КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Кафедра математической статистики

ДИСПЕРСИОННЫЙ АНАЛИЗ

Методические разработки
по специальному курсу

Ч а с т ь I

Оценки метода наименьших квадратов.
Критерий Фишера. Однофакторный
дисперсионный анализ.

Казань – 1998

П Е Ч А Т А Е Т С Я

ПО РЕШЕНИЮ СЕКЦИИ НАУЧНО-МЕТОДИЧЕСКОГО СОВЕТА

КАЗАНСКОГО УНИВЕРСИТЕТА

С о с т а в и т е л ь

доцент С.В.Симушкин

Дисперсионный анализ ($Д_A$) представляет собой набор статистических методов, предназначенных для обработки экспериментальных данных, зависящих от большого количества одновременно действующих факторов. Целью анализа является оценка влияния факторов на результирующий признак и выявление наиболее значимых из них. В качестве примера применения методов $Д_A$ можно привести классическую задачу сравнения урожайности нескольких сортов пшеницы, высеянных на участках земли с различным химическим составом, с применением различных типов удобрений. Методы $Д_A$ позволяют ответить на вопросы о зависимости урожайности от каждого из трех рассматриваемых факторов (сорт, земля, удобрение), о возможном взаимодействии факторов и на ряд других вопросов.

Существенным моментом теории $Д_A$ является представление регрессии отклика статистического эксперимента в виде линейной комбинации функций факторов с неизвестными коэффициентами. В первых двух главах данного курса рассматривается оценка по методу наименьших квадратов неизвестных коэффициентов и выводится ее распределение в предположении нормальности распределения ошибки. В последующих главах строится критерий отношения правдоподобий для проверки различных гипотез относительно уравнения регрессии. Рассматриваются задачи одно-, двух- и многофакторного $Д_A$. Наконец, в последних главах предлагаются методы построения планов статистического эксперимента, оптимизирующих некоторые характеристики дисперсии оценок коэффициентов регрессии.

Оглавление

Г л а в а I

Оценки влияния факторов

§ 1. Классическая модель линейной регрессии

Классическая модель регрессии предполагает, что учитываемые в эксперименте факторы $x_1, \ldots, x_p$ оказывают линейное воздействие на отклик – результат эксперимента y :

$$y = x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p + \varepsilon,$$

где $\beta_1, \ldots, \beta_p$ – неизвестные параметры, в той или иной мере отражающие степень воздействия факторов, ε– случайная ошибка эксперимента, $x_1, \ldots, x_p$ – вектор известных коэффициентов, задаваемых обычно исследователем. Чаще всего эти коэффициенты равны 1 или 0 в зависимости от того, присутствует соответствующий фактор в эксперименте или нет. Иногда, однако, они могут принимать и непрерывный ряд значений, выражая степень предпочтения того или иного фактора. Целью *ДА* является получение выводов относительно значений неизвестных параметров и их комбинаций, а также относительно распределения ошибки ε.

Если было произведено $n\,(> p)$ экспериментов, то результаты измерений могут быть представлены в матричном виде

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

где $\overrightarrow{y} = (y_1, \ldots, y_n)'$ – вектор-столбец наблюденных значений, $\overrightarrow{\beta} = (\beta_1, \ldots, \beta_p)'$ – столбец неизвестных параметров, $\overrightarrow{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ – вектор-столбец ненаблюдаемых ошибок,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ .. & .. & .. & .. \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{pmatrix} -$$

матрица коэффициентов, задаваемых экспериментатором.

Определение 1. Матрица $\mathbf{X}$ называется *матрицей плана.*

Различают три типа моделей *ДА*.

Модель I – модель с постоянными факторами, в которой все параметры $\{\beta_j\}$ постоянны. Иными словами, для этой модели факторы могут разве лишь присутствовать или отсутствовать в эксперименте, но их влияние всюду, где они есть, одинаково.

Если некоторый фактор β_j постоянный и входные коэффициенты $x_{ji} = 1$ для всех $i = 1, \ldots, n$ (то есть фактор присутствует во всех экспериментах), тогда β_j называется *аддитивным фактором.* Легко понять, что такой фактор представляет собой средний ожидаемый отклик в предположении, что все остальные неаддитивные факторы не влияют на результат эксперимента. Поэтому аддитивный фактор может быть только один и его можно интерпретировать как среднее значение случайной составляющей модели ε. Этим, в частности, объясняется введенное ниже условие $\mathbf{E}\,\varepsilon = 0$ на ошибку эксперимента.

Модель II – модель со случайными факторами, где все параметры $\{\beta_j\}$ случайны, кроме, может быть, одного, который является аддитивным.

Модель III – смешанная модель.

Рассмотрим сначала ситуацию, описываемую моделью I *ДА*.

Статистические выводы в значительной степени зависят от предположений относительно распределения ошибок $\overrightarrow{\varepsilon}$. Минимальные требования, которым должно подчиняться это распределение, заключаются, во-первых, в отсутствии систематической ошибки –

$$\mathbf{E}\,\varepsilon_i = 0,\ i = 1, \ldots, n,$$

во-вторых, в некоррелируемости ошибок в различных экспериментах –

$$\mathbf{E}\,\varepsilon_j \varepsilon_k = 0 \text{ при } j \neq k,$$

и, в-третьих, в одинаковой распределенности ошибок. Последнее условие обычно сводится к требованию отсутствия изменчивости дисперсии ошибок от эксперимента к эксперименту: $\mathbf{D}\,\varepsilon_i = \mathbf{E}\,\varepsilon_i^2 = \sigma^2,\ i = 1, \ldots, n$.

Таким образом, основные предположения *ДА* можно записать в виде

$$\Omega: \quad \begin{cases} \overrightarrow{y} = \mathbf{X}'\overrightarrow{\beta} + \overrightarrow{\varepsilon}, \\ \mathbf{E}\,\overrightarrow{\varepsilon} = \overrightarrow{0}, \quad \mathrm{Cov}(\overrightarrow{\varepsilon}) = \mathbf{E}\,\overrightarrow{\varepsilon}\,\overrightarrow{\varepsilon}' = \sigma^2\mathbf{I}. \end{cases} \tag{I.1}$$

Замечание i. Из основных предположений следует, что

$$\mathbf{E}\,\overrightarrow{y} = \mathbf{X}'\beta \quad \text{и} \quad \mathrm{Cov}(\overrightarrow{y}) = \sigma^2\mathbf{I}. \tag{I.2}$$

§ 2. Оценки метода наименьших квадратов

Мера качества оценок $\overrightarrow{b}$ факторных параметров $\overrightarrow{\beta}$ должна, естественно, зависеть от расхождений $e_i = y_i - (x_{1i}b_1 + \ldots + x_{pi}b_p), \quad i = 1, \ldots, n,$ между наблюденными значениями отклика $\overrightarrow{y}$ и ожидаемыми значениями $\mathbf{X}'\overrightarrow{b}$. На практике наиболее популярна мера качества, равная сумме квадратов ошибок –

$$\mathcal{G}(\overrightarrow{b}, \overrightarrow{y}) \;=\; \sum_{i=1}^{n} e_i^2 = \|\overrightarrow{y} - \mathbf{X}'\overrightarrow{b}\|^2,$$

и называемая обычно средне-квадратической ошибкой (СКО).

Определение 2. Вектор $\overrightarrow{\beta^*}$, для которого СКО

$$\|\overrightarrow{y} - \mathbf{X}'\overrightarrow{\beta^*}\|^2 = \min_{\overrightarrow{b} \in \mathcal{R}^p} \|\overrightarrow{y} - \mathbf{X}'\overrightarrow{b}\|^2,$$

называется *оценкой по методу наименьших квадратов* (ОМНК). Величина СКО

$$SS_e \;= \|\overrightarrow{y} - \mathbf{X}'\overrightarrow{\beta^*}\|^2 \tag{I.3}$$

называется *суммой квадратов ошибок*. Иногда, для того, чтобы подчеркнуть, что минимум ошибки найден при основных предположениях, мы будем обозначать SS_e через $\mathcal{G}_\Omega$.

Покажем, что ОМНК всегда существует.

Если ОМНК существует, то, как известно из курса математического анализа, все производные СКО по компонентам вектора $\overrightarrow{b}$ будут равны нулю в точке $\overrightarrow{b} = \overrightarrow{\beta^*}$. Произведя элементарные алгебраические выкладки, легко показать, что если ОМНК $\overrightarrow{\beta^*}$ существует, то она обязана удовлетворять системе линейных уравнений

$$\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ki}x_{ji}\overrightarrow{\beta_i^*} \;=\; \sum_{i=1}^{n} x_{ki}y_i, \quad k = \overline{1,p}, \tag{I.4}$$

или в матричной форме

$$\mathcal{S}\,\overrightarrow{\beta^*} = \mathbf{X}\,\overrightarrow{y}, \tag{I.5}$$

где матрица $\mathcal{S} = \mathbf{X}\,\mathbf{X}'$ – так называемая *информационная матрица* плана.

Определение 3. Уравнения (I.4) и (I.5) называются *нормальными уравнениями.*

Дальнейшие рассуждения будут существенно опираться на следующую геометрическую интерпретацию.

Обозначим через $\overrightarrow{\xi_j}$ j-ый столбец матрицы $\mathbf{X}'$ (– столбец коэффициентов при j-ом факторе). Если $\operatorname{rang}\mathbf{X} = r\,(\leqslant p)$, то линейное пространство, порожденное системой векторов $\{\overrightarrow{\xi_j}\}_{j=1}^{p}$ (– то есть множество линейных комбинаций $c_1\overrightarrow{\xi_1} + \ldots + c_p\overrightarrow{\xi_p}$), будет иметь размерность r. Обозначим это пространство V_r. Тогда, очевидно, $\mathbf{X}'\overrightarrow{b} = b_1\overrightarrow{\xi_1} + \ldots + b_p\overrightarrow{\xi_p} \in V_r$ и $\| \overrightarrow{y} - \mathbf{X}'\overrightarrow{b} \|$ равно расстоянию от вектора $\overrightarrow{y}$ до элемента $\mathbf{X}'\overrightarrow{b}$ пространства V_r.

Мы часто будем пользоваться критерием перпендикулярности векторов к пространству V_r.

Лемма I.1. *Вектор $\overrightarrow{a} \perp V_r$ тогда и только тогда, когда*

$$\overrightarrow{a}'\mathbf{X}' = \overrightarrow{0}' \quad \textit{или} \quad \mathbf{X}\,\overrightarrow{a} = \overrightarrow{0}.$$

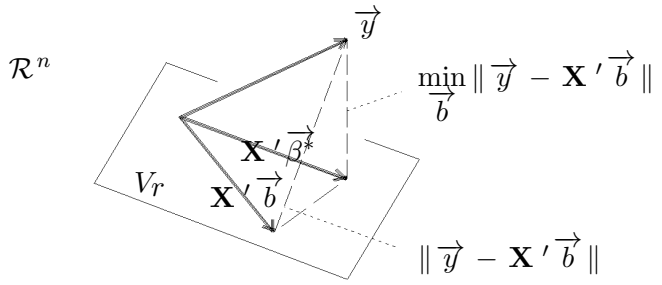
Доказательство. Вектор $\overrightarrow{a} \perp V_r$ только в том случае, если он перпендикулярен всем векторам $\overrightarrow{\xi_j}$, порождающим V_r : $\overrightarrow{a}'\overrightarrow{\xi_j} = 0,\ j = \overline{1,p}$. Поскольку $\overrightarrow{\xi_j}$ образуют столбцы матрицы $\mathbf{X}'$, то отсюда получаем, что перпендикулярность $\overrightarrow{a} \perp V_r$ эквивалентна соотношению

$$\overrightarrow{a}'\mathbf{X}' = \overrightarrow{a}'(\overrightarrow{\xi_1}, \ldots, \overrightarrow{\xi_p}) = (\overrightarrow{a}'\overrightarrow{\xi_1}, \ldots, \overrightarrow{a}'\overrightarrow{\xi_p}) = \overrightarrow{0}'.$$

Второе равенство получается из первого простым транспонированием. $\otimes$

Теорема I.1. *ОМНК всегда существует и удовлетворяет нормальным уравнениям. Любое решение нормальных уравнений будет ОМНК.*

Доказательство. Вектор $\overrightarrow{y}$ принадлежит пространству $\mathcal{R}^n \supset V_r$. Как известно, проекция $\operatorname{Proj}_{V_r}\overrightarrow{y}$ вектора $\overrightarrow{y}$ на пространство V_r является единственным элементом из V_r, минимизирующим расстояние до вектора $\overrightarrow{y}$. Поскольку $\operatorname{Proj}_{V_r}\overrightarrow{y} \in V_r$, то существует такой вектор $\overrightarrow{\beta^*}$, что $\operatorname{Proj}_{V_r}\overrightarrow{y} = \mathbf{X}'\overrightarrow{\beta^*}$ и, следовательно, $\overrightarrow{\beta^*}$ есть ОМНК (см.рис.).



Пусть теперь $\overrightarrow{\beta^*}$ – некоторая ОМНК. Тогда $\mathbf{X}'\overrightarrow{\beta^*} = \mathrm{Proj}_{V_r} \overrightarrow{y}$, т.к. проекция единственна. Следовательно,

$$\mathbf{X}'\overrightarrow{\beta^*} - \overrightarrow{y} \perp V_r .$$

В силу леммы I.1 это возможно только тогда, когда

$$\mathbf{X}\,(\mathbf{X}'\overrightarrow{\beta^*} - \overrightarrow{y}) = \overrightarrow{0},$$

что эквивалентно системе нормальных уравнений (I.5). $\otimes$

Замечание ii. Несмотря на то, что проекция $\mathrm{Proj}_{V_r} \overrightarrow{y}$ определяется единственным образом, ОМНК может быть не единственной. Следующее утверждение содержит условие единственности ОМНК и, кроме того, здесь вычисляется ее ковариационная матрица.

Следствие 1. *Если* $\mathrm{rang}\,\mathbf{X} = p$, *то*
а) существует единственная ОМНК $\overrightarrow{\beta^*} = \mathcal{S}^{-1}\,\mathbf{X}\,\overrightarrow{y}$;
б) $\overrightarrow{\beta^*}$ *есть линейная несмещенная оценка вектора* $\overrightarrow{\beta}$;
в) ковариационная матрица $Cov(\overrightarrow{\beta^*}) = \sigma^2\,\mathcal{S}^{-1}$.

Доказательство. а) Из теории матриц известно, что $\mathrm{rang}\,\mathcal{S} = \mathrm{rang}\,\mathbf{X}\,\mathbf{X}' = \mathrm{rang}\,\mathbf{X} = p$. Следовательно, матрица $\mathcal{S}$ невырождена и имеет обратную матрицу. Первый пункт следствия вытекает теперь из вида нормальных уравнений (I.5).

б) Так как среднее значение $\mathbf{E}\,\overrightarrow{y} = \mathbf{X}'\overrightarrow{\beta}$, то математическое ожидание вектора $\overrightarrow{\beta^*}$ равно

$$\mathbf{E}\,\overrightarrow{\beta^*} = \mathbf{E}\,\mathcal{S}^{-1}\,\mathbf{X}\,\overrightarrow{y} = \mathcal{S}^{-1}\,\mathbf{X}\;\mathbf{E}\,\overrightarrow{y} = \mathcal{S}^{-1}\,\mathbf{X}\,\mathbf{X}'\overrightarrow{\beta} = \mathcal{S}^{-1}\,\mathcal{S}\,\overrightarrow{\beta} = \overrightarrow{\beta}.$$

в) ОМНК $\overrightarrow{\beta^*} = \mathcal{S}^{-1}\,\mathbf{X}\,\overrightarrow{y}$ есть линейная комбинация вектора $\overrightarrow{y}$ с ковариационной матрицей $\mathrm{Cov}(\overrightarrow{y}) = \sigma^2\mathbf{I}$, поэтому

$$\mathrm{Cov}(\overrightarrow{\beta^*}) = \mathcal{S}^{-1}\,\mathbf{X}\;\mathrm{Cov}(\overrightarrow{y})\,(\mathcal{S}^{-1}\,\mathbf{X})' = \sigma^2\,\mathcal{S}^{-1}\,\mathbf{X}\,\mathbf{X}'\,(\mathcal{S}^{-1})'.$$

Очевидно матрица $\mathcal{S}$ симметрична и $\mathbf{X}\,\mathbf{X}' = \mathcal{S}$, следовательно,

$$\mathrm{cov}(\overrightarrow{\beta^*}) = \sigma^2\,\mathcal{S}^{-1}\,\mathcal{S}\,\mathcal{S}^{-1} = \sigma^2\,\mathcal{S}^{-1}.$$

что и требовалось доказать. $\otimes$

§ 3. Функции, допускающие оценку. Теорема Гаусса-Маркова

В реальных статистических задачах зачастую требуется определить не сами значения параметров $\overrightarrow{\beta}$, входящих в уравнение регрессии, а либо оценить некоторую функцию от этих параметров, либо проверить ряд гипотез относительно $\overrightarrow{\beta}$. Как та, так и другая проблема формализуется обычно с помощью линейных функций вектора $\overrightarrow{\beta}$. Здесь будет описан класс линейных функций, для которых (в любом случае) существует единственная ОМНК и которые применяются для целей *ДА*.

