факторного эксперимента можно выделить такой ортогональный базис пространства линейных оцениваемых форм, который обеспечивает наибольшую значимость какого-либо сравнения одного из факторов, а следовательно, и самого фактора. Это позволяет выявить наилучшую в определенном смысле схему проведения эксперимента, обеспечивающую наибольшую значимость выделяемых факторов. Для подтверждения полученных результатов применен метод ранжирования факторов LASSO.

Ключевые слова: дисперсионный анализ; качественные факторы; пространство линейных оцениваемых форм; сравнения; модели полного и неполного ранга; метод LASSO.

ON THE PROBLEM OF FACTOR SIGNIFICANCE IN THE MODELS OF VARIANCE ANALYSIS

© **Nina D. Bekareva, Vladislav Yu. Shchekoldin**

Novosibirsk State Technical University, Novosibirsk, Russia; e-mail: raix@mail.ru

Received February 20, 2018. Revised February 20, 2018. Accepted November 25, 2018.

In analysis of variance models the hypotheses about factor significance form comparisons of the levels of the same factor. If there are no reasons according to the criterion for testing such hypotheses to reject the hypothesis about the equality of some levels of factors, then these levels are considered significantly different. The goal of the article is to find out whether the levels of factors have a significant effect on the response variable. Linear models of analysis of variance (AV) under conditions of complete and incomplete factor design are considered. As a computational scheme for processing the AV models the procedure for transforming incomplete rank models into a full rank models and representing the space of linear estimated forms as a direct sum of orthogonal subspaces corresponding to each of the qualitative facts of the model has been developed. Choosing different groups of linearly independent columns in the initial observation matrix and orthogonalizing this system of vectors (as columns of the matrix) we can obtain various orthogonal bases for the space of linear forms under estimation. Projections of the response vector to the vectors of orthogonal basis corresponding to the same basis determine the contribution of this factor to the total sum of squares obtained as a result of the response projection to the entire space of linear forms being estimated. With different orthogonal bases of the space of linear estimated forms, these contributions of factors change. Under conditions of incomplete factor design, one can distinguish an orthogonal basis of the space of linear estimated forms which provides the greatest significance of any comparison of one of the factors, and, consequently, of the factor itself. This allows determination of the best (in a certain sense) experimental design which ensures the greatest significance of the factors selected. To prove the results obtained the method of ranking factors LASSO was used.

Keywords: analysis of variance; qualitative factors; space of linear estimated forms; comparisons; full and incomplete rank models; LASSO.

Дисперсионный анализ (ДА) как один из методов статистической обработки данных известен с начала прошлого века. Теория этого метода в значительной мере развита в работах Р. Фишера, который и ввел в статистику термины «дисперсия», «дисперсионный анализ». Спецификой моделей ДА является невозможность прямого применения стандартных методов для получения оценок неизвестных параметров моделей вследствие качественной природы входных факторов. За последнее время вычислительные методы ДА претерпели существенные изменения, связанные с их универсализацией. В данной работе рассмотрена вычислительная схема, позволяющая использовать метод наименьших квадратов для получения оценок сравнений, относительно которых традиционно проверяются гипотезы ДА.

Рассмотрим линейные модели ДА с детерминированными факторами. Такие модели в матричной форме обычно записывают в виде [1]

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}. \tag{1}$$

Здесь $\mathbf{Y}$ — n-мерная случайная величина (с.в.), имеющая областью значений некоторое подмножество пространства R^n; $\mathbf{X}$ — матрица наблюдений (планирования, план эксперимента) размерности $n \times p$; $\boldsymbol{\theta}$ — p-мерный вектор неизвестных параметров модели; $\mathbf{e}$ — аддитивная случайная составляющая модели, относительно которой выдвигаются предположения гомоскедастичности и некоррелированности.

Подробнее о планировании эксперимента для моделей ДА можно ознакомиться, например, в работе [2].

При проверке статистических гипотез обычно предполагают, что случайная ошибка распределена по многомерному нормальному закону, т.е.

$$\mathbf{e} \sim N(0, \sigma^2\mathbf{I}), \tag{2}$$

где σ^2 — дисперсия ошибки наблюдения случайной величины $\mathbf{Y}$ в каждой точке наблюдений; $\mathbf{I}$ — единичная матрица порядка n. При нарушении условия (2) целесообразно применять методы непараметрического ДА [3].

В матрице $\mathbf{X}$ каждому фактору соответствует число столбцов, равное числу его уровней. Если число факторов в модели (1) равно m, а число уровней каждого фактора — s_i, $i = \overline{1, m}$, то число столбцов матрицы $\mathbf{X}$ составит

$$p = 1 + \sum_{i=1}^{m} s_i. \tag{3}$$

Первое слагаемое в этой формуле, равное единице, означает, что в модели всегда присутствует некоторый параметр, обозначаемый в векторе $\boldsymbol{\theta}$ через μ, называемый генеральным средним модели. В матрице $\mathbf{X}$ ему соответствует первый столбец, состоящий из единиц. Остальные столбцы этой матрицы состоят из нулей и единиц, причем единицей в s_i-х столбцах i-го фактора отмечается тот столбец, который соответствует уровню фактора, на котором тот присутствует в наблюдении. Принято обозначать следующие за первым столбцом s_1 столбцов матрицы через $\alpha_1, \alpha_2, ..., \alpha_{s1}$, этими же символами обозначают s_1 параметров в векторе $\boldsymbol{\theta}$; следующие s_2 столбцов матрицы — через $\beta_1, \beta_2, ..., \beta_{s2}$, это уровни второго фактора, в векторе $\boldsymbol{\theta}$ им соответствуют также эти s_2 символов, и т.д.

Такое построение матрицы $\mathbf{X}$ приводит к тому, что ее столбцы образуют линейно зависимую систему n-мерных векторов, поскольку сумма столбцов матрицы, соответствующих каждому фактору, равна единичному столбцу. Наибольшее число линейно независимых столбцов матрицы $\mathbf{X}$ бывает в полном факторном эксперименте (ПФЭ). Это число равно рангу матрицы $\mathbf{X}$ и вычисляется по формуле

$$r := rg\mathbf{X} = 1 + \sum_{i=1}^{m} (s_i - 1) = p - m. \tag{4}$$

Полным факторным экспериментом называют дисперсионную модель с матрицей $\mathbf{X}$, содержащей максимальное число различных точек наблюдения (строк матрицы $\mathbf{X}$), при котором каждый уровень каждого фактора встречается с каждым уровнем остальных факторов ровно один раз [4, 5]. Следовательно, число таких наблюдений определяется как $n = s_1 s_2 ... s_m$. Из последнего соотношения очевидно, что $r < p \le n$, при этом левая часть неравенства позволяет называть все модели ДА моделями неполного ранга.