Определение 4. 1) Линейная функция от $\overrightarrow{\beta}$ с известными коэффициентами

$$\psi(\overrightarrow{\beta}) = \sum_{j=1}^{p} c_j\,\beta_j = \overrightarrow{c}'\,\overrightarrow{\beta}$$

называется *параметрической функцией.*

2) Говорят, что параметрическая функция $\psi(\overrightarrow{\beta})$ *допускает оценку*, если существует такой вектор $\overrightarrow{a} = (a_1, \ldots, a_p)'$, что

$$\mathbf{E}\,\overrightarrow{a}'\overrightarrow{y} = \psi(\overrightarrow{\beta}). \tag{I.6}$$

Замечание iii. Условие I.6 означает, по-существу, что для параметрической функции ψ существует линейная несмещенная оценка $\widehat{\psi} = \overrightarrow{a}'\,\overrightarrow{y}$.

Теорема I.2. *Функция $\psi = \overrightarrow{c}'\,\overrightarrow{\beta}$ допускает оценку тогда и только тогда, когда вектор-строка $\overrightarrow{c}'$ является линейной комбинацией строк матрицы $\mathbf{X}'$, т.е. когда существует вектор $\overrightarrow{a}$ такой , что $\overrightarrow{c}' = \overrightarrow{a}'\mathbf{X}'$.*

Доказательство. Функция ψ допускает оценку только тогда, когда существует такой вектор $\overrightarrow{a}$, для которого при любом $\overrightarrow{\beta}$

$$\overrightarrow{c}'\,\overrightarrow{\beta} = \mathbf{E}\,\overrightarrow{a}'\,\overrightarrow{y} = \overrightarrow{a}'\mathbf{X}'\,\overrightarrow{\beta},$$

что равносильно тождеству $\overrightarrow{c}' = \overrightarrow{a}'\mathbf{X}'$. $\otimes$

Найдем теперь наилучшую из возможных оценок параметрической функции. Сравнение оценок будем производить по их дисперсии.

Определение 5. Оценку $\widehat{\psi} = \overrightarrow{a}'\overrightarrow{y}$ назовем МНК-*оценкой* для параметрической функции ψ, если она имеет наименьшую дисперсию среди всех линейных несмещенных оценок ψ.

Теорема I.3. *(Гаусса-Маркова) Для любой параметрической функции $\psi = \overrightarrow{c}'\overrightarrow{\beta}$, допускающей оценку, существует единственная МНК-оценка. Значение этой оценки может быть вычислено по формуле $\widehat{\psi} = \overrightarrow{c}'\overrightarrow{\beta^*}$, где $\overrightarrow{\beta^*}$ – любая ОМНК $\overrightarrow{\beta}$.*

Доказательство. Заметим сначала, что дисперсия любой линейной оценки $\overrightarrow{a}'\overrightarrow{y}$ пропорциональна квадрату длины вектора $\overrightarrow{a}$:

$$\begin{aligned}\mathbf{D}\,\overrightarrow{a}'\overrightarrow{y} &= \mathrm{Cov}(\overrightarrow{a}'\overrightarrow{y}) = \overrightarrow{a}'\,\mathrm{Cov}(\overrightarrow{y})\,\overrightarrow{a} = \\ &= \sigma^2\overrightarrow{a}'\overrightarrow{a} = \sigma^2\|\overrightarrow{a}\|^2.\end{aligned}$$

Пусть теперь $\overrightarrow{a}'\overrightarrow{y}$ – несмещенная оценка ψ и $\overrightarrow{a^*} = \mathrm{Proj}_{V_r}\overrightarrow{a}$ – проекция вектора $\overrightarrow{a}$ на подпространство V_r. Тогда $\overrightarrow{a} = \overrightarrow{a^*} + \overrightarrow{b}$, где $\overrightarrow{b} = \overrightarrow{a} - \overrightarrow{a^*}$, причем $\overrightarrow{b} \perp V_r$. Следовательно, (см.лемму I.1) $\overrightarrow{b}'\mathbf{X}' = \overrightarrow{0}'$. Это равенство позволяет доказать, что статистика $\overrightarrow{a^*}'\overrightarrow{y}$ также образует несмещенную оценку параметрической функции ψ :

$$\begin{aligned}\mathbf{E}\,\overrightarrow{a^*}'\overrightarrow{y} &= \mathbf{E}(\overrightarrow{a} - \overrightarrow{b})'\overrightarrow{y} = \mathbf{E}\,\overrightarrow{a}'\overrightarrow{y} - \mathbf{E}\,\overrightarrow{b}'\overrightarrow{y} = \\ &= \psi - \overrightarrow{b}'\mathbf{E}\,\overrightarrow{y} = \psi - \overrightarrow{b}'\mathbf{X}'\overrightarrow{\beta} = \psi.\end{aligned}$$

Легко видеть, однако, что дисперсия $\overrightarrow{a^*}'\overrightarrow{y}$ меньше дисперсии оценки $\overrightarrow{a}'\overrightarrow{y}$. Действительно, в силу теоремы «Пифагора»

$$\mathbf{D}\,\overrightarrow{a}'\overrightarrow{y} = \sigma^2\|\overrightarrow{a}\|^2 = \sigma^2\left(\|\overrightarrow{a^*}\|^2 + \|\overrightarrow{b}\|^2\right) \leqslant \sigma^2\|\overrightarrow{a^*}\|^2 = \mathbf{D}\,\overrightarrow{a^*}'\overrightarrow{y},$$

причем знак равенства достигается лишь в том случае, если $\overrightarrow{b} = \overrightarrow{0}$, т.е., когда вектор $\overrightarrow{a} \in V_r$.

Покажем теперь, что для любой несмещенной оценки $\overrightarrow{q}'\overrightarrow{y}$ функции ψ проекции $\mathrm{Proj}_{V_r}\overrightarrow{q}$ будут одинаковыми. В самом деле, так

как рассматриваемые оценки несмещены ($\mathbf{E}\,\overrightarrow{a^*}'\overrightarrow{y} = \psi = \mathbf{E}\,\overrightarrow{q^*}'\overrightarrow{y}$), то $\forall\,\overrightarrow{\beta}$

$$0 = \mathbf{E}\,\overrightarrow{a^*}'\overrightarrow{y} - \mathbf{E}\,\overrightarrow{q^*}'\overrightarrow{y} = (\overrightarrow{a^*} - \overrightarrow{q^*})'\mathbf{X}'\overrightarrow{\beta}\,.$$

Отсюда следует, что $(\overrightarrow{a^*} - \overrightarrow{q^*})'\mathbf{X}' = \overrightarrow{0}$, то есть $(\overrightarrow{a^*} - \overrightarrow{q^*}) \perp V_r$. С другой стороны, очевидно, $(\overrightarrow{a^*} - \overrightarrow{q^*}) \in V_r$. Это возможно только в том случае, если $\overrightarrow{a^*} = \overrightarrow{q^*}$, что завершает доказательство первой части теоремы.

Осталось доказать, что значение МНК-оценки $\widehat{\psi}$ может быть получено подстановкой ОМНК $\overrightarrow{\beta^*}$ в уравнение для ψ : $\overrightarrow{a^*}'\overrightarrow{y} = \overrightarrow{c}'\overrightarrow{\beta^*}$.

При доказательстве теоремы I.1 мы показали, что для ОМНК $\overrightarrow{\beta^*}$ вектор $\mathbf{X}'\overrightarrow{\beta^*} = \mathrm{Proj}_{V_r}\,\overrightarrow{y}$. Так как $\overrightarrow{a^*} \in V_r$, то $(\overrightarrow{y} - \mathbf{X}'\overrightarrow{\beta^*}) \perp \overrightarrow{a^*}$ и, следовательно,

$$\overrightarrow{a^*}'\overrightarrow{y} = \overrightarrow{a^*}'\mathbf{X}'\overrightarrow{\beta^*}.$$

С другой стороны, так как $\overrightarrow{a^*}'\overrightarrow{y}$ есть несмещенная оценка ψ, то

$$\psi = \overrightarrow{c}'\overrightarrow{\beta} = \mathbf{E}\,\overrightarrow{a^*}'\overrightarrow{y} = \overrightarrow{a^*}'\mathbf{X}'\overrightarrow{\beta}\,, \quad \forall\,\overrightarrow{\beta}\,.$$

Подставляя в это соотношение оценку $\overrightarrow{\beta^*}$ и сравнивая с предыдущим соотношением, получаем доказательство теоремы. $\otimes$

§ 4. Каноническая форма основных предположений

Вывод распределений оценок параметров и параметрических функций в значительной степени упрощается, если от переменных $\overrightarrow{y}$ перейти к так называемым каноническим переменным. Для этого определим ортонормированный базис $(\overrightarrow{\alpha}_1, \ldots, \overrightarrow{\alpha}_r, \overrightarrow{\alpha}_{r+1}, \ldots, \overrightarrow{\alpha}_n)$ во всём пространстве $\mathcal{R}^n$ таким образом, чтобы его первые r векторов образовывали базис в пространстве V_r. Теперь любой вектор $\overrightarrow{y}$ из $\mathcal{R}^n$ может быть записан в виде

$$\overrightarrow{y} = \sum_{i=1}^{n} z_i\,\overrightarrow{\alpha}_i,$$

где $z_i = \overrightarrow{\alpha}_i'\overrightarrow{y}$, $i = \overline{1,n}$, – координаты $\overrightarrow{y}$ при разложении по «осям» $(\overrightarrow{\alpha}_1, \ldots, \overrightarrow{\alpha}_n)$.

Определение 6. Координаты $\overrightarrow{z} = (z_1, \ldots, z_n)$ разложения наблюдаемого вектора $\overrightarrow{y}$ по базису $(\overrightarrow{\alpha}_1, \ldots, \overrightarrow{\alpha}_n)$, где первые r

базисных векторов совпадают с базисом пространства V_r , называются *каноническими переменными.*

Образуем ортогональную матрицу $P' = (\overrightarrow{\alpha}_1, \ldots, \overrightarrow{\alpha}_n)$, столбцы которой совпадают с векторами построенного базиса. Тогда $PP' = P'P = \mathbf{I}$, и связь между $\overrightarrow{y}$ и $\overrightarrow{z}$ может быть записана в матричном виде

$$\overrightarrow{z} = P\overrightarrow{y}, \qquad \overrightarrow{y} = P'\overrightarrow{z}.$$

Найдем среднее значение и матрицу ковариаций канонических переменных.

Обозначим $\zeta_i = \mathbf{E}\, z_i$, $i = \overline{1,n}$, – средние значения $\overrightarrow{z}$. Так как при $i > r$ базисные векторы $\overrightarrow{\alpha}_i \perp V_r$, то для этих $i > r$ произведение $\overrightarrow{\alpha}_i{}'\mathbf{X}' = \overrightarrow{0}{}'$ и средние значения

$$\zeta_i = \mathbf{E}\, z_i = \overrightarrow{\alpha}_i{}'\mathbf{E}\,\overrightarrow{y} = \overrightarrow{\alpha}_i{}'\mathbf{X}'\overrightarrow{\beta} = 0.$$

Ковариационная матрица

$$\mathrm{Cov}(\overrightarrow{z}) = \mathrm{Cov}(P\overrightarrow{y}) = P\,\mathrm{Cov}(\overrightarrow{y})P' = \sigma^2 PP' = \sigma^2\mathbf{I}$$

в силу ортогональности P.

Таким образом, основные предположения $Д_A$ могут быть переписаны в *канонической форме*:

$$\Omega: \quad \begin{cases} \overrightarrow{z} = \overrightarrow{\zeta} + \overrightarrow{\varepsilon} \\ \mathbf{E}\,\overrightarrow{\varepsilon} = \overrightarrow{0}, \quad \zeta_i = 0,\ i > r, \quad \mathrm{Cov}(\overrightarrow{z}) = \sigma^2\mathbf{I}. \end{cases} \tag{I.7}$$

Пространство ошибок

Оказывается, что в канонических переменных сумма квадратов ошибок SS_e (I.3) записывается через последние $n - r$ координат $\overrightarrow{z}$. Действительно, так как

$$\overrightarrow{y} = \sum_{i=1}^{n} z_i\overrightarrow{\alpha}_i, \qquad \text{то} \qquad \mathrm{Proj}_{V_r}\overrightarrow{y} = \sum_{i=1}^{r} z_i\overrightarrow{\alpha}_i.$$

Следовательно,

$$SS_e = \|\overrightarrow{y} - \mathrm{Proj}_{V_r}\overrightarrow{y}\|^2 = \|\sum_{i=r+1}^{n} z_i\overrightarrow{\alpha}_i\|^2 = \sum_{i=r+1}^{n} z_i^2. \tag{I.8}$$

Это обстоятельство служит одним из оснований для введения следующего определения.

Определение 7. Пространством ошибок называется линейное пространство, порожденное $\{z_{r+1}, \ldots, z_n\}$.

Пространство оценок

Определение 8. Пространство, порожденное $\{z_1, \ldots, z_r\}$, называется пространством оценок.

Введение этого пространства связано с тем, что для параметрической функции $\psi = \overrightarrow{c}' \overrightarrow{\beta}$, допускающей оценку, по теореме Гаусса-Маркова МНК-оценка имеет вид $\widehat{\psi} = \overrightarrow{a}' \overrightarrow{y}$, где $\overrightarrow{a} \in V_r$, т.е. $\overrightarrow{a}' \overrightarrow{\alpha}_j = 0$ для всех $j > r$. Следовательно,

$$\begin{aligned} \widehat{\psi} &= \overrightarrow{a}' \overrightarrow{y} = \overrightarrow{a}' P' \overrightarrow{z} = (\overrightarrow{a}' \overrightarrow{\alpha}_1, \ldots, \overrightarrow{a}' \overrightarrow{\alpha}_n) \overrightarrow{z} = \\ &= (\overrightarrow{a}' \overrightarrow{\alpha}_1, \ldots, \overrightarrow{a}' \overrightarrow{\alpha}_r, 0, \ldots, 0) \overrightarrow{z} = \sum_{i=1}^{r} c_i z_i, \end{aligned}$$

где $c_i = \overrightarrow{a}' \overrightarrow{\alpha}_i$, $i = 1, \ldots, r$.

Таким образом, МНК-оценка любой параметрической функции, допускающей оценку, есть линейная комбинация только первых r канонических переменных $(z_1, \ldots, z_r)$.

Несмещенная оценка σ^2

Хорошей иллюстрацией использования канонических переменных является вычисление математического ожидания SS_e и построение несмещенной оценки σ^2.

Теорема I.4. *Статистика*

$$\mathfrak{S}^2 = \frac{SS_e}{n - r}$$

есть несмещенная оценка σ^2.

Доказательство. В силу (I.8)

$$\mathbf{E}\, SS_e = \sum_{i=r+1}^{n} \mathbf{E}\, z_i^2.$$

Кроме того, по основным предположениям (I.7), $\mathbf{E}\, z_i^2 = \mathbf{D}\, z_i = \sigma^2$ при $i > r$, что, очевидно, доказывает теорему. $\otimes$

Г л а в а II

Дисперсионный анализ нормальных совокупностей

Чтобы построить статистические выводы относительно неизвестных параметров $\overrightarrow{\beta}$, необходимо конкретизировать вид распределения ошибок в основных предположениях $Д_A$. Наиболее продвинутые результаты в $Д_A$ получены в предположении, что ошибки имеют нормальное распределение. Таким образом, основные предположения $Д_A$ переписываются в следующем виде:

$$\Omega: \quad \overrightarrow{y} \ \sim \ \mathcal{N}_n(X'\overrightarrow{\beta}, \sigma^2\mathbf{I}). \tag{II.1}$$

Канонические переменные, как линейные комбинации $\overrightarrow{y}$, также будут распределены нормально с той же ковариационной матрицей $\sigma^2\mathbf{I}$ и нулевыми средними $\zeta_{r+1} = 0, \ldots, \zeta_p = 0$:

$$\Omega: \quad \overrightarrow{z} \ \sim \ \mathcal{N}_n(\overrightarrow{\zeta}, \sigma^2\mathbf{I}), \ \zeta_i = 0, i > r. \tag{II.2}$$

§ 1. Распределение оценок и ошибок

Найдем сначала распределение МНК-оценок параметрических функций, допускающих оценку, и распределение суммы квадратов ошибок SS_e .

Рассмотрим q параметрических функций, допускающих оценку,

$$\psi_i = \overrightarrow{c_i}'\overrightarrow{\beta} = \sum_{j=1}^{p} c_{ij}\beta_j, \quad i = 1, \ldots, q.$$

Пусть $\psi_i^* = \overrightarrow{a_i}'\overrightarrow{y}$ – МНК-оценка функции ψ_i, где вектор $\overrightarrow{a_i} \in V_r$. Обозначим через C и A матрицы, строки которых совпадают с вектор-строками $\overrightarrow{c_i}'$ и $\overrightarrow{a_i}', i = 1, \ldots, q$, соответственно. Тогда можно записать

$$\overrightarrow{\psi} = (\psi_1, \ldots, \psi_q)' = C\overrightarrow{\beta} \quad \text{и} \quad \overrightarrow{\psi^*} = (\psi_1^*, \ldots, \psi_q^*)' = A\overrightarrow{y}\,.$$

Теорема II.1. *Если q линейно независимых параметрических функций $\overrightarrow{\psi}$ допускают оценку, тогда*

1) МНК-оценки $\overrightarrow{\psi^*}$ не зависят от суммы квадратов ошибок SS_e ;

2) $\overrightarrow{\psi^*} \sim \mathcal{N}_q(\overrightarrow{\psi}, \sigma^2 B)$ с некоторой известной матрицей $B > 0$;

3) $\dfrac{SS_e}{\sigma^2} \sim \chi^2_{n-r}$.

Доказательство. 1) Оценки $\overrightarrow{\psi^*}$ принадлежат пространству оценок, а SS_e есть функция переменных пространства ошибок, другими словами, они зависят от $(z_1, \ldots, z_r)$ и $(z_{r+1}, \ldots, z_n)$, соответственно. В силу основных предположений эти совокупности случайных величин независимы, что доказывает первую часть теоремы.

2) Оценки $\overrightarrow{\psi^*}$ несмещены и являются линейными комбинациями $\overrightarrow{y}$, следовательно, вектор $\overrightarrow{\psi^*}$ также распределен нормально с вектором средних $\mathbf{E}\,\overrightarrow{\psi^*} = \overrightarrow{\psi}$ (в силу несмещенности). Вычислим его ковариационную матрицу

$$\mathrm{Cov}(\overrightarrow{\psi^*}) = \mathrm{Cov}(A\,\overrightarrow{y}) = A\,\mathrm{Cov}(\overrightarrow{y})A' = \sigma^2 AA'.$$

Для завершения доказательства второго пункта теоремы осталось показать, что матрица $B = AA'$ имеет ранг q.

Из следующей цепочки равенств

$$C\,\overrightarrow{\beta} = \overrightarrow{\psi} = \mathbf{E}\,\overrightarrow{\psi^*} = \mathbf{E}\,A\,\overrightarrow{y} = AX'\,\overrightarrow{\beta},$$

верной для всех $\overrightarrow{\beta}$, получаем, что $C = AX'$. Следовательно, ранг $\mathrm{rang}(C) \leqslant \mathrm{rang}(A) \leqslant q$. Однако, так как все q параметрических функций $\overrightarrow{\psi}$ линейно независимы, то $\mathrm{rang}(C) = q$. Поэтому $\mathrm{rang}(A) = q$ и $\mathrm{rang}(AA') = q$.

3) Как уже отмечалось, SS_e принадлежит пространству ошибок и

$$SS_e = \sum_{i=r+1}^{n} z_i^2,$$

причем все $z_i \sim \mathcal{N}_1(0, \sigma^2)$ и независимы. Таким образом, $SS_e\,/\sigma^2$ есть сумма квадратов $n-r$ независимых $\mathcal{N}_1(0,1)$ случайных величин. По определению эта сумма имеет хи-квадрат распределение с $n-r$ степенями свободы. $\otimes$

§ 2. Доверительные множества для функций, допускающих оценку. Критерий проверки гипотез

Утверждения теоремы II.1 позволяют строить доверительные множества для параметрических функций, допускающих оценку.

Пусть $\psi_1, \ldots, \psi_q$ – q линейно независимых параметрических функций, допускающих оценку и $\overrightarrow{\psi^*}$- вектор их МНК-оценок. По теореме II.1 $\overrightarrow{\psi^*} \sim \mathcal{N}_q(\overrightarrow{\psi}, \sigma^2 B)$ с некоторой матрицей B полного ранга q. Тогда, как известно, квадратическая форма

$$\frac{(\overrightarrow{\psi^*} - \overrightarrow{\psi})' B^{-1} (\overrightarrow{\psi^*} - \overrightarrow{\psi})}{\sigma^2} \sim \chi^2_q, \tag{II.3}$$

то есть распределена как хи-квадрат случайная величина с q степенями свободы. Кроме того, в силу утверждений 1) и 3) теоремы II.1, $SS_e\,/\sigma^2 \sim \chi^2_{n-r}$ и не зависит от квадратической формы (II.3). Следовательно, отношение

$$\frac{(\overrightarrow{\psi^*} - \overrightarrow{\psi})' B^{-1} (\overrightarrow{\psi^*} - \overrightarrow{\psi})/q}{SS_e\,/(n-r)} \sim F_{q,n-r}, \tag{II.4}$$

то есть имеет распределение Фишера с $(q, n-r)$ степенями свободы. Поэтому, если $F^{\alpha}_{q,n-r}$ – верхняя α-квантиль соответствующего распределения Фишера, тогда

$$\mathbf{P}\left\{ \frac{(\overrightarrow{\psi^*} - \overrightarrow{\psi})' B^{-1} (\overrightarrow{\psi^*} - \overrightarrow{\psi})/q}{\mathfrak{S}^2} \leqslant F^{\alpha}_{q,n-r} \right\} = 1 - \alpha \tag{II.5}$$

с $\mathfrak{S}^2 = SS_e\,/(n-r)$. Таким образом, имеет место

Теорема II.2. *В основных предположениях (II.1) множество тех* $\overrightarrow{\psi}$, *для которых выполняется неравенство*

$$(\overrightarrow{\psi^*} - \overrightarrow{\psi})' B^{-1} (\overrightarrow{\psi^*} - \overrightarrow{\psi}) \leqslant q\,\mathfrak{S}^2\, F^{\alpha}_{q,n-r}, \tag{II.6}$$

образует $(1-\alpha)$*-доверительное множество (доверительный эллипсоид) для вектора линейно независимых параметрических функций* $\overrightarrow{\psi} = (\psi_1, \ldots, \psi_q)$.

Случай зависимых параметрических функций сводится к только что рассмотренному путем сокращения количества оцениваемых

функций до максимального линейно независимого числа. Остальные функции будут линейно выражаться через выбранные.

В частном случае $q = 1$ распределение Фишера $\mathcal{F}_{1,n-r}$ совпадает с распределением квадрата «стьюдентовской» случайной величины с $n - r$ степенями свободы и доверительный эллипсоид представляет собой стандартный интервал числовой оси

$$\psi^* - t^{\alpha/2}_{n-r}\widehat{\sigma} \leqslant \psi \leqslant \psi^* + t^{\alpha/2}_{n-r}\widehat{\sigma},$$

где $\widehat{\sigma} = \mathfrak{S}|a|$ – оценка стандартного отклонения ψ^*, $t^{\alpha/2}_{n-r}$ – верхняя $\alpha/2$-квантиль распределения Стьюдента с $n - r$ степенями свободы.

Критерий проверки гипотезы, построенный по доверительному эллипсоиду

Доверительный эллипсоид, предложенный в теореме II.2, можно использовать для построения критерия проверки любых простых гипотез относительно набора параметрических функций. В частности, имеет место

Теорема II.3. *При проверке гипотезы*

$$\mathbf{H} \ : \ \psi_1 = \psi_2 = \cdots = \psi_q = 0$$

относительно q линейно независимых параметрических функций критерий, отвергающий гипотезу $\mathbf{H}$, *если* $\overrightarrow{\psi^*}' B^{-1} \overrightarrow{\psi^*} > q\mathfrak{S}^2 F^{\alpha}_{q,n-r}$, *имеет уровень* α.

Доказательство. Указанным критерием гипотеза $\mathbf{H}$ принимается тогда и только тогда, когда вектор $\overrightarrow{\psi} = 0$ принадлежит доверительному множеству (II.6). Следовательно, вероятность принятия гипотезы, если она верна, равна вероятности накрытия доверительным эллипсоидом (II.6) истинной точки $\overrightarrow{\psi} = 0$. По построению, эта вероятность равна $1 - \alpha$. $\otimes$

§ 3. Критерий отношения правдоподобий. Критерий Фишера

Как известно, наиболее мощный критерий проверки гипотез основан обычно на статистике отношения правдоподобий (ОП)

$$\lambda = \frac{\max\limits_{\omega} f(\overrightarrow{y})}{\max\limits_{\Omega} f(\overrightarrow{y})},$$

где f – плотность расределения наблюдений $\overrightarrow{y}$ и максимумы берутся по параметрам, определяющим плотность и принадлежащим либо основным предположениям Ω, либо множеству ω, характеризующему предположения гипотезы.

В рамках рассмотренной нами геометрической интерпретации основные предположения Ω задают r-мерное подпространство V_r (– пространство всех линейных комбинаций $\mathbf{X}'\overrightarrow{\beta}$) пространства $\mathcal{R}^n$. Предположения ω определяют $(r-q)$-мерное подпространство V_{r-q} пространства V_r . (Для примера, условия $y=0, z=0$ задают одномерное подпространство – ось OX трехмерного пространства $\mathcal{R}^3$).

Для вычисления статистики ОП выпишем плотность совместного распределения наблюдений $f(\overrightarrow{y}) = f_1(y_1)\cdots f_n(y_n)$. В наших предположениях каждое наблюдение получено из нормального распределения, то есть

$$f_i(y_i) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(y_i-\sum_{j=1}^{p}x_{ji}\beta_j\right)^2\right\}$$

и

$$f(\overrightarrow{y}) = \frac{1}{(2\pi\sigma^2)^{-n/2}}\exp\left\{-\frac{\mathcal{G}(\overrightarrow{y},\overrightarrow{\beta})}{2\sigma^2}\right\}, \tag{II.7}$$

где $\mathcal{G}(\overrightarrow{y},\overrightarrow{\beta}) = \|\overrightarrow{y}-\mathbf{X}'\overrightarrow{\beta}\|^2$ равно как раз той сумме квадратов, которая минимизировалась при построении ОМНК.

При фиксированном σ^2 максимум плотности (II.7) достигается в точках достижения минимума $\mathcal{G}$. В случае Ω это есть точки $\overrightarrow{\beta^*}$ (обозначим их здесь $\overrightarrow{\beta_\Omega}$), для которых $\mathbf{X}'\overrightarrow{\beta_\Omega} = \mathrm{Proj}_{V_r}\overrightarrow{y}$. Аналогично, точки достижения максимума плотности по ω суть множество точек $\overrightarrow{\beta_\omega}$, для которых $\mathbf{X}'\overrightarrow{\beta_\omega} = \mathrm{Proj}_{V_{r-q}}\overrightarrow{y}$. Соответствующие минимумы функции $\mathcal{G}$обозначим $\mathcal{G}_\Omega$ и $\mathcal{G}_\omega$. Таким образом, для нахождения максимума плотности (по Ω и ω) необходимо найти теперь максимум по σ^2 выражений вида

$$\frac{1}{(2\pi\sigma^2)^{-n/2}}\exp\left\{-\frac{\mathcal{G}^*}{2\sigma^2}\right\}, \tag{II.8}$$

где $\mathcal{G}^*$ равно $\mathcal{G}_\Omega$ или $\mathcal{G}_\omega$. Стандартным способом легко устанавливается, что максимум (II.8) достигается при

$$\sigma^2 = \frac{\mathcal{G}^*}{n}.$$

Подставляя полученные значения в (II.7), получаем статистику ОП

$$\lambda = \left(\frac{\mathcal{G}_\omega}{\mathcal{G}_\Omega}\right)^{-n/2}. \tag{II.9}$$

Для построения критерия нам необходимо теперь найти распределение этой статистики. Оказывается, что следующее взаимно однозначное преобразование статистики λ будет иметь хорошо известное распределение Фишера и, следовательно, более удобно с практической точки зрения.

Теорема II.4. *Если верна гипотеза* **H**, *то статистика*

$$\mathcal{F} = \frac{n-r}{q} \cdot \frac{\mathcal{G}_\omega - \mathcal{G}_\Omega}{\mathcal{G}_\Omega} \sim F_{q,n-r}.$$

Доказательство. Для доказательства теоремы несколько переопределим базис, построенный при определении канонических переменных. Рассмотрим три пространства $V_{r-q} \subset V_r \subset \mathcal{R}^n$ и выберем ортонормированный базис в пространстве $\mathcal{R}^n$ по следующей схеме

$$\underbrace{\underbrace{\overrightarrow{\alpha}_1, \ldots, \overrightarrow{\alpha}_q, \underbrace{\overrightarrow{\alpha}_{q+1}, \ldots, \overrightarrow{\alpha}_r}_{\text{базис } V_{r-q}}}_{\text{базис } V_r}, \overrightarrow{\alpha}_{r+1}, \ldots, \overrightarrow{\alpha}_n}_{\text{базис } \mathcal{R}^n}.$$

Как было показано ранее, средние значения канонических переменных $\mathbf{E}\, z_i = 0$ для $i = r+1, \ldots, n$. Аналогично показывается, что, если верна гипотеза **H** (то есть регрессия $\mathbf{X}'\overrightarrow{\beta} \in V_{r-q}$), то средние значения $\mathbf{E}\, z_i = 0$ и для $i = 1, \ldots, r$.

Далее, так как соответствующие ОМНК найдены через проекции на пространства V_r и V_{r-q} , то

$$\mathcal{G}_\Omega = \| \overrightarrow{y} - \mathrm{Proj}_{V_r} \overrightarrow{y} \|^2 = \| \sum_{i=1}^{n} z_i \overrightarrow{\alpha}_i - \sum_{i=1}^{r} z_i \overrightarrow{\alpha}_i \|^2 = \sum_{i=r+1}^{n} z_i^2.$$

и, аналогично,

$$\mathcal{G}_\omega = \sum_{i=1}^{q} z_i^2 + \sum_{i=r+1}^{n} z_i^2.$$

Таким образом, статистика отношения правдоподобий записывается через канонические переменные в виде

$$\mathcal{F} = \frac{n-r}{q} \frac{\sum\limits_{i=1}^{q} z_i^2}{\sum\limits_{i=r+1}^{n} z_i^2} = \frac{n-r}{q} \frac{\sum\limits_{i=1}^{q} \left(\frac{z_i}{\sigma}\right)^2}{\sum\limits_{i=r+1}^{n} \left(\frac{z_i}{\sigma}\right)^2}.$$

Если верна гипотеза $\mathbf{H}$, то, как отмечалось выше, переменные z_i в числителе имеют среднее 0. Следовательно, числитель статистики $\mathcal{F}$ есть сумма квадратов независимых $\mathcal{N}_1(0,1)$ случайных величин и поэтому имеет хи-квадрат распределение с q степенями свободы. Знаменатель $\mathcal{F}$ (в любом случае) имеет хи-квадрат распределение с $n-r$ степенями свободы, что в силу условия независимости канонических переменных доказывает утверждение теоремы. $\otimes$

Замечание i. В знаменателе статистики $\mathcal{F}$ стоит сумма квадратов ошибок SS_e. Среднее значение этой суммы $SS_e/(n-r)$ (равное $\mathfrak{S}^2$ – несмещенной оценке σ^2) принято обозначать $\overline{SS_e}$. Сумму квадратов в числителе $\mathcal{F}$ обозначают SS_H, а её среднее $\overline{SS_H} = SS_H/q$. В этих обозначениях статистика

$$\mathcal{F} = \frac{\overline{SS_H}}{\overline{SS_e}}.$$

F-критерий проверки линейных гипотез

С учетом полученного распределения статистики отношения правдоподобий можно предложить следующий критерий проверки гипотез.

Определение 1. Критерий проверки гипотезы $\mathbf{H}: \psi_1 = \psi_2 = \cdots = \psi_q = 0$, относительно q линейно независимых параметрических функций, отвергающий $\mathbf{H}$ при значениях статистики

$$\mathcal{F} > F^{\alpha}_{q,n-r},$$

где $F^{\alpha}_{q,n-r}$ – верхняя α-квантиль распределения Фишера с $(q, n-r)$ степенями свободы, называется *F-критерием.*

Замечание ii. Алгебраическими методами легко показывается, что F-критерий эквивалентен построенному выше критерию, основанному на доверительном эллипсоиде для вектора параметрических функций. Не вдаваясь в подробности, заметим только, что гипотезу $\mathbf{H}$ можно записать в эквивалентной форме $\mathbf{H}: \zeta_1 = \cdots = \zeta_q = 0$ относительно средних значений канонических переменных. Несмещенными оценками с минимальной дисперсией этих неизвестных параметров являются, очевидно, сами канонические переменные – $\zeta_1^* = z_1, \ldots, \zeta_q^* = z_q$. Ковариационная матрица вектора $(z_1, \ldots, z_q)'$ пропорциональна единичной матрице, поэтому числитель статистики II.6 полностью совпадает с числителем статистики F-критерия. Знаменатели этих статистик равны по определению.