В качестве примера рассмотрим модель ПФЭ (1) с $m = 2$, $s_1 = 2$, $s_2 = 3$. Матрицу $\mathbf{X}$ для моделей вида (1) задают обычно в виде $\mathbf{X}_R$ размерности $n \times$ m, в которой i-я строка содержит номера тех уровней факторов, на которых они присутствуют в i-м наблюдении. Тогда

$$\mathbf{X}_R = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 2 & 1 \\ 2 & 2 \\ 2 & 3 \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \ \boldsymbol{\theta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Очевидно, что для этой модели $r = 4$, $p = 6$, $n = 6$.

Основной задачей при исследовании моделей ДА является выяснение значимости влияния каждого фактора на исследуемую с.в. $\mathbf{Y}$.

Обобщенное обращение матриц в моделях дисперсионного анализа

Критерии, позволяющие оценить значимость как модели (1) в целом, так и каждого из m факторов, содержат статистики, определяемые МНК-оценками вектора неизвестных параметров модели $\hat{\boldsymbol{\theta}}$. Поэтому в первую очередь должна быть решена задача получения вектора $\hat{\boldsymbol{\theta}}$. В силу особенности матрицы $\mathbf{X}$ применение стандартной формулы Гаусса $\hat{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ для получения таких оценок невозможно.

В качестве универсального метода вычисления оценок $\hat{\boldsymbol{\theta}}$ в модели ДА используется процедура обобщенного обращения матриц. Следуя работам [6, 7], кратко опишем схему этой вычислительной процедуры.

Определение. Обобщенно обратной матрицей некоторой матрицы $\mathbf{A}$ называют любую матрицу $\mathbf{A}^-$, такую, что $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$.

Решением системы нормальных уравнений

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^{\mathrm{T}}\mathbf{Y} \qquad (5)$$

в случае вырожденной матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ служит вектор

$$\tilde{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^-\mathbf{X}^{\mathrm{T}}\mathbf{y} + (\mathbf{H} - \mathbf{I})\mathbf{z}, \qquad (6)$$

где $\mathbf{H} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^-\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\mathbf{z}$ — произвольный вектор.

В этой схеме матрица $\mathbf{H}$ важна не только тем, что линейная оболочка ее строк определяет пространство линейных оцениваемых форм (ЛОФ) данной модели при заданной матрице $\mathbf{X}$. Наибольший интерес в этом пространстве вызывают сравнения — линейные функции уровней одного и того же фактора, сумма коэффициентов которых равна нулю. В ДА именно сравнения формируют статистические гипотезы. Но пространство ЛОФ может и не содержать сравнений — линейных функций уровней каких-либо факторов или даже всех факторов, это зависит от матрицы наблюдений $\mathbf{X}$. Ответ на вопрос, имеются ли в пространстве ЛОФ сравнения, т.е. могут ли быть сформированы гипотезы о значимости факторов модели, получают с помощью матрицы $\mathbf{H}$. Справедливо утверждение [6]: линейная функция $\mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}$ является сравнением тогда и только тогда, когда выполняется соотношение $\mathbf{c}^{\mathrm{T}}\mathbf{H} = \mathbf{c}^{\mathrm{T}}$.

Отметим, что полученная оценка $\tilde{\boldsymbol{\theta}}$ (6) является смещенной:

$$M\tilde{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^-\mathbf{X}^{\mathrm{T}}M\mathbf{y} + (\mathbf{H} - \mathbf{I})\mathbf{z} = \mathbf{H}\boldsymbol{\theta} + (\mathbf{H} - \mathbf{I})\mathbf{z}$$

даже при $\mathbf{z} = 0$. Но важным является следующее утверждение: для произвольного вектора $\mathbf{c}$ величина $\mathbf{c}^{\mathrm{T}}\tilde{\boldsymbol{\theta}}$ сохраняется одной и той же для всех решений системы нормальных уравнений (5) [1, 6]. Это утверждение эквивалентно тому, что $\mathbf{c}^{\mathrm{T}}\mathbf{H} = \mathbf{c}^{\mathrm{T}}$ и существует линейная функция $\mathbf{y}$, математическое ожидание которой равно $\mathbf{c}^{\mathrm{T}}\boldsymbol{\theta}$.

Построение моделей дисперсионного анализа полного ранга

В работе предлагается такая процедура получения обобщенной обратной матрицы $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^-$ [6], которая значительно упрощает вычислительную схему в моделях ДА и позволяет перейти от модели неполного ранга (1) к модели ДА полного ранга. Опишем эту вычислительную схему.

I. Получаем скелетное разложение матрицы $\mathbf{X}$ [8], т.е. ее представление в виде произведения двух матриц:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{A}, \qquad (7)$$

где $\mathbf{T}$ — матрица полного столбцового ранга размерности $n \times r$, состоящая из линейно независимых столбцов матрицы $\mathbf{X}$; $\mathbf{A}$ — матрица размерности $r \times p$.

Такое разложение можно получить по схеме Гаусса с обратными ходами всякий раз, как только очередной столбец матрицы $\mathbf{X}$ окажется зависимым от предыдущих ее столбцов. Пусть q первых столбцов матрицы $\mathbf{X}$ образуют линейно независимую систему векторов, $1 \le q < r$. Они переписываются в матрицу $\mathbf{T}$, одновременно с этим заполняются q первых столбцов матрицы $\mathbf{A}$, которые образуют q нулевых векторов с единицей в каждом из них на q-м месте. Как только $(q + 1)$-й столбец матрицы $\mathbf{X}$ оказывается линейно зависимым от предыдущих q столбцов, обратным ходом метода Гаусса он представляется в виде их линейной комбинации, коэффициенты которой записываются в $(q + 1)$-й столбец матрицы $\mathbf{A}$. Далее из матрицы $\mathbf{X}$ удаляется $(q + 1)$-й столбец, и продолжает работать схема Гаусса. Результат скелетного разложения матрицы из рассмотренного выше примера имеет вид

$$\mathbf{T}=\begin{pmatrix}1&1&1&0\\1&1&0&1\\1&1&0&0\\1&0&1&0\\1&0&0&1\\1&0&0&0\end{pmatrix},\ \mathbf{A}=\begin{pmatrix}1&0&1&0&0&1\\0&1&-1&0&0&0\\0&0&0&1&0&-1\\0&0&0&0&1&-1\end{pmatrix}.$$

Для удобства последующих вычислений изменим порядок столбцов матрицы $\mathbf{A}$ и, следовательно, порядок компонент вектора $\boldsymbol{\theta}$, таким образом, чтобы столбцы матрицы $\mathbf{A}$ с единственными единицами в них располагались на первых r местах.