Замечание iii. Мощность F-критерия равна

$$\mathbf{P}\{\mathcal{F} > F^{\alpha}_{q,n-r}\},$$

где вероятность вычисляется при общих предположениях Ω. В этих предположениях средние значения $\zeta_i = \mathbf{E}\, z_i \, (i = 1, \ldots, q)$ уже не равны 0 и числитель статистики отношения правдоподобий будет иметь нецентральное хи-квадрат распределение, а сама статистика нецентральное распределение Фишера. Параметр нецентральности равен

$$\delta^2 = \frac{\sum\limits_{i=1}^{q} \zeta_i^2}{\sigma^2}.$$

Таким образом, если неизвестные значения $\zeta_1, \ldots, \zeta_q$ и σ^2, входящие в δ^2, заменить на их соответствующие оценки, то мощность F-критерия может быть приближенно вычислена с помощью таблиц нецентрального распределения Фишера.

Последнее, что необходимо теперь сделать, – сравнить мощность построенного F-критерия с мощностью других возможных критериев уровня α. Утверждать, что F-критерий имеет наибольшую мощность среди всех критериев уровня α нельзя, однако ему присущ ряд оптимальных свойств, которые приведем здесь без доказательства.

Во-первых, для произвольного критерия W проверки гипотезы $\mathbf{H}$ определим мощность $\mathfrak{b}(\theta, W)$ как вероятность отвергнуть гипотезу $\mathbf{H}$, когда истинное значение вектора неизвестных параметров есть θ. Критерий уровня α имеет мощность $\mathfrak{b}(\theta, W) \leqslant \alpha$ для всех θ, удовлетворяющих условиям гипотезы $\mathbf{H}$ (коротко, $\theta \in \mathbf{H}$).

Теорема II.5. *Среди всех критериев уровня α, мощность которых зависит от неизвестных параметров только через значения параметра нецентральности δ^2, F-критерий имеет равномерно наибольшую мощность при $\theta \notin \mathbf{H}$.*

Далее, пусть $\mathcal{W}_\alpha$ – класс критериев уровня α. Рассмотрим огибающую функцию мощности

$$\tilde{\mathfrak{b}}(\theta) = \sup_{W \in \mathcal{W}_\alpha} \mathfrak{b}(\theta, W)$$

как (недостижимый) эталон мощности критерия. Назовем силой критерия W максимум расхождения между огибающей функцией мощности и мощностью W:

$$\sup_{\theta \notin \mathbf{H}} (\tilde{\mathfrak{b}}(\theta) - \mathfrak{b}(\theta, W)). \tag{II.10}$$

Критерий, минимизирующий (II.10), называется *наиболее сильным.*

Теорема II.6. *F-критерий – наиболее сильный критерий уровня α.*

Г л а в а III

Модели дисперсионного анализа

§ 1. Однофакторный дисперсионный анализ

Простейшим случаем *ДА* является *однофакторный анализ*. Этот термин относится к сравнению средних нескольких одномерных популяций. Обозначим эти средние через $\beta_1, \ldots, \beta_p$ и предположим, что наблюдения в каждой популяции имеют нормальное распределение с одинаковой дисперсией σ^2. Пусть в каждой из популяций было произведено $n_1, \ldots, n_p$ наблюдений, соответственно. Тогда i-ое наблюдение в j-ой популяции может быть записано в виде

$$y_{ji} = \beta_j + \varepsilon_{ji}, \tag{III.1}$$

где ε_{ji} – ошибка наблюдения. К данным подобного рода может быть применена общая теория, разработанная в предыдущих главах.

Представим соотношение III.1 в матричной форме. Для этого рассмотрим p-мерный вектор $\overrightarrow{u_j} = (0, \ldots, 0, 1, 0, \ldots, 0)'$ с единственным ненулевым элементом, стоящим на j-ом месте, и матрицу

$$\mathbf{X} = (\overrightarrow{u_1}, \ldots, \overrightarrow{u_1}, \ldots, \overrightarrow{u_p}, \ldots, \overrightarrow{u_p}),$$

у которой первые n_1 столбцов совпадают с вектором $\overrightarrow{u_1}$, последующие n_2 столбцов совпадают с $\overrightarrow{u_2}$ и т.д., последние n_p столбцов совпадают с вектором $\overrightarrow{u_p}$. Тогда вектор наблюдений

$$\overrightarrow{y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{pn_p} \end{pmatrix} = \mathbf{X}'\overrightarrow{\beta} + \overrightarrow{\varepsilon} = \begin{pmatrix} \overrightarrow{u_1}' \\ \vdots \\ \overrightarrow{u_1}' \\ \vdots \\ \overrightarrow{u_p}' \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \overrightarrow{\varepsilon}$$

с соответствующим вектором ошибок $\overrightarrow{\varepsilon}$.

Ранее было установлено, что ОМНК $\overrightarrow{\beta^*}$ параметров $\overrightarrow{\beta}$ удовлетворяют нормальным уравнениям $\mathbf{X}\,\mathbf{X}'\overrightarrow{\beta^*} = \mathbf{X}\,\overrightarrow{y}$, причем, если $r = \operatorname{rang}\mathbf{X} = p$, то $\overrightarrow{\beta^*} = \mathcal{S}^{-1}\,\mathbf{X}\,\overrightarrow{y}$ с информационной матрицей $\mathcal{S} = \mathbf{X}\,\mathbf{X}'$. В рассматриваемом случае информационная матрица

$$\mathcal{S} = \mathbf{X}\,\mathbf{X}' = (\overrightarrow{u_1}, \ldots, \overrightarrow{u_1}, \ldots, \overrightarrow{u_p}) \begin{pmatrix} \overrightarrow{u_1}' \\ \vdots \\ \overrightarrow{u_1}' \\ \vdots \\ \overrightarrow{u_p}' \end{pmatrix} = \sum_{i=1}^{n_1} \overrightarrow{u_1}\overrightarrow{u_1}' + \ldots + \sum_{i=1}^{n_p} \overrightarrow{u_p}\overrightarrow{u_p}'.$$

Произведение $\overrightarrow{u_j}\,\overrightarrow{u_j}'$ образует матрицу $p \cdot p$, у которой все элементы равны 0, кроме j-ого диагонального элемента, который равен 1. Таким образом, информационная матрица равна диагональной матрице

$$\mathcal{S} = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & n_p \end{pmatrix} \quad \text{и} \quad \mathcal{S}^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{1}{n_p} \end{pmatrix}.$$

Легко видеть, что

$$\mathbf{X}\,\overrightarrow{y} = \left(\sum_{i=1}^{n_1} y_{1i}, \ldots, \sum_{i=1}^{n_p} y_{p\,i} \right).$$

Следовательно,

$$\overrightarrow{\beta^*} = (y_{1\bullet}, \ldots, y_{p\bullet})' \,, \tag{III.2}$$

где с целью сокращения записи введено обозначение $y_{j\bullet} = \frac{1}{n_j} \sum\limits_{i=1}^{n_j} y_{ji}$.

Сумма квадратов ошибок равна

$$\mathcal{G}_\Omega = SSe = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ji} - y_{j\bullet})^2 = \sum_{ji} y_{ji}^2 - \sum_{j=1}^{p} n_j y_{j\bullet}^2.$$

Последнее равенство является следствием известного утверждения (сравните с двумя способами вычисления дисперсии), которое мы будем часто использовать в дальнейшем изложении.

Лемма III.1. *Для любого набора чисел $z_1, \ldots, z_m$ сумма*

$$\sum_{i=1}^{m} (z_i - z_\bullet)^2 = \sum_{i=1}^{m} z_i^2 - m\, z_\bullet^{\,2} \,.$$

Следуя общей теории, можно найти распределение вектора ОМНК. Заметим сначала, что так как существует обратная матрица $\mathcal{S}^{-1}$, то ранг нашей задачи $r = p$. Следовательно, в силу следствия 1 ковариационная матрица $\mathrm{Cov}\,\overrightarrow{\beta^*} = \sigma^2\,\mathcal{S}^{-1} = \sigma^2\mathbf{I}$. Далее, параметрические функции $\psi_j(\overrightarrow{\beta}) = \beta_j\,(j = 1, \ldots, p)$ допускают оценку $(= y_{j\bullet})$, отсюда, как следствие теоремы (II.1), получаем следующее утверждение, которое здесь легче доказать непосредственно.

Теорема III.1. *В основных предположениях с нормальным распределением ошибок ОМНК* $\beta_j^* = y_{j\bullet} \sim \mathcal{N}_1(\beta_j, \sigma^2/n_j)\ (j = 1, \ldots, p)$ *и независимы.*

Перейдем теперь к вопросу проверки гипотез. Обычно в данной ситуации проверяется стандартная гипотеза о равенстве средних значений всех популяций

$$\mathbf{H}: \beta_1 = \beta_2 = \ldots = \beta_p.$$

В терминах равенства нулю параметрических функций эта гипотеза записывается в виде

$$\mathbf{H}: \psi_1(\overrightarrow{\beta}) = \psi_2(\overrightarrow{\beta}) = \ldots = \psi_{p-1}(\overrightarrow{\beta}) = 0,$$

где функции $\psi_j(\overrightarrow{\beta}) = \beta_j - \beta_p$.

Для построения F-критерия необходимо найти теперь ОМНК в предположениях $\mathbf{H}$. Вычислим их непосредственно.

Если справедлива гипотеза $\mathbf{H}$, тогда $\mathbf{X}'\overrightarrow{\beta} = (\beta, \ldots, \beta)'$ и, следовательно, для нахождения ОМНК необходимо найти минимум функции

$$\| \overrightarrow{y} - \mathbf{X}'\overrightarrow{b} \|^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(y_{ji} - b)^2$$

по параметру b. Другими словами, необходимо найти константу b, дающую наилучший среднеквадратический прогноз всего вектора данных $\overrightarrow{y}$. Хорошо известно, что оптимальная константа равна среднему значению всех данных. Таким образом, ОМНК неизвестного параметра β в предположениях гипотезы равна

$$\beta_\omega = y_{\bullet\bullet} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j} y_{ji} = \frac{1}{n}\sum_{j=1}^{p} n_j y_{j\bullet}\,.$$

Сумма квадратов ошибок (см.Лемму III.1)

$$\mathcal{G}\omega = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(y_{ji} - y_{\bullet\bullet})^2 = \sum_{ji} y_{ji}^2 - n y_{\bullet\bullet}^2$$

и

$$SS_H = \mathcal{G}_\omega - \mathcal{G}_\Omega = \sum_{j=1}^{p} n_j y_{j\cdot}^2 - n y_{\cdot\cdot}^2 = \sum_{j=1}^{p} n_j (y_{j\cdot} - y_{\cdot\cdot})^2 \ .$$

В рассматриваемом случае гипотеза задается $q = p - 1$ линейно независимыми функциями. Следовательно, статистика F-критерия

$$\mathcal{F} = \frac{n-p}{p-1} \cdot \frac{\sum\limits_{j=1}^{p} n_j (y_{j\cdot} - y_{\cdot\cdot})^2}{\sum\limits_{ji} y_{ji}^2 - \sum\limits_{j=1}^{p} n_j y_{j\cdot}^2} \ \sim F_{p-1,n-p}.$$

Сведем полученные результаты в единую таблицу.

Таблица однофакторного *ДА*

Источник разброса	SS	Ст.свободы	$\overline{SS}$	$\mathcal{F}$	α
Между группами	SS_H	$\nu_H = p-1$	$\overline{SS_H}$	$\dfrac{\overline{SS_H}}{\overline{SS_e}}$	$1 - F_{\nu_H,\nu_e}(\mathcal{F})$
Внутри групп	SS_e	$\nu_e = n-p$	$\overline{SS_e}$	–	–
Полная сумма квадратов	SS_Π	$n-1$	–	–	–

Разбиение полной суммы квадратов

Легко видеть, что справедливо соотношение

$$\sum_{ji} y_{ji}^2 - n y_{\cdot\cdot}^2 = SS_e \ + SS_H \ ,$$

которое называется разбиением полной суммы квадратов. Левая часть этого равенства обозначается обычно SS_Π . С помощью этого равенства легче вычислить $SS_e = SS_\Pi - SS_H$.

§ 2. Сравнения

Если гипотеза о равенстве всех средних отвергается, то возникает задача получения дополнительной информации о различающихся популяциях. Для этой цели рекомендуется использовать метод сравнений. Введем

Определение 1. Сравнением параметров $\beta_1, \ldots, \beta_p$ называется параметрическая функция $\psi = \sum c_j \beta_j$ с коэффициентами, удовлетворяющими условию $\sum c_j = 0$.

Пример. Если требуется сравнить две популяции (например, первую и вторую), естественно рассмотреть параметрическую функ-

цию $\psi = \beta_1 - \beta_2$, которая будет, очевидно, сравнением. С другой стороны, если первые k и, соответственно, последние $p-k$ популяций можно объединить по какому-либо признаку в две группы, то для оценки различий в этих группах можно использовать сравнение $\psi = \sum_{j=1}^{k} \beta_j/k - \sum_{j=k+1}^{p} \beta_j/(p-k)$.

В однофакторном *ДА* любое сравнение допускает оценку. А именно, для сравнения $\psi = \sum c_j \beta_j$ МНК-оценкой (по теореме Гаусса-Маркова) будет статистика $\psi^* = \sum c_j y_{j\cdot}$. Дисперсия этой оценки

$$\mathbf{D}(\psi^*) = \sum_{j=1}^{p} c_j^2 \, \mathbf{D}(y_{j\cdot}) = \sigma^2 \sum_{j=1}^{p} \frac{c_j^2}{n_j},$$

а её оценка

$$\mathfrak{D}(\psi^*) = \mathfrak{S}^2 \sum_{j=1}^{p} \frac{c_j^2}{n_j}. \tag{III.3}$$

В обычной практике применения статистических методов сразу после отвержения основной гипотезы рассматривается некоторая вспомогательная гипотеза о равенстве нулю какого-либо сравнения, к проверке которой применяются аналогичные методы (F-критерий). Однако такой составной вывод будет уже иметь уровень значимости выше ожидаемого α. Г.Шеффе предложил метод (называемый в честь автора S-методом), позволяющий оценить сразу все сравнения, связанные с основной гипотезой.

Рассмотрим набор из q линейно независимых сравнений $\psi_1, \ldots, \psi_q$, (описывающих гипотезу H) и пространство $\mathfrak{L}$, порожденное этими функциями. Справедлива

Теорема III.2. *Положим* $\Delta = qF^{\alpha}_{q,n-r}$. *Вероятность того, что*

$$\psi^* - \sqrt{\Delta \mathfrak{D}(\psi^*)} \leqslant \psi \leqslant \psi^* + \sqrt{\Delta \mathfrak{D}(\psi^*)} \tag{III.4}$$

одновременно для всех сравнений $\psi \in \mathfrak{L}$, *равна* $1-\alpha$.

Связь этой теоремы с F-критерием объясняет следующая

Теорема III.3. *F-критерий отвергает гипотезу H тогда и только тогда, когда для какого-либо сравнения* ψ *интервал III.4 не накрывает точку* $\psi = 0$.

Таким образом, если гипотеза **H** отвергается F-критерием, то, применяя доверительный интервал III.4, можно попытаться отыскать сравнение, способствовавшее этому выводу.

Рассмотрим следующий

Пример. Сравниваются веса новорожденных поросят в восьми опоросах. Были получены следующие данные.

1	2	3	4	5	6	7	8
2.0	3.5	3.3	3.2	2.6	3.1	2.6	2.5
2.8	2.8	3.6	3.3	2.6	2.9	2.2	2.4
3.3	3.2	2.6	3.2	2.9	3.1	2.2	3.0
3.2	3.5	3.1	2.9	2.0	2.5	2.5	1.5
4.4	2.3	3.2	3.3	2.0		1.2	
3.6	2.4	3.3	2.5	2.1		1.2	
2.9	2.0	2.9	2.6				
2.5	1.6	3.4	2.8				
2.8		3.2					
2.1		3.2					

Требуется с 10%-ым уровнем значимости ($\alpha = 0.1$)

а) проверить гипотезу отсутствия различий между средними весами в восьми приплодах;

б) в предположении, что опоросы 1,3 и 4 получены от одной свиноматки, а остальные пять от другой, установить значимость различия между средними весами в этих двух группах;

в) проверить значимость различия между средними весами в больших опоросах (1,2,3,4) и меньших (5,6,7,8).

С целью упрощения вычислений преобразуем данные, умножая их на 10 (убираем десятичную точку) и вычитая 25 (сдвигаем к приблизительному общему центру). В следующей таблице наряду с преобразованными данными указаны и их квадраты:

1		2		3		4		5		6		7		8	
Y :	Y^2	Y :	Y^2	Y :	Y^2	Y :	Y^2	Y :	Y^2	Y :	Y^2	Y :	Y^2	Y :	Y^2
−5 :	25	10 :	100	8 :	64	7 :	49	1 :	1	6 :	36	1 :	1	0 :	0
3 :	9	3 :	9	11 :	121	8 :	64	1 :	1	4 :	16	−3 :	9	−1 :	1
8 :	64	7 :	49	1 :	1	7 :	49	4 :	16	6 :	36	−3 :	9	5 :	25
7 :	49	10 :	100	6 :	36	4 :	16	−5 :	25	0 :	0	0 :	0	−10 :	100
19 :	361	−2 :	4	7 :	49	8 :	64	−5 :	25	:		−13 :	169	:	
11 :	121	−1 :	1	8 :	64	0 :	0	−4 :	16	:		−13 :	169	:	
4 :	16	−5 :	25	4 :	16	1 :	1	:		:		:		:	
0 :	0	−9 :	81	9 :	81	3 :	9	:		:		:		:	
3 :	9	:		7 :	49	:		:		:		:		:	
−4 :	16	:		7 :	49	:		:		:		:		:	
10		8		10		8		6		4		6		4	
46 :		13 :		68 :		38 :		−8 :		16 :		−31 :		−6 :	
:	670	:	369	:	530	:	252	:	84	:	88	:	357	:	126
4.6 :		1.63 :		6.8 :		4.75 :		−1.3 :		4 :		−5.2 :		−1.5 :	
:	21.16	:	2.64	:	46.24	:	22.56	:	1.77	:	16	:	26.73	:	2.25
:	211.6	:	21.13	:	462.4	:	180.5	:	10.7	:	64	:	160.2	:	9

Кроме того, в этой таблице после выборочных данных приведены:

в 1-ой строке – количества данных в популяциях – n_j;
во 2-ой строке – суммы первых степеней – $\sum\limits_{i=1}^{n_j} y_{ji}$;
в 3-ей строке – суммы квадратов в популяциях – $\sum\limits_{i=1}^{n_j} y_{ji}^2$;
в 4-ой строке – выборочные средние в популяциях – $y_{j\bullet}$;
в 5-ой строке – их квадраты – $y_{j\bullet}^2$;
в 6-ой строке – произведения $n_j y_{j\bullet}^2$.
Суммируя значения в последних шести строках, получаем

$n = 56$ – сумма в 1-ой строке;
$y_{\bullet\bullet} = \frac{136}{56} = 2.43$ – во 2-ой строке;
$y_{\bullet\bullet}^2 = 5.9 \quad n y_{\bullet\bullet}^2 = 330.3$
$SS_{\text{П}} = 2476 - 330.3 = 2145.7$ – в 3-ей строке;
$SS_e = 2476 - 1119.5 = 1356.5$ – в 6-ой строке;
$\mathfrak{S}^2 = \overline{SS_e} = 1356.3/(56-8) = 28.26$ – оценка дисперсии;
$SS_H = 2145.7 - 1356.5 = 789.2$.

Таким образом, найдя по таблицам распределения Фишера критический уровень значимости $\alpha_{\text{кр}} = 1 - F_{7,48}(3.99) = 0.0016$, получаем следующую таблицу:

Таблица однофакторного *ДА*

Источник разброса	SS	Ст. свободы	$\overline{SS}$	$\mathcal{F}$	α
Между группами	789.2	7	112.7	3.99	0.0016
Внутри групп	1356.5	48	28.26		
Полная сумма квадратов	2145.7	55			

Гипотеза о равенстве средних значений во всех опоросах отвергается с уровнем значимости $0.001 < 0.1$. Перейдем теперь к решению следующих двух поставленных задач.

Для решения второй задачи (сравнение среднего веса в 1,3 и 4 опоросах с остальными) рассмотрим сравнение

$$\psi = \frac{1}{3}(\beta_1 + \beta_3 + \beta_4) - \frac{1}{5}(\beta_2 + \beta_5 + \beta_6 + \beta_7 + \beta_8).$$

ОМНК этой параметрической функции равна $\psi^* = 5.85$
Оценка дисперсии (по формуле III.3) $\mathfrak{D}(\psi^*) = 28.01 \cdot 0.074 = 2.07$.
По таблице распределения Фишера со степенями свободы 7 и 48 находим верхнюю 10%-ую точку $F_{7,48}^{0.1} = 1.85$. Поэтому $\Delta = 7 \cdot 1.85 =$

12.95 , $\sqrt{\Delta\mathfrak{D}(\psi^*)} = \sqrt{26.6} = 5.16$ и доверительный интервал для сравнения равен $(0.69; 11.01)$. Так как этот интервал не содержит 0, то можно утверждать, что выбор свиноматки влияет на вес приплода.

Решая последнюю поставленную задачу, рассмотрим сравнение

$$\psi = \frac{1}{4}(\beta_1 + \beta_2 + \beta_3 + \beta_4) - \frac{1}{4}(\beta_5 + \beta_6 + \beta_7 + \beta_8).$$

Вычисления, подобные вышеприведенным, дают следующие результаты:

$$\psi^* = 5.45, \quad \mathfrak{D}(\psi^*) = 2.24, \quad \sqrt{\Delta\mathfrak{D}(\psi^*)} = 5.39\,.$$

Доверительный интервал для ψ – $(0.06, 9.94)$ также не содержит 0 и, следовательно, можно утверждать, что в более крупных опоросах вес приплода значимо выше.

§ 3. Непараметрические критерии однородности

В том случае, когда нет оснований предполагать нормальность распределения ошибок, применяют так называемые непараметрические (или свободные от распределения) методы анализа однородности различных групп. Эти методы чаще всего основываются на ранговых статистиках, то есть не на абсолютных значениях наблюдаемого отклика, а на их относительных местах при расположении в единый ряд.

Пусть r_{ji}– ранг (место) наблюдения y_{ji} в упорядоченной по возрастанию совокупности всех наблюденных значений отклика $\{y_{ji}\}$. Очевидно, сумма всех рангов от 1 до n равна $n(n+1)/2$, а их общее среднее значение равно $(n+1)/2$. Поэтому в качестве меры отклонения j-ой группы от среднего уровня может быть взята величина $(r_{j\cdot} - (n+1)/2)^2$, где, как обычно, $r_{j\cdot}$ – средний ранг наблюдений, попавших в j-ую группу. Если гипотеза однородности групп верна, то следует ожидать, что значения этой меры во всех группах будут близки к нулю.

Определение 2. Статистика

$$KW = \frac{12}{n(n+1)} \sum_{j=1}^{p} n_j \left(r_{j\cdot} - \frac{n+1}{2} \right)^2 = \frac{12}{n(n+1)} \sum_{j=1}^{p} n_j r_{j\cdot}^2 - 3(n+1)$$

называется *статистикой Краскела-Уоллиса.*

Теорема III.4. *1) Если наблюдения во всех p группах имеют одинаковое распределение, то есть группы однородны, тогда*

$$KW \rightsquigarrow \chi^2_{p-1}, \; n \to \infty.$$

2) Пусть K_{p-1} – функция распределения хи-квадрат с $(p-1)$-ой степенью свободы, тогда критерий, отвергающий гипотезу однородности при критическом уровне значимости

$$\alpha_{кр} = 1 - K_{p-1}(KW) < \alpha,$$

имеет асимптотический уровень α.

Замечание i. При малых объемах выборок распределение статистики KW может быть найдено точно. Идея его построения основывается на том факте, что в предположениях гипотезы все возможные расположения рангов в группах равновероятны.

Замечание ii. На практике чаще всего встречается ситуация, когда наблюденные значения отклика содержат совпадающие элементы. В этом случае при малом числе совпадений рекомендуется использовать средние ранги. Так, например, если четыре равных значения отклика занимают места с 7-ого по 10-е, то ранги всех этих наблюдений выбираются равными 8.5. Если совпадений слишком много, то статистику KW необходимо модернизировать следующим образом

$$KW' = \frac{KW}{1 - \sum_{k=1}^{g} T_k/(n^3 - n)},$$

где g – число блоков совпадающих наблюдений, $T_k = (t_k^3 - t_k), t_k$ – число совпадающих наблюдений в k-ом блоке.

Замечание iii. При $p = 2$ статистика Краскела-Уоллиса по своему действию эквивалентна статистике Уилкоксона и статистике Манна-Уитни.

Нередко на практике известно, что группы упорядочены по возрастанию влияния фактора. В этом случае можно предложить критерий с более высокой мощностью.

Рассмотрим две произвольные группы j и k. Вычислим статистику Манна-Уитни для этих групп

$$U_{jk} = \sum_{il} \phi(y_{ji}, y_{kl}),$$

где

$$\phi(u, v) = \begin{cases} 1 \text{ , если } u < v; \\ 1/2 \text{ , если } u = v; \\ 0 \text{ , если } u > v. \end{cases}$$

Определение 3. Статистика

$$I = \sum_{1 \leqslant j < k \leqslant p} U_{jk}$$

называется *статистикой Джонкхиера.*

Теорема III.5. *1) Если наблюдения во всех группах однородны, тогда статистика Джонкхиера асимптотически нормальна со средним*

$$m = \frac{1}{4}(n^2 - \sum_{j=1}^{k} n_j^2)$$

и дисперсией

$$d^2 = \frac{1}{72}\left[n^2(2n+3) - \sum_{j=1}^{p} n_j^2(2n_j+3)\right].$$

2) Критерий, отвергающий гипотезу однородности в пользу альтернативы возрастающего влияния эффектов при критическом уровне значимости

$$\alpha_{кр} = 1 - \Phi\left(\frac{I-m}{d}\right) < \alpha,$$

имеет асимптотический уровень α.

Оценка сравнений

Определение 4. Медиана всевозможных разностей пар выборочных данных, попавших в j-ую и k-ую группы,

$$z_{jk} = \mathrm{med}(y_{ji} - y_{kl}, \quad i = \overline{1, n_j},\ l = \overline{1, n_k}),$$

называется статистикой Ходжеса-Лемана.

Замечание iv. Статистика Ходжеса-Лемана представляет собой непараметрическую оценку сдвига между эффектами j-ой и k-ой групп $\tau_{jk} = \beta_j - \beta_k$. Эта оценка предпочтительна в ситуациях, когда ничего неизвестно о распределении ошибки наблюдений. Кроме того, оценка Ходжеса-Лемана в сравнении со стандартной оценкой, равной разности выборочных средних в группах, менее чувствительна к наличию резко выделяющихся наблюдений.

К сожалению оценка Ходжеса-Лемана не удовлетворяет естественному требованию аддитивности: $z_{jk} + z_{km} \neq z_{jm}$, в то время как сдвиги $\tau_{jk} + \tau_{km} = \tau_{jm}$. Поэтому рассматривают скорректированную оценку сдвига

$$\Delta_j = \frac{1}{n}\sum_{k=1}^{p} n_k z_{jk}, \quad j = \overline{1, p},$$

где $z_{kk} = 0$. Величина Δ_j отражает сдвиг j-ой группы относительно всех остальных групп.

С помощью этой оценки легко построить оценку любого сравнения. Для этого представим сравнение в эквивалентной форме:

$$\psi = \sum_{j=1}^{p} c_j \beta_j = \sum_{j=1}^{p} \sum_{k=1}^{p} \frac{c_j}{p} (\beta_j - \beta_k),$$

которое легко проверить непосредственно. Тогда оценка сравнения

$$\psi^* = \sum_{j=1}^{p} \sum_{k=1}^{p} \frac{c_j}{p} (\Delta_j - \Delta_k).$$

Пример. Рассмотрим задачу сравнения весов новорожденных поросят, используя критерий Краскела-Уоллиса. Расположив данные на странице 27 в порядке возрастания, поставим каждому наблюдению в соответствие его ранг с учетом повторов:

	6.5	52.5	48	42	23	37	23	18
	27.5	27.5	54.5	48	23	32	11.5	14.5
	48	42	23	42	32	37	11.5	35
	42	52.5	37	32	6.5	18	18	3
	56	13	42	48	6.5		1.5	
	54.5	14.5	48	18	9.5		1.5	
	32	6.5	32	23				
	18	4	51	27.5				
	27.5		42					
	9.5		42					
n_j	10	8	10	8	6	4	6	4
	321.5	212.5	419.5	280.5	100.5	124	67	70.5
r_j.	32.15	26.56	41.95	35.06	16.75	31	11.17	17.63

Статистика Краскела-Уоллиса равна $KW = 20.473$. По таблице распределения хи-квадрат с 7 степенями свободы находим $\alpha_{\text{кр}} = 0.0046$, то есть гипотезу однородности групп следует отвергнуть.

Модернизированная статистика Краскела-Уоллиса $KW' = 20.59$ практически не отличается от KW.

Критерий Джонкхиера к данной задаче, конечно, можно применить, однако в этом нет большого смысла, поскольку нет никаких оснований считать, что группы упорядочены по возрастанию среднего веса.

§ 4. Полный двухфакторный анализ

В предыдущем разделе были представлены методы обработки экспериментов, в которых отклик (результат эксперимента) зависел только от одной управляющей переменной. Предположим теперь, что уже два фактора A и B изменяются в эсперименте, и наша задача состоит в оценке влияния этих факторов на отклик.

Пусть фактор A разбит на p_1 уровней, а фактор B на p_2 уровней. Данные статистического эксперимента, таким образом, распадаются в матрицу с p_1 строками и p_2 столбцами. Наблюдение, попавшее в $(i,j)-$ую ячейку этой матрицы, обозначим через y_{ij}. Разброс данных обусловлен как чисто случайными (неучтенными) факторами, так и влиянием на результат эксперимента рассматриваемых нами факторов. Обозначим через η_{ij} истинное среднее значение отклика в $(i,j)-$ой ячейке. Степень различия между этими средними указывает на степень влияния факторов на результат. Основным моментом излагаемой далее теории 2-факторного *Д*А является представление среднего η_{ij} в виде четырех слагаемых:

$$\left.\begin{array}{c} \eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \\ \sum\limits_{i=1}^{p_1} \alpha_i = 0, \quad \sum\limits_{j=1}^{p_2} \beta_j = 0, \quad \sum\limits_{i=1}^{p_1} \gamma_{ij} = 0, \quad \sum\limits_{j=1}^{p_2} \gamma_{ij} = 0, \end{array}\right\} \qquad \text{(III.5)}$$

Покажем справедливость этого представления и, заодно, дадим интерпретацию входящих в него слагаемых.

Обозначим через $A_i = \eta_{i\bullet}$ – истинное *среднее i-ого уровня фактора A* (– среднее в i-ой строке). Аналогично, через $B_j = \eta_{\bullet j}$ обозначим истинное *среднее j-ого уровня фактора B* (– среднее по j-ому столбцу).

Общее среднее значение ожидаемых результатов экспериментов по всем ячейкам

$$\mu = \eta_{\bullet\bullet} = \frac{1}{p_1 p_2} \sum_{ij} \eta_{ij} \;=\; \frac{1}{p_1} \sum_{i=1}^{p_1} A_i \;=\; \frac{1}{p_2} \sum_{j=1}^{p_2} B_j$$

называется *генеральным средним.*

Отличие среднего i-ого уровня фактора A от генерального среднего естественно взять в качестве меры влияния фактора A на отклик. Соответствующая разность $\alpha_i = A_i - \mu$ называется *главным эффектом i-ого уровня фактора* A. Аналогично, *главным эффектом j-ого уровня фактора* B называется отклонение $\beta_j = B_j - \mu$.

Очевидно, оба эти параметра удовлетворяют соотношениям (III.5).

Заметим далее, что главный эффект α_i есть среднее (по столбцам) величин $\eta_{ij} - B_j$: $\alpha_i = \sum_j (\eta_{ij} - B_j)/p_2$. Указанные величины можно трактовать как эффекты уровней A по отношению к j-ому уровню B. Поэтому превышение этой величины над своим средним естественно назвать *взаимодействием i-ого уровня A с j-ым уровнем B*:

$$\gamma_{ij} = \eta_{ij} - B_j - \alpha_i = \eta_{ij} - A_i - B_j + \mu.$$

Легко проверить, что взаимодействие удовлетворяет всем соотношениям из представления (III.5). В частности,

$$\eta_{ij} = A_i + B_j + \gamma_{ij} - \mu = \mu + (A_i - \mu) + (B_j - \mu) + \gamma_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

Замечание v. Следует сказать, что пять уравнений (III.5) единственным образом определяют компоненты $(\mu, \alpha_i, \beta_j, \gamma_{ij})$ по значениям истинных средних η_{ij}. В частности, количество линейно независимых параметров уравнения (III.5) равно $p = p_1 p_2$ – числу неизвестных средних η_{ij}.

Замечание vi. Случай $\gamma_{ij} = 0$ при всех i, j называется случаем аддитивности эффектов или отсутствия взаимодействия, так как, по определению, в этом случае истинное среднее

$$\eta_{ij} = \mu + \alpha_i + \beta_j \quad \forall i, j.$$

Количество линейно независимых параметров при отсутствии взаимодействия сокращается до $p_1 + p_2 - 1$.

Замечание vii. Легко понять, что если общее число наблюдений n не больше общего числа параметров p, то статистическими методами невозможно выявить влияние факторов на отклик. Такая ситуация возникает, например, при проведении экспериментов только с одним наблюдением в каждой ячейке. В этом случае общее число наблюдений $n = p$ и для проведения статистического анализа необходимы дополнительные предположения относительно вида представления (III.5). Обычно делается допущение, что отсутствует взаимодействие и рассматривается дисперсионный анализ с аддитивными эффектами.

Двухфакторный анализ с одним наблюдением в ячейке.