В рассматриваемом примере это 1, 2, 4 и 5-й столбцы. В матрицах $\mathbf{T}$ и $\mathbf{A}$ обозначим буквенными символами столбцы в соответствии с новым порядком компонент вектора $\boldsymbol{\theta} = (\mu, \alpha_1, \beta_1, \beta_2, \alpha_2, \alpha_3)^{\mathrm{T}}$ и представим матрицу $\mathbf{A}$ в блочной форме. Тогда скелетное разложение примет вид

$$\mathbf{X}=\mathbf{T}\cdot\mathbf{A}=$$

$$=\begin{array}{c}\begin{array}{cccc}\mu&\alpha_1&\beta_1&\beta_2\end{array}\\ \begin{pmatrix}1&1&1&0\\1&1&0&1\\1&1&0&0\\1&0&1&0\\1&0&0&1\\1&0&0&0\end{pmatrix}\end{array}\cdot\begin{array}{c}\begin{array}{cccc|cc}\mu&\alpha_1&\beta_1&\beta_2&\alpha_2&\beta_3\end{array}\\ \left(\begin{array}{cccc|cc}1&0&0&0&1&1\\0&1&0&0&-1&0\\0&0&1&0&0&-1\\0&0&0&1&0&-1\end{array}\right)\end{array}=$$

$$=\mathbf{T}\cdot(\mathbf{I}|\mathbf{K}).$$

II. Полученное представление матрицы $\mathbf{X}$ в виде $\mathbf{X} = \mathbf{T}\cdot\mathbf{A}$ подставим в модель (1):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e} = \mathbf{T}\mathbf{A}\boldsymbol{\theta} + \mathbf{e} = \mathbf{T}\boldsymbol{\psi} + \mathbf{e}, \qquad (8)$$

где $\boldsymbol{\psi} = \mathbf{A}\cdot\boldsymbol{\theta}$ — вектор размерности r.

Покажем, что верхний левый блок матрицы $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}$ можно представить в виде $(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}$, а остальные все элементы этой матрицы равны нулю, т.е.

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}=\begin{bmatrix}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}&0\\0&0\end{bmatrix}.$$

Действительно, используя представление матрицы $\mathbf{A}$ в виде $\mathbf{A} = (\mathbf{I}|\mathbf{K})$, имеем

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}(\mathbf{X}^{\mathrm{T}}\mathbf{X})=$$

$$=\begin{pmatrix}\mathbf{T}^{\mathrm{T}}\mathbf{T}&\mathbf{T}^{\mathrm{T}}\mathbf{TK}\\(\mathbf{TK})^{\mathrm{T}}\mathbf{T}&(\mathbf{TK})^{\mathrm{T}}\mathbf{TK}\end{pmatrix}\times\begin{pmatrix}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}&0\\0&0\end{pmatrix}\times$$

$$\times\begin{pmatrix}\mathbf{T}^{\mathrm{T}}\mathbf{T}&\mathbf{T}^{\mathrm{T}}\mathbf{TK}\\(\mathbf{TK})^{\mathrm{T}}\mathbf{T}&(\mathbf{TK})^{\mathrm{T}}\mathbf{TK}\end{pmatrix}=\begin{pmatrix}\mathbf{T}^{\mathrm{T}}\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}&0\\(\mathbf{TK})^{\mathrm{T}}\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}&0\end{pmatrix}\times$$

$$\times\begin{pmatrix}\mathbf{T}^{\mathrm{T}}\mathbf{T}&\mathbf{T}^{\mathrm{T}}\mathbf{TK}\\(\mathbf{TK})^{\mathrm{T}}\mathbf{T}&(\mathbf{TK})^{\mathrm{T}}\mathbf{TK}\end{pmatrix}=$$

$$=\begin{pmatrix}\mathbf{T}^{\mathrm{T}}\mathbf{T}&\mathbf{T}^{\mathrm{T}}\mathbf{TK}\\(\mathbf{TK})^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{T}&(\mathbf{TK})^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{TK}\end{pmatrix}=$$

$$=\begin{pmatrix}\mathbf{T}^{\mathrm{T}}\mathbf{T}&\mathbf{T}^{\mathrm{T}}\mathbf{TK}\\(\mathbf{TK})^{\mathrm{T}}\mathbf{T}&(\mathbf{TK})^{\mathrm{T}}\mathbf{TK}\end{pmatrix}=\mathbf{X}^{\mathrm{T}}\mathbf{X}.$$

Полученная матрица $\mathbf{A}$ обладает следующими свойствами [4, 5].

1. Она является ненулевой частью матрицы $\mathbf{H}$:

$$\mathbf{H}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{X}=$$

$$=\begin{pmatrix}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}&0\\0&0\end{pmatrix}\times\begin{pmatrix}\mathbf{T}^{\mathrm{T}}\mathbf{T}&\mathbf{T}^{\mathrm{T}}\mathbf{TK}\\(\mathbf{TK})^{\mathrm{T}}\mathbf{T}&(\mathbf{TK})^{\mathrm{T}}\mathbf{TK}\end{pmatrix}=$$

$$=\begin{pmatrix}\mathbf{I}&\mathbf{K}\\0&0\end{pmatrix}=\begin{pmatrix}\mathbf{A}\\0\end{pmatrix},$$

а это означает, что ее строки содержат коэффициенты базисных линейных форм пространства ЛОФ.

2. В случае полного ранга матрицы $\mathbf{X}$ базисные линейные формы представляют собой сравнения, относительно которых традиционно проверяются гипотезы в ДА. Следовательно, если число строк матрицы $\mathbf{A}$ равно полному столбцовому рангу матрицы $\mathbf{X}$ (см. формулу (4)), то нет сомнения в том, что основную гипотезу о значимости факторов можно проверить, записав ее через сравнения.