Итак, рассмотрим 2-факторный *ДА* с одним наблюдением в каждой ячейке и предположим, что взаимодействия отсутствуют. Тогда основные предположения могут быть записаны в виде

$$\Omega : \begin{cases} y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \\ \sum_i \alpha_i = \sum_j \beta_j = 0, \\ \{e_{ij}\} \text{ независимы и распределены } \mathcal{N}(0, \sigma^2). \end{cases}$$

Данные подобного рода можно представить в виде $\overrightarrow{y} = \mathbf{X}'\beta + \overrightarrow{\varepsilon}$, если в качестве матрицы $\mathbf{X}$ взять следующую матрицу:

	y_{11}	y_{12}	$\dots$	y_{1p_2}	y_{21}	$\dots$	y_{2p_2}	$\dots$	$y_{p_1 1}$	$y_{p_1 2}$	$\dots$	$y_{p_1 p_2}$
μ	1	1	$\dots$	1	1	$\dots$	1	$\dots$	1	1	$\dots$	1
α_1	1	1	$\dots$	1	0	$\dots$	0	$\dots$	0	0	$\dots$	0
$\vdots$			$\vdots$			$\vdots$		$\vdots$			$\vdots$	
α_{p_1}	0	0	$\dots$	0	0	$\dots$	0	$\dots$	1	1	$\dots$	1
β_1	1	0	$\dots$	0	1	$\dots$	0	$\dots$	1	0	$\dots$	0
$\vdots$			$\vdots$			$\vdots$		$\vdots$			$\vdots$	
β_{p_2}	0	0	$\dots$	1	0	$\dots$	1	$\dots$	0	0	$\dots$	1

Первая строка и первый столбец здесь приписаны для объяснения способа построения этой матрицы: чтобы получить y_{ij}, нужно умножить соответствующий столбец на первый столбец.

Таким образом, имеем:

1) общее число наблюдений $n = p_1 p_2$;

2) общее число параметров $p = p_1 + p_2 + 1$;

3) ранг задачи $r = \operatorname{rang} \mathbf{X} = p_1 + p_2 - 1$ (– первая строка матрицы $\mathbf{X}$ равна сумме p_2 строк β, а строка α_1 равна сумме всех строк β минус сумма оставшихся строк α; остальные строки линейно независимы).

Т.к. $r < p$, то для нахождения ОМНК не применим способ, связанный с обращением информационной матрицы. Рассмотрим функцию

$$\mathcal{G} = \sum_{ij} (y_{ij} - \mu - \alpha_i - \beta_j)^2$$

и найдем ее производные по параметрам с учетом связей (III.5):

$$\mathcal{G}'_\mu = -2 \sum_{ij} (y_{ij} - \mu - \alpha_i - \beta_j) = -2(\sum_{ij} y_{ij} - p_1 p_2 \mu).$$

Следовательно, нормальное уравнение $\mathcal{G}'_\mu = 0$ дает ОМНК

$$\mu^* = \frac{1}{n} \sum_{ij} y_{ij} = y_{\cdot\cdot} \, .$$

Далее,

$$\mathcal{G}'_{\alpha_i} = -2 \sum_{j=1}^{p_2} (y_{ij} - \mu - \alpha_i - \beta_j) = -2(\sum_{j=1}^{p_2} y_{ij} - p_2 \mu - p_2 \alpha_i).$$

Отсюда, учитывая, что $\mu^* = y_{\cdot\cdot}$, получаем ОМНК

$$\alpha_i^* = \frac{1}{p_2} \sum_{j=1}^{p_2} y_{ij} - \mu^* = y_{i\cdot} - y_{\cdot\cdot} \, .$$

Аналогично, ОМНК

$$\beta_j^* = \frac{1}{p_1}\sum_{i=1}^{p_1} y_{ij} - \mu^* = y_{\cdot j} - y_{\cdot\cdot}\ .$$

Замечание viii. Несмотря на то, что для данной задачи ранг $r < p$, мы получили единственные решения нормальных уравнений. Это обусловлено тем, что здесь на параметры наложены два дополнительных условия и число линейно независимых параметров $p-2$ совпадает с рангом задачи.

Таким образом, сумма квадратов ошибок, которая в данном случае, в силу того, что мы не учитываем взаимодействия, иногда называется суммой квадратов взаимодействий, равна

$$SS_e\ = \sum_{ij}(y_{ij} - y_{i\cdot} - y_{\cdot j} + y_{\cdot\cdot})^2.$$

Её число степеней свободы равно

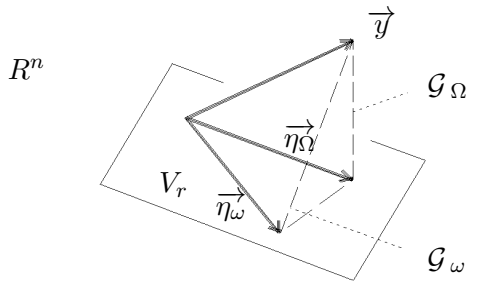
$$\nu_e = n - r = p_1p_2 - (p_1 + p_2 - 1) = (p_1 - 1)(p_2 - 1).$$

Проверка стандартных гипотез

В данной ситуации чаще всего рассматриваются гипотезы об отсутствии воздействия факторов на отклик

$$\mathbf{H}_A:\ \alpha_i = 0,\ \forall i, \quad \mathbf{H}_B:\ \beta_j = 0,\ \forall j.$$

Для нахождения ОМНК при гипотезе $\mathbf{H}_A$ необходимо минимизировать функцию $\mathcal{G} = \sum(y_{ij} - \mu - \beta_j)^2$. Непосредственными вычислениями легко показывается, что ОМНК $\mu_\omega^* = \mu^* = y_{\cdot\cdot}$ и $\beta_{j\omega}^* = \beta_j^* = y_{\cdot j}$. Следовательно, сумма квадратов расхождений $\mathcal{G}_\omega = \sum_{ij}(y_{ij} - y_{\cdot j} - y_{\cdot\cdot})^2$. Для вычисления суммы квадратов SS_H воспользуемся следующей геометрической иллюстрацией:



где, как обычно, $\overrightarrow{\eta_\Omega}$ и $\overrightarrow{\eta_\omega}$ – оценки регрессии при основных предположениях и предположениях гипотезы, представляющие собой векторы с компонентами $\mu^* + \alpha_i^* + \beta_j^*$ и $\mu^* + \beta_j^*$, соответственно. Отсюда легко видеть (в силу теоремы Пифагора), что $\mathcal{G}_\omega - \mathcal{G}_\Omega = \|\overrightarrow{\eta_\Omega} - \overrightarrow{\eta_\omega}\|^2$ и

$$SS_H\ = \mathcal{G}_\omega - \mathcal{G}_\Omega = \sum_{ij}(\mu^* + \alpha_i^* + \beta_j^* - \mu^* - \beta_j^*)^2 = \sum_{ij}(\alpha_i^*)^2 = p_2\sum_{i=1}^{p_1}(\alpha_i^*)^2.$$

Поскольку указанная сумма квадратов связана с гипотезой относительно фактора A, её обычно обозначают SS_A и называют суммой квадратов главных эффектов фактора A. Подставляя в это соотношение выражение для α_i^* и воспользовавшись утверждением леммы III.1, получаем

$$SS_A = p_2 \sum_i (y_{i\cdot} - y_{\cdot\cdot})^2 = p_2 \sum_i y_{i\cdot}^2 - p_1 p_2 y_{\cdot\cdot}^2 \ .$$

Хотя гипотеза $\mathbf{H}_A$ и задается p_1 параметрическими функциями, среди них всего $q = p_1 - 1$ линейно независимых, так как на них накладывается одна линейная связь. Поэтому критическая область F-критерия при проверке гипотезы $\mathbf{H}_A$ определяется неравенством

$$\mathcal{F}_A = \overline{SS_A}/\overline{SS_e} \ > \ F^{\alpha}_{p_1-1,\nu_e},$$

где $\overline{SS}_A = SS_A/(p_1 - 1)$.

Аналогично, F-критерий будет отвергать гипотезу $\mathbf{H}_B$, если

$$\mathcal{F}_B = \overline{SS}_B/\overline{SS_e} \ > \ F^{\alpha}_{p_2-1,\nu_e},$$

где $\overline{SS}_B = SS_B/(p_2-1)$ и $SS_B = p_1 \sum_j (y_{\cdot j} - y_{\cdot\cdot})^2 = p_1 \sum_j y_{\cdot j}^2 - p_1 p_2 y_{\cdot\cdot}^2$.

Разбиение полной суммы квадратов

Как и в однофакторном *ДА* здесь также полная сумма квадратов, символизирующая разброс всех данных, может быть разложена в сумму, компоненты которой обусловлены соответствующими факторами.

Теорема III.6. *Полная сумма квадратов*

$$SS_{\text{П}} \ = \sum_{ij} (y_{ij} - y_{\cdot\cdot})^2 = SS_A + SS_B + SS_e$$

Доказательство. Заметим, что $SS_e = \| \overrightarrow{y} - \overrightarrow{\eta_\Omega} \|^2 = \| \overrightarrow{y} \|^2 - \| \overrightarrow{\eta_\Omega} \|^2$. Следовательно,

$$\begin{aligned} SS_e \ &= \ \sum_{ij} y_{ij}^2 - \sum_{ij} (\mu^* + \alpha_i^* + \beta_j^*)^2 = \sum_{ij} y_{ij}^2 - \sum_{ij} \left[(\mu^*)^2 + (\alpha_i^*)^2 + (\beta_j^*)^2 \right] = \\ &= \sum_{ij} y_{ij}^2 - p_1 p_2 y_{\cdot\cdot}^2 - SS_A - SS_B, \end{aligned}$$

где удвоенные произведения в предпоследней сумме пропали в силу дополнительных условий $\sum \alpha_i^* = \sum \beta_j^* = 0$. $\otimes$

С помощью разложения полной суммы квадратов упрощается вычисление остаточной суммы квадратов $SS_e = SS_{\text{П}} - SS_A - SS_B$.

Соберем результаты в единую таблицу.

Таблица двухфакторного *ДА*

Источник разброса	SS	Ст.свободы	$\overline{SS}$	$\mathcal{F}$	α
Фактор A (строки)	SS_A	$\nu_A = p_1 - 1$	$\overline{SS}_A$	$\dfrac{\overline{SS}_A}{\overline{SS_e}}$	$1 - F_{\nu_A,\nu_e}(\mathcal{F})$
Фактор B (столбцы)	SS_B	$\nu_B = p_2 - 1$	$\overline{SS}_B$	$\dfrac{\overline{SS}_B}{\overline{SS_e}}$	$1 - F_{\nu_B,\nu_e}(\mathcal{F})$
Ошибки	SS_e	$\nu_e = (p_1 - 1)(p_2 - 1)$	$\overline{SS_e}$	–	–
Полная сумма квадратов	$SS_{\text{П}}$	$p_1 p_2 - 1$	–	–	–

Замечание ix. Предположение о равенстве нулю взаимодействий можно интерпретировать как добавление этих взаимодействий к ошибке. Поэтому в данном случае сумму квадратов ошибок иногда называют суммой квадратов взаимодействий SS_{AB}. Забегая вперед, отметим, что при полном уравнении регрессии формула для вычисления SS_{AB} и её число степеней свободы совпадают с формулами для SS_e при отсутствии взаимодействий.

Сравнения главных эффектов

Как и при 1-факторном *ДА* здесь можно уточнить вывод в случае, если гипотеза об отсутствии эффекта воздействия фактора на отклик отвергается. Равенство нулю всех главных эффектов фактора A эквивалентно условию постоянства средних значений всех уровней фактора: $A_i = \mu$. Поэтому для уточнения причин отвержения гипотезы $\mathbf{H}_A$ необходимо рассмотреть всевозможные сравнения средних уровней: $\psi = \sum c_i A_i$. МНК-оценка ψ равна $\psi^* = \sum c_i y_{i\bullet}$. Так как случайные величины $y_{i\bullet}\,(i = 1, \ldots, p_1)$ независимы, то дисперсия

$$\mathbf{D}(\psi^*) = \sum_{i=1}^{p_1} c_i^2\, \mathbf{D}(y_{i\bullet}) = \frac{\sigma^2}{p_2} \sum_{i=1}^{p_1} c_i^2. \qquad \text{(III.6)}$$

Для того, чтобы применить теперь теорему III.2 и формулу (III.4), достаточно заменить в формуле (III.6) σ^2 ее оценкой $\mathfrak{S}^2 = \overline{SS_e}$ и выбрать $q = p_1 - 1, n - r = \nu_e$.

Пример. Исследовалась поглощаемость 5-ти сортов масла при изготовлении пончиков. С этой целью в течение 6 рабочих дней на каждом из сортов масла приготавливалось в день по одной партии (24

штуки) пончиков. Были получены следующие результаты:

Сорт / День	1	2	3	4	5
1	164	172	163	150	164
2	177	197	177	172	169
3	168	167	144	146	145
4	146	161	165	141	149
5	172	180	166	169	170
6	196	190	178	183	167

Требуется с 10% уровнем значимости

1) проверить гипотезу об отсутствии различия между истинными средними количества поглощенного масла при использовании 5 сортов масла;

2) проверить гипотезу о неизменности среднего количества поглощенного масла по дням недели;

Вычисления иллюстрирует следующая таблица.

−6:	2:	−7:	−20:	−6:	↦	−37:	: 1369
: 36	: 4	: 49	: 400	: 36	↦	: 525:	
7:	27:	7:	2:	−1:	↦	42:	: 1764
: 49	: 729	: 49	: 4	: 1	↦	: 832:	
−2:	−3:	−26:	−24:	−25:	↦	−80:	: 6400
: 4	: 9	:676	: 576	: 625	↦	:1890:	
−24:	−9:	−5:	−29:	−21:	↦	−88:	: 7744
: 576	: 81	: 25	: 841	: 441	↦	:1964:	
2:	10:	−4:	−1:	0:	↦	7:	: 49
: 4	: 100	: 16	: 1	: 0	↦	: 121:	
26:	20:	8:	13:	−3:	↦	64:	: 4096
: 676	: 400	: 64	: 169	: 9	↦	:1318:	
↓ : ↓	↓ : ↓	↓ : ↓	↓ : ↓	↓ : ↓		↓ : ↓	: ↓
3:	47:	−27:	−59:	−56:	↦	−92:	:21422
:1345	:1323	:879	:1991	:1112	↦	:6650:	
9:	2209:	729:	3481:	3136:	↦	9564:	:

В эту таблицу сначала заносятся преобразованные данные (сдвинутые на 170) и их квадраты (по диагонали). После этого суммируются все 10 столбцов (по строкам) и 12 строк (по столбцам) и результаты записываются, соответственно, в правые 2 колонки и нижние 2 строки, которые также суммируются. Последний столбец и последняя строка содержат квадраты соответствующих сумм первых степеней по строкам и столбцам. Процесс суммирования для наглядности показан стрелками. Таким образом, получаем следующие результаты:

1. Число неизвестных параметров $p_1 = 6,\ p_2 = 5$.

2. Общий объем выборки $n = 30$.

3. Сумма всех значений $\sum_{ij} y_{ij} = -92$ и оценка генерального среднего

$$\mu^* = y_{\cdot\cdot} = -3.07 \ .$$

4. $ny_{\cdot\cdot}^2 = 282.13$.

5. Сумма всех квадратов $\sum_{ij} y_{ij}^2 = 6650$ и полная сумма квадратов

$$SS_{\text{П}} = 6650 - 282.13 = 6367.87 \ .$$

6. Сумма квадратов главных эффектов фактора A

$$SS_A = \sum_{i=1}^{p_1} \left(\sum_{j=1}^{p_2} y_{ij} \right)^2 - ny_{\cdot\cdot}^2 = 21422/5 - 282.13 = 4002.27 \ .$$

7. Сумма квадратов главных эффектов фактора B

$$SS_B = \sum_{j=1}^{p_2} \left(\sum_{i=1}^{p_1} y_{ij} \right)^2 - ny_{\cdot\cdot}^2 = 9564/6 - 282.13 = 1311.87 \ .$$

8. Сумма квадратов ошибок

$$SS_e = SS_{\text{П}} - SS_A - SS_B = 1053.73$$

Запишем результаты в таблицу.

Таблица двухфакторного *ДА*

Источник разброса	SS	Ст. свободы	$\overline{SS}$	$\mathcal{F}$	α	%
По дням	4002.27	5	800.454	15.2	< 0.0001	63%
По сортам	1311.87	4	327.97	6.23	0.002	20%
Ошибки	1053.73	20	52.67	-	-	17%
Полная сумма квадратов	6367.87	29	-	-	-	100%

Критический уровень значимости α найден по таблице распределения Фишера со степенями свободы (5,20) для первого фактора и (4,20) для второго. Таким образом, налицо значимое изменение поглощаемости масла как по сортам, так и по дням недели.

Последний столбец таблицы содержит долю соответствующей суммы квадратов в процентах к полной сумме квадратов. В соответствии с устоявшейся на практике терминологией можно утверждать, что на 60% изменчивость поглощаемости масла обусловлена разбросом по дням недели и на 20% влиянием сорта масла.

КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Кафедра математической статистики

ДИСПЕРСИОННЫЙ АНАЛИЗ

Методические разработки
по специальному курсу

Ч а с т ь 2

Многофакторный дисперсионный анализ.
Оптимальное планирование факторных
экспериментов

Казань – 1998

Двухфакторный дисперсионный анализ с одним наблюдением в ячейке с учетом взаимодействий

Обсудим теперь вопрос учета взаимодействий при проведении *ДА* с одним наблюдением в каждой ячейке. Как уже отмечалось, это возможно только при определенных предположениях относительно неизвестных параметров. Выше рассматривался случай полного отсутствия взаимодействий. Если все же мы хотим каким-либо образом учесть возможные взаимодействия при проведении *ДА*, нам необходимо построить новую модель регрессии с взаимодействиями, содержащими меньшее число параметров. Одна из таких моделей предполагает, что взаимодействия есть произведения главных эффектов

$$\gamma_{ij} = G\alpha_i\beta_j,$$

где G неизвестная константа. Таким образом, регрессия принимает вид $\eta_{ij} = \mu + \alpha_i + \beta_j + G\alpha_i\beta_j$, то есть является нелинейной по параметрам. Для этой регрессии задача нахождения ОМНК становится нетривиальной. Можно поступить следующим образом. Если предположить, что главные эффекты известны, тогда МНК-оценка для G легко находится и равна

$$G^* = \frac{\sum_{ij} \alpha_i\beta_j y_{ij}}{\sum_i \alpha_i^2 \sum\limits_j \beta_j^2}.$$

Теперь, положив формально $\alpha_i = \alpha_i^*, \beta_j = \beta_j^*$ и $\gamma_{ij}^* = G^*\alpha_i^*\beta_j^*$, можно рассмотреть формальную сумму квадратов взаимодействий

$$SS_{\text{взаим}} = \sum_{ij}(\gamma_{ij}^*)^2 = \frac{\left[\sum_{ij}\alpha_i^*\beta_j^* y_{ij}\right]^2}{\sum_i(\alpha_i^*)^2 \sum_j(\beta_j^*)^2} = \frac{p_1p_2K^2\left[\sum_{ij}\alpha_i^*\beta_j^* y_{ij}\right]^2}{SS_A\, SS_B}.$$

Основываясь на значениях $SS_{\text{взаим}}$, мы можем проверить гипотезу об отсутствии взаимодействий, отвергая гипотезу при больших $SS_{\text{взаим}}$.

Теорема III.7. *Если уравнение регрессии не содержит взаимодействий – $\eta_{ij} = \mu + \alpha_i + \beta_j$, тогда*

1) $SS_{взаим}/\sigma^2 \sim \chi_1^2$;

2) Остаточная сумма квадратов

$$SS_{ост} = \sum_{ij}(y_{ij} - y_{i\cdot} - y_{\cdot j} + y_{\cdot\cdot})^2 - SS_{взаим} = SS_e - SS_{взаим}$$

имеет хи-квадрат распределение с $p_1p_2 - p_1 - p_2$ степенями свободы.

3) $SS_{взаим}$ и $SS_{ост}$ независимы.

Эту теорему можно стандартным образом использовать для проверки гипотезы об отсутствии взаимодействий. Рассмотрим статистику

$$\mathcal{F} = \frac{SS_{\text{взаим}}}{\overline{SS}_{\text{ост}}},$$

которая, в силу предыдущей теоремы, имеет распределение Фишера с одной и $p_1p_2 - p_1 - p_2$ степенями свободы. Таким образом, гипотезу об отсутствии взаимодействий следует отвергать, если критический уровень значимости, вычисленный по распределению Фишера ($\alpha_{\text{кр.}} = 1 - F_{1,p_1p_2-p_1-p_2}(\mathcal{F})$) будет меньше заданного уровня α.

При подтверждении гипотезы дальнейший дисперсионный анализ проводится для модели регрессии без взаимодействий. Значительные сложности начинаются, если гипотеза об отсутствии взаимодействий отвергается. Мы не будем вдаваться в подробности статистического вывода в этом случае. Заметим только, что здесь приходится ограничиваться приближенным критерием.

Двухфакторный анализ с одинаковым числом наблюдений в ячейке

Пусть теперь в каждой ij-ой ячейке производилось по K_{ij} наблюдений (например, каждый день на каждом сорте масла изготавливалось несколько партий пончиков). Обозначим y_{ijk} – k-ое наблюдение в ij-ой ячейке. Тогда основные предположения записываются в виде

$$\Omega : \begin{cases} y_{ijk} = \eta_{ij} + e_{ij}, \\ \{e_{ij}\} \text{ независимы и распределены } \mathcal{N}(0, \sigma^2), \\ k = 1, \ldots, K_{ij};\ (ij) \in D, \end{cases}$$

где через D обозначено множество всех пар (– ячеек) (ij), для которых было произведено хотя бы одно наблюдение ($K_{ij} > 0$). Итак, мы имеем $n = \sum_{ij} K_{ij}$ наблюдений, а общее число неизвестных параметров (η_{ij}) равно количеству элементов множества D. Для нахождения ОМНК при Ω необходимо минимизировать по η_{ij} функцию

$$\mathcal{G} = \sum_{ij \in D} \sum_{k=1}^{K_{ij}} (y_{ijk} - \eta_{ij})^2.$$

Очевидно, здесь существует единственный минимум, и ОМНК $\eta^*_{ij} = y_{ij\bullet}$ для всех непустых ячеек (легко проверить дифференцированием по η_{ij}). Отсюда получаем, что, во-первых, в силу единственности ОМНК ранг задачи $r = p$ (см. следствие 1 из теоремы о существовании ОМНК) и, во-вторых, сумма квадратов ошибок

$$SS_e = \mathcal{G}_\Omega = \sum_{ij \in D} \sum_{k=1}^{K_{ij}} (y_{ijk} - y_{ij\bullet})^2,$$

а её число степеней свободы $\nu_e = n - p$.

Как отмечалось, все главные эффекты однозначно определяются значениями η_{ij}. Поэтому их МНК-оценки легко получить простой подстановкой ОМНК η^*_{ij} в уравнения для главных эффектов. Иными словами, МНК-оценки

$$\mu^* = \eta^*_{\cdot\cdot} = y_{\cdot\cdot\cdot},$$
$$\alpha^*_i = \eta^*_{i\cdot} - \eta^*_{\cdot\cdot} = y_{i\cdot\cdot} - y_{\cdot\cdot\cdot}, \quad \beta^*_j = \eta^*_{\cdot j} - \eta^*_{\cdot\cdot} = y_{\cdot j\cdot} - y_{\cdot\cdot\cdot},$$
$$\gamma^*_{ij} = \eta^*_{ij} - \eta^*_{i\cdot} - \eta^*_{\cdot j} + \eta^*_{\cdot\cdot} = y_{ij\cdot} - y_{i\cdot\cdot} - y_{\cdot j\cdot} + y_{\cdot\cdot\cdot} \ .$$

Проверка гипотез

Рассмотрим теперь задачу проверки стандартных гипотез *ДА*:

$\mathbf{H}_A$: все $\alpha_i = 0$ – фактор A не влияет на отклик;

$\mathbf{H}_B$: все $\beta_j = 0$ – фактор B не влияет на отклик;

$\mathbf{H}_{AB}$: все $\gamma_{ij} = 0$ – между факторами отсутствует взаимодействие.

Дальнейшие выкладки резко усложнятся, если не все K_{ij} одинаковы. Поэтому, с целью упрощения, предположим, что все $K_{ij} = K > 1$. В этом случае общее число наблюдений $n = Kp_1p_2$ при $p = p_1p_2$ параметрах. Число степеней свободы SS_e равно $\nu_e = n - p = p_1p_2(K-1)$.

Для каждой из гипотез необходимо минимизировать функцию

$$\mathcal{G} = \sum_{ijk} (y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2.$$

Сначала представим эту функцию в удобной для наших целей форме.

Лемма III.2. *Сумма квадратов расхождений*

$$\begin{aligned}\mathcal{G} \ = \ & SS_e + p_1p_2K(\mu^* - \mu)^2 + \\ & + p_2K\sum_i(\alpha^*_i - \alpha_i)^2 + p_1K\sum_j(\beta^*_j - \beta_j)^2 + K\sum_{ij}(\gamma^*_{ij} - \gamma_{ij})^2.\end{aligned}$$

Доказательство. Указанное разложение легко получается возведением в квадрат и суммированием следующего очевидного представления

$$\begin{aligned}y_{ij} - \mu - \alpha_i - \beta_j - \gamma_{ij} = & (y_{ij} - \mu^* - \alpha^*_i - \beta^*_j - \gamma^*_{ij}) + \\ & + (\mu^* - \mu) + (\alpha^*_i - \alpha_i) + (\beta^*_j - \beta_j) + (\gamma^*_{ij} - \gamma_{ij}).\end{aligned}$$

Как обычно, здесь удвоенные суммы квадратов пропадают в силу дополнительных соотношений для главных эффектов. $\otimes$

Чтобы найти теперь минимум $\mathcal{G}$ при гипотезе $\mathbf{H}_A$, необходимо положить все $\alpha_i = 0$ и минимизировать по оставшимся переменным с учетом дополнительных условий. Однако ясно, что этот минимум будет достигаться, если выбрать параметры так, чтобы соответствующие слагаемые в разложении $\mathcal{G}$ обратились в нуль, то есть положить $\mu = \mu^*$, $\beta_j = \beta_j^*$, $\gamma_{ij} = \gamma_{ij}^*$. При этом, сумма квадратов

$$\mathcal{G}_{\omega_A} = SS_e + p_2 K \sum_{i=1}^{p_1} (\alpha_i^*)^2 \qquad SS_A = p_2 K \sum_{i=1}^{p_1} (\alpha_i^*)^2 \ .$$

Гипотеза H_A описывается p_1 линейными функциями от неизвестных параметров, между которыми существует одна линейная связь. Поэтому число степеней свободы SS_A равно $p_1 - 1$. Аналогично, для проверки гипотезы $\mathbf{H}_B$ сумма квадратов ошибок

$$SS_B = p_1 K \sum_{j=1}^{p_2} \left(\beta_j^*\right)^2$$

с $p_2 - 1$ степенью свободы.

Для гипотезы $\mathbf{H}_{AB}$ сумма квадратов $SS_{AB} = K \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \left(\gamma_{ij}^*\right)^2$.

Число степеней свободы SS_{AB} легко найти, если расположить все функции, описывающие эту гипотезу, в матрицу $p_1 \cdot p_2$ и заметить, что в силу линейных связей на γ_{ij} первая строка и первый столбец этой матрицы линейно выражаются через элементы оставшейся $(p_1 - 1) \cdot (p_2 - 1)$ матрицы, которые между собой уже линейно независимы. Таким образом, число степеней свободы SS_{AB} равно $(p_1 - 1)(p_2 - 1)$. Приведем упрощенный способ проведения вычислений, необходимых для $\mathit{Д}A$.

Вычисления

Во-первых, наряду с таблицей данных полезно составить новую таблицу из средних значений каждой ячейки $y_{ij\bullet}$ и их квадратов. Эта новая таблица похожа на таблицу данных 2-факторного $\mathit{Д}A$ с одним наблюдением в ячейке. Сумма квадратов ошибок SS_e для таблицы средних, вычисленная по формуле для $\mathit{Д}A$ с одним наблюдением в ячейке, совпадает с суммой квадратов взаимодействий, вычисленной по формуле для $\mathit{Д}A$ с K наблюдениями в ячейке с точностью до множителя K (вспомним, что сумма квадратов ошибок при $\mathit{Д}A$ с одним наблюдением в ячейке называлась суммой квадратов взаимодействий). Так же, с точностью до множителя K, совпадают формулы для SS_A и SS_B при построении $\mathit{Д}A$ с одним наблюдением и с K наблюдениями в ячейке. Если суммы квадратов, относящиеся к таблице средних,

снабдить штрихом, то будет иметь место разбиение

$$SS_{\text{П}}{}' = \sum_{ij} y_{ij\bullet}^2 - p_1 p_2 y_{\dots}^2 = SS_A{}' + SS_B{}' + SS_e{}'.$$

Положим $SS_{\text{ячеек}} = K\, SS_{\text{П}}{}'$. Тогда, как отмечалось,

$$SS_A = K\, SS_A{}', \quad SS_B = K\, SS_B{}', \quad SS_{AB} = K\, SS_e{}'$$

и

$$SS_{AB} = SS_{\text{ячеек}} - SS_A - SS_B. \tag{III.7}$$

Для вычисления SS_e заметим, что полная сумма квадратов исходной таблицы $SS_{\text{П}} = SS_e + SS_{\text{ячеек}}$.

Теперь мы можем представить результаты в виде единой таблицы.

Таблица двухфакторного *ДА* с K наблюдениями в ячейке

Источник разброса	SS	Ст.свободы	$\overline{SS}$	$\mathcal{F}$	α
Фактор A (строки)	SS_A	$\nu_A = p_1 - 1$	$\overline{SS}_A$	$\dfrac{\overline{SS}_A}{\overline{SS_e}}$	$1 - F_{\nu_A,\nu_e}(\mathcal{F})$
Фактор B (столбцы)	SS_B	$\nu_B = p_2 - 1$	$\overline{SS}_B$	$\dfrac{\overline{SS}_B}{\overline{SS_e}}$	$1 - F_{\nu_B,\nu_e}(\mathcal{F})$
Взаимодействия	SS_{AB}	$\nu_{AB} = (p_1 - 1)(p_2 - 1)$	$\overline{SS}_{AB}$	$\dfrac{\overline{SS}_{AB}}{\overline{SS_e}}$	$1 - F_{\nu_{AB},\nu_e}(\mathcal{F})$
Ошибки	SS_e	$\nu_e = p_1 p_2 (K - 1)$	$\overline{SS_e}$	–	–
Полная сумма квадратов	$SS_{\text{П}}$	$Kp_1p_2 - 1$	–	–	–

Рекомендуемый процесс вычислений будет состоять из следующих шагов.

1) Вычисляются средние значения по каждой ячейке $y_{ij\bullet}$ и заносятся в новую таблицу вместе со своими квадратами $y_{ij\bullet}^2$.

2) По новой таблице вычисляются $y_{\dots}$, $SS_{\text{ячеек}}$, SS_A, SS_B.

3) Вычисляется SS_{AB} по формуле (III.7).

4) По исходной таблице находится сумма квадратов всех наблюдений и вычисляется $SS_{\text{П}}$.

5) Вычисляется $SS_e = SS_{\text{П}} - SS_{\text{ячеек}}$.

Попутно нами доказано стандартное разложение для полной суммы квадратов

$$SS_{\text{П}} = SS_A + SS_B + SS_e\,.$$

Пример 2-факторного дисперсионного анализа с 3 наблюдениями в ячейке

На фабрике по изготовлению мясных консервов наполнение банок происходит на 6 ротационных машинах. Партии мяса поступают от 5 поставщиков. Для проверки наполняемости банок от каждой партии и каждой машины были случайно отобраны 3 наполненных банки и взвешены. Результаты представлены в следующей таблице.

Поставщик / Цилиндр	1	2	3	4	5
1	501, 501, 502	504, 503, 505	506, 503, 507	503, 501, 503	501, 503, 503
2	499, 503, 499	498, 501, 500	501, 503, 505	502, 501, 500	501, 500, 501
3	501, 501, 501	502, 500, 501	502, 503, 504	501, 503, 503	503, 503, 503
4	498, 503, 500	498, 500, 501	503, 503, 504	500, 500, 502	500, 501, 501
5	501, 501, 499	502, 501, 505	500, 501, 502	501, 500, 499	498, 503, 501
6	501, 501, 500	500, 503, 500	503, 503, 504	503, 500, 502	501, 503, 502

Проверить с уровнем значимости 10% гипотезы о том, что

а) средний вес наполненной банки не зависит от поставщика;

б) средний вес наполненной банки не изменяется от машины к машине;

в) отсутствует взаимодействие между типом поставщика и ротационной машиной.

§ 5. Полный многофакторный анализ с взаимодействиями

В этом разделе на примере 3-факторного *ДА* мы распространим понятия, введенные для 2-факторного анализа, на случай многих факторов, влияющих на отклик.