В рассматриваемом примере компоненты вектора $\boldsymbol{\psi}$ определяются как

$$\psi_1 = \mu + \alpha_2 + \beta_3,\ \psi_2 = \alpha_1 - \alpha_2,$$

$$\psi_3 = \beta_1 - \beta_3,\ \psi_4 = \beta_2 - \beta_3.$$

Модель ДА, представленная в виде (8), позволяет определить МНК-оценки вектора сравнений ψ по обычной формуле Гаусса – Маркова как

$$\hat{\boldsymbol{\psi}} = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{Y}, \qquad (9)$$

при этом $\hat{\psi}_1 = \hat{m} = \bar{\mathbf{Y}}$ — среднее значение вектора $\mathbf{Y}$.

Критерии проверки значимости дисперсионной модели и ее факторов

После вычисления вектора МНК-оценок неизвестных параметров модели (8) следует перейти к процедуре проверки значимости построенной модели. Для этого введем некоторые обозначения [1]: $SSP = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} = |\mathbf{Y}|^2$ — полная сумма квадратов модели, где $|\mathbf{Y}|$ — длина вектора; $\hat{\mathbf{Y}}$ — оценка вектора $\mathbf{Y}$, полученная по формуле $\hat{\mathbf{Y}} = \mathbf{T}\hat{\boldsymbol{\psi}}$. С геометрической точки зрения вектор $\hat{\mathbf{Y}}$ является ортогональной проекцией вектора $\mathbf{Y}$ на пространство линейных оцениваемых форм, т.е. $\hat{\mathbf{Y}} = \mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}\mathbf{Y}$, а $\mathbf{T}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}$ — ортогональный проектор на пространство ЛОФ. Покажем, что $\hat{\mathbf{Y}}^{\mathrm{T}}\hat{\mathbf{Y}} = \hat{\boldsymbol{\psi}}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{T}\hat{\boldsymbol{\psi}}$.

Квадратичная форма в этом выражении численно равна той части суммы SSP, которая обуславливается моделью, обозначим ее через SS, т.е.

$$SS = \hat{\boldsymbol{\psi}}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{T}\hat{\boldsymbol{\psi}} = \hat{\mathbf{Y}}^{\mathrm{T}}\hat{\mathbf{Y}} = |\hat{\mathbf{Y}}|^2. \qquad (10)$$

Квадрат длины вектора остатков $\hat{\mathbf{e}} = (\mathbf{Y} - \hat{\mathbf{Y}})$, ортогонального пространству ЛОФ, обозначим через $SSO = (\mathbf{Y} - \hat{\mathbf{Y}})^{\mathrm{T}}(\mathbf{Y} - \hat{\mathbf{Y}}) = |\mathbf{Y} - \hat{\mathbf{Y}}|^2$. Условия, определяющие векторы $\hat{\mathbf{Y}}$ и $(\mathbf{Y} - \hat{\mathbf{Y}})$, позволяют записать выражение, связывающее рассматриваемые суммы квадратов:

$$SSP = SS + SSO. \qquad (11)$$

Согласно [1] основная гипотеза о значимости факторов модели ДА может быть записана в виде

$$H_0\colon\ \psi = 0. \qquad (12)$$

Статистика, используемая для проверки гипотезы (12), вычисляется как

$$F = \frac{\hat{\bar{\psi}}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})\bar{\psi}/r}{SSO/(n-r)} = \frac{n-r}{r}\,\frac{SS}{SSO}. \qquad (13)$$

Она имеет распределение Снедекора – Фишера с r и $n - r$ степенями свободы. Тогда критерием проверки гипотезы (12) является неравенство

$$F > F_{\alpha;\, r;\, n-r}, \qquad (14)$$

где $F_{\alpha;\, r;\, n-r}$ — α-квантиль распределения Снедекора – Фишера, $\alpha \in (0; 1)$ — выбираемый уровень значимости. Если неравенство (14) верно, то гипотеза H_0 отвергается, в противном случае можно утверждать, что основная гипотеза не отвергается. При отклонении гипотезы (12) модель (8) признается значимой, однако это совсем не означает, что все входные факторы, содержащиеся в модели, оказывают значимое влияние на отклик. Поэтому с практической точки зрения, например, для повышения интерпретационных свойств модели, имеет смысл проверять на значимость не только модель целиком, но и каждый включаемый в нее входной фактор в отдельности.

С этой целью приведенная выше вычислительная схема может быть применена для каждого входного фактора по отдельности. Тогда (13) преобразуется к виду

$$F_k = \frac{\hat{\bar{\psi}}_k^{\mathrm{T}}(\mathbf{T}_k^{\mathrm{T}}\mathbf{T}_k)\hat{\bar{\psi}}_k/q_k}{SSO/(n-r)}, \qquad (15)$$

где $\mathbf{T}_k$ — часть матрицы $\mathbf{T}$, соответствующая k-му фактору, $k = \overline{1, m}$; $q_k = s_k - 1$ — число степеней свободы квадратичной формы числителя.

Аналогично можно проверять на значимость и каждое сравнение каждого фактора. В этом случае в (15) следует положить $q_k \equiv 1$, а в качестве $\mathbf{T}_k$ будем рассматривать один из столбцов матрицы $\mathbf{T}$, соответствующий проверяемому сравнению, обозначим его $\mathbf{t}^{(l)}$, $l = \overline{1, r}$.

Как с вычислительной точки зрения, так и для целей интерпретации получаемых результатов идеальным был бы случай, когда сумма квадратов, обусловленная моделью, могла бы быть представлена в виде

$$SS = SS_1 + SS_2 + \ldots + SS_r, \qquad (16)$$

где SS_l — сумма квадратов, вносимая в SS каждым l-м сравнением, $l = \overline{1, r}$.

Ортогонализация столбцов матрицы T

В критерии (14) используются статистики (13) и (15), определяемые суммами квадратов SS_k, однако для них равенство (16) не имеет место. Это означает, что входные факторы в модели (8) зависимы, причем форма зависимости может быть как линейной, так и нелинейной. От линейной зависимости факторов можно избавиться путем ортогонализации столбцов матрицы $\mathbf{T}$ методом Грама – Шмидта [8, 9].

Результатом ортогонализации матрицы $\mathbf{T}$ будет матрица $\tilde{\mathbf{T}}$. Тогда модель (8) примет вид

$$\mathbf{Y} = \tilde{\mathbf{T}}\boldsymbol{\varphi} + \mathbf{e}, \qquad (17)$$

где векторы параметров моделей (8) и (17) связаны соотношением

$$\boldsymbol{\psi} = \mathbf{P}\boldsymbol{\varphi}, \qquad (18)$$

а $\mathbf{P}$ — матрица перехода от базиса $\boldsymbol{\psi}$ к базису $\boldsymbol{\varphi}$ в пространстве ЛОФ. Основная гипотеза в этом случае будет записана как H_0: $\mathbf{P}\boldsymbol{\varphi} = 0$, что при условии невырожденности матрицы $\mathbf{P}$ равносильно

$$H_0\colon\ \boldsymbol{\varphi} = 0. \qquad (19)$$

В предположении (2) векторы оценок $\hat{\boldsymbol{\psi}}$ и $\hat{\boldsymbol{\varphi}}$ нормально распределены $\hat{\boldsymbol{\psi}} \sim N(\mathbf{T}\psi,\ \sigma^2(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1})$, $\hat{\boldsymbol{\varphi}} \sim N(\tilde{\mathbf{T}}\hat{\boldsymbol{\varphi}}),\ \sigma^2(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}})^{-1})$. Из курса теории вероятностей [10] известно, что ковариационные матрицы оценок $\hat{\boldsymbol{\psi}}$ и $\hat{\boldsymbol{\varphi}}$, обозначаемые $\boldsymbol{\Gamma}_{\hat{\psi}}$ и $\boldsymbol{\Gamma}_{\hat{\varphi}}$, вследствие (18) связаны соотношением $\boldsymbol{\Gamma}_{\hat{\psi}} = \mathbf{P}\boldsymbol{\Gamma}_{\hat{\varphi}}\mathbf{P}^{\mathrm{T}}$ или, что то же самое, $\sigma^2(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1} = \sigma^2\mathbf{P}(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}})^{-1}\mathbf{P}^{\mathrm{T}}$.

Отсюда $((\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1})^{-1} = (\mathbf{P}(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}})^{-1}\mathbf{P}^{\mathrm{T}})^{-1}$ или $\mathbf{T}^{\mathrm{T}}\mathbf{T} = (\mathbf{P}(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}})^{-1}\mathbf{P}^{\mathrm{T}})^{-1}$. Тогда квадратичная форма в (10) может быть представлена следующим образом:

$$\hat{\boldsymbol{\varphi}}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})\hat{\boldsymbol{\varphi}} = (\mathbf{P}\hat{\boldsymbol{\varphi}})^{\mathrm{T}}(\mathbf{P}(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}})^{-1}\mathbf{P}^{\mathrm{T}})^{-1}(\mathbf{P}\hat{\boldsymbol{\varphi}}) =$$

$$= \hat{\boldsymbol{\varphi}}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}(\mathbf{P}^{\mathrm{T}})^{-1}(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}})\mathbf{P}^{-1}\mathbf{P}\hat{\boldsymbol{\varphi}} = \hat{\boldsymbol{\varphi}}^{\mathrm{T}}(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}})\hat{\boldsymbol{\varphi}}.$$

Последнее равенство означает, что квадратичные формы $\hat{\boldsymbol{\varphi}}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})\hat{\boldsymbol{\varphi}}$ и $\hat{\boldsymbol{\varphi}}^{\mathrm{T}}(\widetilde{\mathbf{T}}^{\mathrm{T}}\widetilde{\mathbf{T}})\hat{\boldsymbol{\varphi}}$ численно равны, следовательно, статистики критериев проверки гипотез (12) и (19) совпадают. Тогда очевидно, что (12) и (19) — это одна и та же гипотеза, определяемая разными равенствами. Поэтому все выводы относительно вектора параметров $\boldsymbol{\varphi}$ с соответствующей матрицей $\widetilde{\mathbf{T}}$ могут быть однозначно перенесены относительно вектора параметров $\boldsymbol{\psi}$ с соответствующей матрицей $\mathbf{T}$.

Введение процедуры ортогонализации существенно упрощает вычислительную схему: матрица $\mathbf{T}^{\mathrm{T}}\mathbf{T}$ становится диагональной, отпадает необходимость вычисления обратной к ней матрицы для нахождения оценок параметров или сравнений, поскольку в этом случае достаточно просто заменить ее диагональные элементы обратными к ним величинами. Обозначим диагональ такой обратной матрицы через вектор $\mathbf{d} = (d_1, d_2, ..., d_r)$.

Обращаясь к (8) и (9), заметим, что, вычислив вектор $\mathbf{T}^{\mathrm{T}}\mathbf{Y}$, можно получить оценки вектора сравнений $\hat{\boldsymbol{\psi}}$ как слагаемые скалярного произведения $(\mathbf{d}, \mathbf{T}^{\mathrm{T}}\mathbf{Y})$. Поскольку в условиях ортогональности $\mathbf{T}$ равенство (16) выполняется, то сначала можно вычислять суммы SS_l, $l = \overline{1, r}$, а затем, объединяя их по признаку принадлежности к одному и тому же фактору и суммируя, определить суммы квадратов, обусловленные каждым входным фактором в отдельности.

Это позволяет изменить всю вычислительную схему метода в принципе. Дело в том, что из-за особенностей моделей ДА сами по себе оценки вектора сравнений $\hat{\boldsymbol{\psi}}$ при исследовании моделей не используются, важны лишь значения соответствующих сумм квадратов, которые вносятся в общую сумму каждым отдельным фактором (или соответствующим сравнением).

После проведения ортогонализации матрицы $\mathbf{T}$ ее столбцы $\mathbf{t}^{(l)}$, $l = \overline{1, r}$, образуют ортогональный базис пространства ЛОФ. Геометрически это означает, что все пространство ЛОФ можно представить в виде прямой суммы ортогональных подпространств, а следовательно, вектор прогнозов — в виде $\hat{\mathbf{Y}} = \bigoplus_{l=1}^{r} Z_l$, где $\mathbf{Z}_l$ — проекция вектора $\mathbf{Y}$ на соответствующий вектор ортогонального базиса $\mathbf{t}^{(l)}$. Отсюда следует справедливость равенства (16). Известно, что поочередно вычисляемые проекции можно рассчитать по формулам [1, 8]

$$\mathbf{Z}_l = \frac{(\mathbf{Y}_l, \mathbf{t}^{(l)})}{|\mathbf{Y}_l|^2} \cdot \mathbf{t}^{(l)},\ l = \overline{1, r}, \tag{20}$$

где $\mathbf{Y}_l = \mathbf{Y} - \sum_{i=1}^{l} \mathbf{Z}_i$. При этом несложно заметить, что

$$SS_l = |Z_l|^2, \tag{21}$$

а сама величина $|Z_l|$ есть l-я координата вектора $\hat{\mathbf{Y}}$ в ортогональном базисе, состоящем из столбцов матрицы $\mathbf{T}$. В результате для вычисления сумм SS_l можно применить следующую простую схему.

Шаг 0. Положить $\mathbf{Y}_1 = \mathbf{Y}$, $l = 1$.

Шаг 1. Определить очередную проекцию $\mathbf{Z}_l$ по формуле (20).

Шаг 2. Вычислить сумму квадратов (21).

Шаг 3. Вычислить остаток от проектирования вектора $\mathbf{Y}$ на $\mathbf{t}^{(l)}$:

$$\mathbf{Y}_{l+1} = \mathbf{Y}_l - \mathbf{Z}_l.$$

Шаг 4. Если $l < r$, то положить $l = l + 1$ и перейти на Шаг 1, иначе закончить.

Заметим, что по окончании схемы последнее значение остаточной суммы совпадет с SSO, т.е. $SSO = |\mathbf{Y}_r - \mathbf{Z}_r|^2$.

Как будет показано ниже, результаты исследования моделей ДА при помощи разработанной вычислительной схемы неоднозначны: в моделях неполного факторного эксперимента (НФЭ) на значимость уровней факторов могут оказать влияние порядок следования факторов и их уровней, вследствие чего изменяется базис пространства ЛОФ. В целях снижения уровня неопределенности выводов относительно моделей ДА авторы предлагают использовать регуляризационный подход.

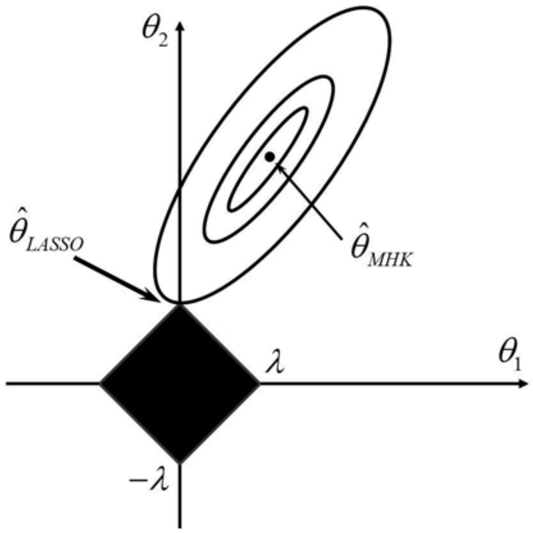
Метод LASSO для выявления значимых факторов в моделях ДА

Изначально метод LASSO был разработан Р. Тибширани [11] в целях решения проблемы плохой обусловленности информационной матрицы, однако позже обнаружились дополнительные полезные свойства метода, позволившего существенно расширить его возможности. В начале XXI века в зарубежной литературе появилось множество публикаций, посвященных разработке разнообразных версий метода LASSO, которые не прекращаются и в настоящее время [12]. В отечественной литературе метод LASSO упоминается выборочно и, как правило, в методической литературе на практике активно не используется.

Кратко изложим идеи метода LASSO. Пусть за основу взята линейная по параметрам регрессионная модель

$$\mathbf{Y} = \theta_0 + \theta_1 x_1 + ... + \theta_m x_m + \mathbf{e}, \tag{22}$$

где $x_1, x_2, ..., x_m$ — заданные входные переменные, $\theta_0, \theta_1, ..., \theta_m$ — неизвестные параметры мо-



Графическая интерпретация метода LASSO

дели, $\mathbf{e}$ — аддитивная случайная составляющая. Для устранения эффекта масштаба обычно рекомендуется проводить центрирование и нормирование входных переменных. Тогда модель (22) будет лишена свободного члена и, с точностью до обозначений, может быть переписана как

$$\mathbf{Y} = \theta_1 x_1 + \ldots + \theta_m x_m + \mathbf{e} = \mathbf{x}\boldsymbol{\theta} + \mathbf{e}. \qquad (23)$$

В рамках решаемой задачи осуществлять указанное преобразование нет необходимости, поскольку заранее известно, что в моделях ДА все элементы матрицы наблюдений $\mathbf{X}$ могут принимать только значения 0 или 1, т.е. масштабирование не требуется.

При оценивании параметров модели (23) при плохой обусловленности матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ как значения получаемых оценок, так и их дисперсия могут быть достаточно большими. Это негативно сказывается на результатах проводимого статистического анализа, поскольку появляются существенные ошибки в прогнозировании значений отклика [11, 12]. Для контроля таких ошибок предложено решать задачу условной оптимизации

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}) \to \min_{\theta}, \; Pe(\boldsymbol{\theta}) \le \lambda, \qquad (24)$$

где $Pe(\boldsymbol{\theta})$ — функция штрафа (penalty function); λ — некоторый параметр регуляризации. Если в задаче (24) штрафная функция есть сумма квадратов оценок параметров модели (23), то получится известная задача оценивания ридж-регрессии, а если $Pe(\boldsymbol{\theta}) = |\theta_1| + \ldots + |\theta_m|$, то возникает метод LASSO.

Несмотря на внешнюю схожесть ридж- и LASSO-регрессий, получаемые на их основе результаты будут иметь существенные различия. Полезно следующее свойство оценок LASSO: если при некотором значении параметра регуляризации $\lambda = \lambda_0$ какой-либо из параметров регрессии обращается в нуль, то для всех значений $\lambda < \lambda_0$ этот параметр также остается равным нулю, что несвойственно ридж-регрессии [11, 12]. А поскольку нулевое значение параметра автоматически означает незначимость соответствующего ему фактора, то, уменьшая величину параметра регуляризации, можно наблюдать постепенное исключение из исходной модели малозначимых входных факторов и сохранения в ней тех, которые даже при наличии внешних ограничений принимают ненулевые значения. Тогда, решая последовательно задачи типа (24) с разными значениями λ, можно проранжировать входные факторы, не прибегая к стандартным статистическим критериям типа Стьюдента, Фишера или Пирсона [13].

Идею метода LASSO удобно проинтерпретировать графически на примере модели с двумя параметрами. На рисунке изображены линии уровня функции, представляющей собой сумму квадратов отклонений исходных и прогнозных значений отклика в зависимости от параметров модели. Стрелками указаны оценки параметров по МНК и методу LASSO, черным выделена область допустимых решений при $Pe(\boldsymbol{\theta}) \le \lambda$ — ромб с диагоналями 2λ. Видно, что оптимальное решение задачи (24) лежит на границе ромба и при уменьшении его площади (параметра регуляризации) оно будет по-прежнему находиться на оси θ_2, т.е. соответствовать ситуации, когда $\theta_1 = 0$.

Результаты вычислительных экспериментов

Корректность разработанного в работе подхода и изучения построения моделей ДА проверена при помощи метода статистического моделирования. Процесс вычислений был условно разбит на три этапа.

Этап 1. Для проверки того, что в ПФЭ порядок следования факторов (и их уровней) не играет роли, рассмотрен эксперимент с двумя факторами — с двумя и тремя уровнями соответственно. Для обеспечения полноты ранга матрицы $\mathbf{X}$ из нее были удалены столбцы, соответствующие второму уровню первого фактора и третьему уровню второго. Значения отклика были зашумлены случайной ошибкой с нормальным распределением, $e \sim N(0{,}2)$. Отметим, что для обеспечения достаточного числа степеней свободы вектор откликов $\mathbf{Y}$ представляет собой результат усреднения двух наблюдений, проведенных в каждой точке плана.

Результаты вычислений по предложенной схеме представлены в табл. 1. Жирным шрифтом выделены значения тех статистик, которые оказались выше критических значений распределения Фишера, следовательно, соответствующие им уровни факторов являются значимыми. Полная и остаточная суммы квадратов в обоих случаях были одинаковыми. Полученные результаты подтверждают тот факт, что ПФЭ инвариантен порядку извлечения факторов.

Этап 2. Возможности предложенного в работе подхода были продемонстрированы на примере анализа модели ДА в условиях НФЭ. Для этого из ПФЭ было удалено первое наблюдение (без потери общности можно было выбрать любое другое) и рассмотрены различные случаи последовательного извлечения факторов и их уровней, т.е. процесс проектирования вектора $\mathbf{Y}$ на столбцы ортогональной матрицы $\widetilde{\mathbf{T}}$. Рассмотрены шесть схем. Схема 1 — с прямым порядком извлечения факторов (сначала уровни первого фактора, потом — уровни второго); схема 2 — с обратным порядком извлечения факторов (сначала — уровни второго фактора, потом — первого); схема 3 — с прямым порядком извлечения факторов с условием, что для обеспечения полноты ранга матрицы наблюдений вместо уровня β_3 удален уровень β_1; схема 4 — то же самое, что и схема 3, но удален уровень β_2; схемы 5 и 6 аналогичны схемам 3 и 4, но при обратном порядке извлечения факторов. Результаты представлены в табл. 2 – 4. Здесь жирным шрифтом выделены значения статистики Фишера, оказавшиеся значимыми, т.е. соответствующие значимым уровням факторов.

Анализируя полученные результаты, можно заключить, что изменение схемы проведения эксперимента позволяет выделять уровни входных факторов с различной значимостью. Это означает, что существует возможность планирования экспериментов для получения факторов (или их уровней) с наибольшей значимостью. В рассмотренном примере такую значимость обеспечивают эксперименты, проведенные по схемам 1 или 4.

Этап 3. Для проверки корректности выделения значимых факторов в моделях ДА был применен метод LASSO. В качестве базовой рассматривалась модель ДА без генерального среднего, поскольку в методе LASSO необходимо, чтобы средние значения переменных были нулевыми [11]. В используемой схеме НФЭ считается, что из схемы ПФЭ удалено первое наблюдение. Для обеспечения полноты ранга матрицы наблюдений был удален второй уровень первого фактора, поэтому в модели присутствуют параметры $\alpha_1 - \alpha_2$, β_1, β_2, β_3. В табл. 5 приведены результаты применения метода LASSO для эксперимента, проведенного по схеме 1 из этапа 2. Жирным шрифтом выделены ненулевые значения оценок параметров: чем их больше в столбце, тем более значимое влияние на отклик оказывает соответствующий уровень фактора или сравнение. Видно, что метод LASSO предлагает следующий способ ранжирования уровней второго фактора от самого сильного (по степени влияния на отклик) до самого слабого: $\beta_1 \succ \beta_2 \succ \beta_3$. Как видно из табл. 2, рассмотренная схема 1 дает значимость сравнения $\psi_3 = \beta_1 - \beta_3$ (на основе сравнения зна-

Таблица 1. Результаты моделирования в случае ПФЭ

Шаги схемы, l	Объясненная сумма, SS_l	Значение статистики, F_l	Остаточная сумма, SSY_l
	Прямой порядок факторов		
1	206,930	532,911	54,923
2	1,252	3,225	53,671
3	47,360	121,967	6,311
4	3,981	10,253	2,330
	Обратный порядок факторов		
1	206,930	532,911	54,923
2	47,360	121,967	7,563
3	3,981	10,253	3,582
4	1,252	3,225	2,330

Таблица 2. Схемы с прямым и обратным порядком извлечения факторов

l	Схема 1			Схема 2		
	SS_l	F_l	SSY_l	SS_l	F_l	SSY_l
1	51,200	136,533	8,800	51,200	136,533	8,800
2	0,300	0,800	8,500	4,050	**10,800**	4,750
3	6,000	**16,000**	2,500	0,250	0,667	4,500
4	0,250	0,667	2,250	2,250	**6,000**	2,250

Таблица 3. Схемы с прямым порядком извлечения факторов при отсутствии уровней второго фактора β_1 и β_2

l	Схема 3			Схема 4		
	SS_l	F_l	SSY_l	SS_l	F_l	SSY_l
1	51,200	136,533	8,800	51,200	136,533	8,800
2	0,300	0,800	8,500	0,300	0,800	8,500
3	1,929	5,143	6,571	6,000	**16,000**	2,500
4	4,321	**11,524**	2,250	0,250	0,667	2,250

Таблица 4. Схемы с обратным порядком извлечения факторов при отсутствии уровней второго фактора β_1 и β_2

l	Схема 5			Схема 6		
	SS_l	F_l	SSY_l	SS_l	F_l	SSY_l
1	51,200	136,533	8,800	51,200	136,533	8,800
2	1,633	4,356	7,167	4,050	**10,800**	4,750
3	2,667	**7,111**	4,500	0,250	0,667	4,500
4	2,250	**6,000**	2,250	2,250	**6,000**	2,250

Таблица 5. Оценки параметров метода LASSO

Параметр регуляризации λ	Оценки параметров			
	$\alpha_1 - \alpha_2$	β_1	β_2	β_3
3,300	**0,300**	**1,800**	**1,000**	**–0,200**
3,135	**0,281**	**1,763**	**0,926**	**–0,163**
2,970	**0,263**	**1,727**	**0,853**	**–0,127**
2,805	**0,245**	**1,690**	**0,780**	**–0,090**
2,640	**0,227**	**1,653**	**0,707**	**–0,053**
2,475	**0,208**	**1,617**	**0,633**	**–0,017**
2,310	**0,187**	**1,574**	**0,548**	0,000
2,145	**0,163**	**1,527**	**0,454**	0,000
1,980	**0,140**	**1,480**	**0,360**	0,000
1,815	**0,116**	**1,433**	**0,266**	0,000
1,650	**0,093**	**1,386**	**0,171**	0,000
1,485	0,069	**1,339**	**0,077**	0,000
1,320	0,000	**1,307**	**0,013**	0,000
1,155	0,000	**1,154**	**0,001**	0,000
0,990	0,000	**0,990**	0,000	0,000
0,825	0,000	**0,825**	0,000	0,000
0,660	0,000	**0,660**	0,000	0,000
0,495	0,000	**0,495**	0,000	0,000
0,330	0,000	**0,330**	0,000	0,000
0,165	0,000	**0,165**	0,000	0,000
0,000	0,000	**0,000**	0,000	0,000

чений F-статистик). Сопоставляя этот вывод с результатом применения метода LASSO, заключаем, что из двух уровней второго фактора (β_1 и β_3) бо́льшую значимость имеет уровень β_1.

Таким образом, проведенное моделирование позволило определить, что в схемах неполного факторного эксперимента порядок извлечения факторов и их уровней оказывает значимое влияние на получаемые результаты, т.е. приводит к изменению значимости этих уровней и, возможно, самих факторов.

Предложенная вычислительная схема анализа моделей ДА свободна от недостатков схем, в которых используют другие способы получения обобщенных обратных матриц, а применение процедуры ортогонализации позволяет устранить линейную зависимость входных факторов. В целях планирования экспериментов разработанный подход позволяет определять такую схему проведения эксперимента, при которой обеспечивается выделение наиболее значимых факторов.

ЛИТЕРАТУРА

1. **Шеффе Г.** Дисперсионный анализ. — М.: Физматлит, 1980. — 512 с.
2. **Маркова Е. В., Грошовый Т. А., Головкин В. А.** Математическое моделирование эксперимента в фармацевтической технологии: планы дисперсионного анализа. — Киев: Высшая школа, 1992. — 187 с.
3. **Холлендер М., Вулф Д. А.** Непараметрические методы статистики. — М.: Финансы и статистика, 1983. — 518 с.
4. **Бекарева Н. Д.** Анализ и планирование ковариационного эксперимента / Автореферат дис. ... канд. техн. наук. — Новосибирск, 1981. — 24 с.
5. **Бекарева Н. Д., Денисов В. И., Попов А. А.** Оптимальное планирование дисперсионного и ковариационного экспериментов / Заводская лаборатория. 1986. Т. 52. № 5. С. 60 – 63.
6. **Рао С. Р.** Линейные статистические методы и их применения. — М.: Наука, 1968. — 548 с.
7. **Сирл С., Госман У.** Матричная алгебра в экономике. — М.: Статистика, 1974. — 375 с.
8. **Гантмахер Ф.** Теория матриц. — М.: Физматлит, 2004. — 560 с.
9. **Бекарева Н. Д., Раскосова М. А.** Статистическая обработка моделей дисперсионного анализа / Сборник научных трудов Новосибирского государственного технического университета. 2012. № 4. С. 73 – 78.
10. **Гнеденко Б. В.** Курс теории вероятностей. — М.: Едиториал УРСС, 2005. — 448 с.
11. **Tibshirani R.** Regression shrinkage and selection via the LASSO / Journal of the Royal Statistical Society. 1996. N 58(1). P. 267 – 288.
12. **Hastie T., Tibshirani R., Friedman J.** The elements of statistical learning: data mining, inference, and prediction. — Springer-Verlag, 2009. — 746 p.
13. **Мельников В. В., Усманов Т. Б., Щеколдин В. Ю.** Факторы государственных закупок: оценка влияния в рамках международных сопоставлений / Вопросы регулирования экономики. 2016. Т. 7. № 2. С. 82 – 93.

REFERENCES

1. **Sheffe G.** Analysis of variance. — Moscow: Fizmatlit, 1980. — 512 p. [in Russian].
2. **Markova E., Groshevyi T., Golovkin V.** Mathematical modeling of the experiment in pharmaceutical technology: designs analysis of variance. — Kiev: Vysshaya shkola, 1992. — 187 p. [in Russian].
3. **Hollender M., Wolffe D.** Nonparametical methods of statistics. — Moscow: Finansy i statistika, 1983. — 518 p. [in Russian].
4. **Bekareva N.** Analysis and planning for covariance experiment: author’s abstract of candidate’s thesis. — Novosibirsk, 1981. — 24 p. [in Russian].
5. **Bekareva N., Denisov V., Popov A.** Optimal planning for variance and covariance experimental design / Zavod. Lab. 1986. Vol. 52. N 5. P. 60 – 63 [in Russian].
6. **Rao S.** Linear statistical inference and its applications. — Moscow: Nauka, 1968. — 548 p. [in Russian].
7. **Searle S., Hausman W.** Matrix algebra for business and economics. — Moscow: Statistika, 1974. — 375 p. [in Russian].
8. **Gantmakher F.** Matrix theory. — Moscow: Fizmatlit, 2004. — 560 p. [in Russian].
9. **Bekareva N., Raskosova M.** Statistical processing for models of analysis of variance / Sb. Nauch. Tr. Novosib. Gos. Tekhn. Univ. 2012. N 4. P. 73 – 78 [in Russian].
10. **Gnedenko B.** Course in probability theory. — Moscow: Editorial URSS, 2005. — 448 p. [in Russian].
11. **Tibshirani R.** Regression shrinkage and selection via the LASSO / Journal of the Royal Statistical Society. 1996. N 58(1). P. 267 – 288.
12. **Hastie T., Tibshirani R., Friedman J.** The elements of statistical learning: data mining, inference, and prediction. — Springer-Verlag, 2009. — 746 p.
13. **Melnikov V., Usmanov T., Shchekoldin V.** Factors of public procurement: evaluation of an influence into the international comparison / Vopr. Regulir. Ékon. 2016. Vol. 7. N 2. P. 82 – 93 [in Russian].