Пусть на результат эксперимента действуют три фактора A, B и C, причем, фактор A разбит на p_A уровней, фактор B на p_B уровней, фактор C на p_C уровней. Первым шагом построения статистического вывода представим регрессию в виде суммы главных эффектов и взаимодействий. С целью унификации будем обозначать все компоненты регрессии буквой α с верхним индексом, указывающим на соответствующий фактор. Регрессия равна

$$\eta_{ijk} = \mu + \alpha_i^A + \alpha_j^B + \alpha_k^C + \alpha_{ij}^{AB} + \alpha_{ik}^{AC} + \alpha_{jk}^{BC} + \alpha_{ijk}^{ABC}.$$

Здесь

$$\mu = \eta_{...} \text{ — генеральное среднее;} \qquad \text{(III.8)}$$

$$\left.\begin{aligned} \alpha_i^A &= \eta_{i\bullet\bullet} - \eta_{\dots} \\ \alpha_j^B &= \eta_{\bullet j\bullet} - \eta_{\dots} \\ \alpha_k^C &= \eta_{\bullet\bullet k} - \eta_{\dots} \end{aligned}\right\} \begin{array}{l}\text{главные эффекты} \\ \text{факторов A, B, C}\,;\end{array} \tag{III.9}$$

$$\left.\begin{aligned} \alpha_{ij}^{AB} &= \eta_{ij\bullet} - \eta_{i\bullet\bullet} - \eta_{\bullet j\bullet} + \eta_{\dots} \\ \alpha_{ik}^{AC} &= \eta_{i\bullet k} - \eta_{i\bullet\bullet} - \eta_{\bullet\bullet k} + \eta_{\dots} \\ \alpha_{jk}^{BC} &= \eta_{\bullet jk} - \eta_{\bullet j\bullet} - \eta_{\bullet\bullet k} + \eta_{\dots} \end{aligned}\right\} \begin{array}{l}\text{двухфакторные} \\ \text{взаимодействия}\,;\end{array} \tag{III.10}$$

$$\alpha_{ijk}^{ABC} = \eta_{ijk} - \eta_{ij\bullet} - \eta_{i\bullet k} - \eta_{\bullet jk} + \eta_{i\bullet\bullet} + \eta_{\bullet j\bullet} + \eta_{\bullet\bullet k} - \eta_{\dots} \\ \text{— трехфакторное взаимодействие}\,. \tag{III.11}$$

Из этих формул легко вывести схему построения всех взаимодействий (на случай более чем трех факторов):

а) взаимодействие разбивается на группы слагаемых с чередующимися знаками “$+$“, “$-$“, $\dots$;

б) первая группа (со знаком “$+$“) содержит всего одно слагаемое, равное регрессии, усредненной по индексам, соответствующим факторам, не участвующим в рассматриваемом взаимодействии (например, $\eta_{ij\bullet}$);

в) вторая группа слагаемых (со знаком “$-$“) содержит усреднения регрессии из пункта а) по всем сочетаниям одного индекса;

г) третья группа (со знаком “$+$“) усреднения регрессии из пункта а) по двум индексам;

д) и т.д.

Предположим теперь, что при каждом сочетании уровней (в каждой ячейке) было произведено по M наблюдений и пусть y_{ijkm} есть m-ое наблюдение на i-ом уровне фактора A, j-ом уровне фактора B и k-ом уровне фактора C. Таким образом, всего получено $Mp_Ap_Bp_C$ наблюдений, по которым требуется оценить регрессию η_{ijk} с $p = p_Ap_Bp_C$ параметрами. ОМНК для параметров регрессии находятся стандартным способом с помощью подстановки в уравнения (III.8–III.11) оценки $\eta^*_{ijk} = y_{ijk\bullet}$. Например, ОМНК

$$\hat\alpha_{ij}^{AB} = y_{ij\bullet\bullet} - y_{i\dots} - y_{\bullet j\bullet} + y_{\dots\bullet}\,.$$

Суммы квадратов ошибок при проверке гипотез о главных эффектах и взаимодействиях легко вычислить, воспользовавшись следующей мнемонической формулой

$$SS = \sum_{ijkm} (\hat\alpha)^2\,.$$

Так, например,

$$SS_{AC} = \sum_{ijkm} \left(\hat{\alpha}_{ik}^{AC}\right)^2 = Mp_B \sum_{ik} \left(\hat{\alpha}_{ik}^{AC}\right)^2 .$$

Число степеней свободы суммы квадратов главных эффектов, как обычно, на единицу меньше числа уровней фактора (например, $p_A - 1$ для фактора A), а число степеней свободы суммы квадратов взаимодействий равно произведению степеней свободы факторов, участвующих в этом взаимодействии ($(p_B - 1)(p_C - 1)$ для SS_{BC}).

Сумма квадратов при числе наблюдений в ячейке $M > 1$ равна

$$SS_e = \sum_{ijkm} (y_{ijkm} - y_{....})^2 = \sum_{ijkm} y_{ijkm}^2 - ny_{....}^2$$

и её число степеней свободы равно $\nu_e = n - p = p_A p_B p_C (M - 1)$.

В случае $M = 1$ *ДА* возможен только при определенных предположениях относительно значений параметров регрессии. Сокращение числа параметров начинают, обычно, с предположения, что равны нулю все старшие взаимодействия: $\alpha_{ijk}^{ABC} = 0,\ \forall i, j, k$. Сумма квадратов ошибок в этом случае равна сумме квадратов самого старшего взаимодействия (– SS_{ABC} при 3-факторном *ДА*). Можно отбросить несколько групп взаимодействий. Тогда SS_e будет равна сумме SS всех отброшенных взаимодействий, а число степеней свободы SS_e – сумме соответствующих степеней свободы.

Разбиение полной суммы квадратов

Иногда *ДА* строят на формальном разбиении полной суммы квадратов, которое для 3-факторной модели принимает вид

$$\begin{aligned} SS_{\text{П}} &= \sum_{ijkm} y_{ijkm}^2 - ny_{....}^2 = \\ &= SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_{ABC} + SS_e . \end{aligned}$$

К этому разложению применяют следующую теорему Cochran'a.

Теорема III.8. *Пусть Q есть квадратичная форма полного ранга n от n независимых нормальных $\mathcal{N}(0,1)$ случайных величин, причем имеет место разложение Q на s квадратичных форм*

$$Q = Q_1 + \ldots + Q_s.$$

Обозначим через n_j ранг квадратичной формы Q_j. Тогда случайные величины $Q_1, \ldots, Q_s$ независимы и имеют χ^2-квадрат распределение с $n_1, \ldots, n_s$ степенями свободы, соответственно, в том и только в том случае, если $n = n_1 + \ldots + n_s$.

Предположим, что нам необходимо проверить гипотезу относительно взаимодействий факторов A и B. Проверим выполнимость условия теоремы Cochran'a при разложении полной суммы квадратов. Очевидно, сумма степеней свободы правой части (III.12) равна $(p_A p_B p_C M - 1)$ – степени свободы $SS_{\text{П}}$. Следовательно, отношение $\mathcal{F} = \overline{SS}_{AB} / \overline{SS}_e$ имеет распределение Фишера с $(p_A - 1)(p_B - 1)$ и $p_A p_B p_C (M - 1)$ степенями свободы, и гипотеза $\mathbf{H}_{AB} : \alpha_{ij}^{AB} = 0$ об отсутствии взаимодействий между A и B отвергается, если значение статистики $\mathcal{F}$ достаточно велико.

Пример. Провести анализ влияния на содержание влаги в пищевом продукте четырех факторов:

фактор A – сорт соли – 3 уровня;

фактор B – количество соли – 3 уровня;

фактор C – количество кислоты – 2 уровня;

фактор D – пищевые примеси – 2 уровня.

По результатам нижеприведенных измерений необходимо:

а) составить таблицу *ДА*. Полезно рассмотреть все $\overline{SS}$, даже если часть из них будет объединена с SS_e : относительная величина $\overline{SS}$ дает дополнительную информацию о влиянии факторов;

б) объединить SS всех трехфакторных и четырехфакторных взаимодейсвий с SS ошибок. Выявить значимость влияния оставшихся взаимодействий и главных эффектов на отклик.

Уровни A	Уровни B	Уровни C			
		1		2	
		Уровни D		Уровни D	
		1	2	1	2
1	1	8	5	8	4
	2	17	11	13	10
	3	22	16	20	15
2	1	7	3	10	5
	2	26	17	24	19
	3	34	32	34	29
3	1	10	5	9	4
	2	24	14	24	16
	3	39	33	36	34

Обработку этих данных произведем воспользовавшись возможностями пакета «STATGRAFICS». Данные занесем в файл «VLAGA»с переменной Y, содержащей количество влаги, и переменными SolSort, SolKol, Kislota и Primesi, содержащими номера уровней соответствующих факторов. Для примера приведем строку редактора данных *SG*,

в которую внесено количество влаги в пищевом продукте при использовании второго сорта соли, первого уровня количества соли, первого же уровня кислоты и второго уровня примесей:

Row	Y	SolSort	SolKol	Kislota	Primesi
13	3	2	1	1	2

Для проведения четырехфакторного $Д_A$ следует в разделе «Analysis of Variance» главного меню S_G выбрать подраздел «Multifactor Analysis of Variance» . Поясним некоторые графы, которые необходимо заполнить перед запуском программы на выполнение.

1) В графу «Data» заносится имя переменной, содержащей наблюдаемый в эксперименте отклик – «VLAGA.Y», в графу «Factors» управляемые факторы – «VLAGA.SolSort», «VLAGA.SolKol», «VLAGA.Kislota»,«VLAGA.Primesi».

2) Графа «Covariates» в данном случае остается пустой, поскольку должна содержать отсутствующие в нашем эксперименте непрерывные составляющие регрессии.

3) В графе «Interactions» (взаимодействия) необходимо убрать или оставить значки $*$ в зависимости от того, предполагается учет соответствующего двухфакторного взаимодействия или нет. Взаимодействия более высокого порядка в пакете S_G автоматически игнорируются. Для данной задачи все значки могут быть оставлены без изменения.

Выбрав метод Шеффе построения доверительных интервалов, произведем вычисления (клавиша **F6**). На экран будет выведена следующая таблица $Д_A$:

Source of var.	Sum of Sq.	d.f.	Mean sq.	F-ratio	Sig. level	%
MAIN EFFECTS	3633.83	6	605.63	304.93	.00	90.1
SolSort	495.05	2	247.52	124.62	.00	12.2
SolKol	2905.38	2	1452.69	731.42	.00	72.0
Kislota	3.36	1	3.36	1.69	.21	0.0
Primesi	230.02	1	230.02	115.81	.00	5.7
INTERACTIONS	364.69	13	28.053419	14.12	.00	9.0
SolSort SolKol	333.11	4	83.27	41.93	.00	8.2
SolSort Kislota	3.72	2	1.86	.93	.41	0.0
SolKol Kislota	6.05	2	3.02	1.52	.24	0.1
SolSort Primesi	4.05	2	2.02	1.02	.38	0.1
SolKol Primesi	14.38	2	7.19	3.62	.05	0.3
Kislota Primesi	3.36	1	3.36	1.69	.21	0.0
RESIDUAL	31.78	16	1.99			0.8
TOTAL (CORR.)	4030.31	35				100

Анализ предпоследнего столбца этой таблицы (Sig.Level – Уровень значимости) показывает, что все факторы, за исключением третьего, значимо влияют на содержание влаги. Из взаимодействий значимое отличие от нуля имеют только сочетания факторов AB (сорт соли – количество соли) и BD(количество соли – примеси). Исключив незначимые факторы и взаимодействия, получаем следующее уравнение регрессии

$$\mathbf{E}\, y_{ijkl} = \eta_{ijkl} = \mu + \alpha_i^A + \alpha_j^B + \alpha_l^D + \alpha_{ij}^{AB} + \alpha_{jl}^{BD}.$$

Последний столбец таблицы (в пакете $\mathcal{SG}$ не вычисляется) содержит долю в процентах суммы квадратов того или иного фактора в полной сумме квадратов. Этот показатель наглядно демонстрирует степень влияния фактора на отклик. Отсюда можно сделать вывод, что основное влияние на отклик оказывает количество соли.

Из дополнительных возможностей, вызываемых по клавише **F5** рассмотрим только опцию «Means Table», по которой выдается таблица средних значений с доверительными интервалами. Приведем первую страницу этой таблицы для регрессии, содержащей все факторы.

Level	Count	Average	Stnd. Err (internal)	Stnd.Err (pooled s)	95% Conf. for mean	
SolSort						
1	12	12.416	1.658	.406	11.554	13.279
2	12	19.833	3.267	.406	18.970	20.695
3	12	20.666	3.654	.406	19.804	21.529
SolKol						
1	12	6.500	.712	.406	5.637	7.362
2	12	17.916	1.588	.406	17.054	18.779
3	12	28.500	2.343	.406	27.637	29.362
Kislota						
1	18	17.944	2.652	.332	17.240	18.648
2	18	17.333	2.474	.332	16.628	18.037
Primesi						
1	18	20.167	2.478	.332	19.462	20.871
2	18	15.111	2.504	.332	14.406	15.815
Total	36	17.638	.2348	.2348	17.140	18.136

Из этой таблицы также видно, что главные эффекты уровней третьего фактора не отличаются между собой (их доверительные интервалы накладываются друг на друга). Значимое влияние первого фактора обеспечивается только эффектом первого уровня.

К сожалению, из этих таблиц нельзя установить значимость влияния взаимодействий высокого порядка. Для этого придется проводить дополнительные вычисления.

Глава IV

Модели со случайными факторами

Для моделей со случайными факторами (модель II $Д_A$) не существует общей теории, аналогичной вышеразвитой теории для моделей с постоянными факторами. Здесь мы рассмотрим два случая $Д_A$ со случайными компонентами – однофакторный и двухфакторный $Д_A$.

§ 1. Однофакторный дисперсионный анализ

Представим следующий производственный эксперимент. На контроль поступают образцы изделий, произведенные по одной и той же технологии (например, лопатки турбины, изготовленные в одной и той же печи). Над каждым образцом проводится по несколько контрольных замеров. Результат k-ого измерения i-ого образца можно представить в виде

$$y_{ik} = \theta_i + \varepsilon_{ik},$$

где θ_i – истинное значение измеряемой характеристики качества, случайно изменяющееся от образца к образцу, ε_{ik} – случайная ошибка измерения. Если определить эффект i-ого образца как разность между θ_i и теоретическим средним $\mu = \mathbf{E}\,\theta_i$ случайной величины θ_i : $\alpha_i = \theta_i - \mu$, то наша модель может быть переписана в стандартном для однофакторного $Д_A$ виде $y_{ik} = \mu + \alpha_i + \varepsilon_{ik}$ с единственным отличием, что эффекты случайны. Мы увидим далее, что таблица однофакторного $Д_A$ в модели II совпадает с таблицей однофакторного $Д_A$ в модели I, однако обоснования для этих таблиц разнятся.

Предположим, как обычно, что эффекты и ошибки распределены нормально и независимы. Кроме того, предположим, что всего было исследовано I образцов и на каждом образце было сделано по K замеров. В этом случае основные предположения принимают вид

$$\Omega : \begin{cases} y_{ik} = \mu + \alpha_i + \varepsilon_{ik},\ i = \overline{1,I}, k = \overline{1,K}, \\ \{\alpha_i\},\ \{\varepsilon_{ik}\}\ \text{— независимы в совокупности}, \\ \alpha_i \sim \mathcal{N}(0, \sigma_A^2), \qquad \varepsilon_{ik} \sim \mathcal{N}(0, \sigma_e^2). \end{cases}$$

Замечание i. Другое существенное отличие от модели I заключается в том, что измерения на одном образце зависимы. Действительно, при различных индексах j и k ковариация $\mathrm{Cov}(y_{ij}, y_{ik}) = \mathbf{E}(y_{ij}-\mu)(y_{ik}-\mu) = \mathbf{E}(\alpha_i+\varepsilon_{ij})(\alpha_i+\varepsilon_{ik}) = \mathbf{E}\,\alpha_i^2 = \sigma_A^2$, так как остальные математические ожидания равняются нулю в силу независимости рассматриваемых элементов и равенства нулю их средних значений.

Замечание ii. Так же другое значение, отличное от дисперсии ошибки, принимает дисперсия наблюдений

$$\mathbf{D}(y_{ik}) = \mathbf{D}(\alpha_i + \varepsilon_{ik}) = \mathbf{D}\,\alpha_i + \mathbf{D}\,\varepsilon_{ik} = \sigma_A^2 + \sigma_e^2.$$

Определение 1. Коэффициент корреляции между наблюдениями на одном образце

$$\rho = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}$$

называется *внутриклассовой корреляцией.*

Сумма квадратов ошибок в основных предположениях для этой модели находится так же, как для модели I:

$$SS_e = \sum_{ik}(y_{ik} - y_{i\cdot})^2.$$

Для того чтобы найти ее распределение, подставим модельные предположения из Ω в выражение для SS_e. Получим

$$SS_e = \sum_{ik}(\varepsilon_{ik} - \varepsilon_{i\cdot})^2. \qquad \text{(IV.1)}$$

Рассмотрим теперь фиктивную стандартную модель однофакторного *ДА* с постоянными факторами

$$\Omega' : \begin{cases} \varepsilon_{ik} = a_i + e_{ik},\ i = \overline{1,I}, k = \overline{1,K}, \\ \{e_{ik}\} \ \text{— независимы в совокупности}, \\ e_{ik} \sim \mathcal{N}(0, \sigma_e^2). \end{cases}$$

Для этой модели $\mathcal{G}_{\Omega'}$ совпадает с выше рассмотренной суммой квадратов (IV.1): $\mathcal{G}_{\Omega'} = SS_e$. Исходя из общей теории, можно утверждать, что $\mathcal{G}_{\Omega'}/\sigma_e^2 \sim \chi^2_{\nu_e}$, где $\nu_e = IK - I = I(K-1)$. Таким образом, сумма квадратов ошибок $SS_e \sim \sigma_e^2\chi^2_{\nu_e}$ и

$$\overline{SS_e} = \frac{1}{I(K-1)}\sum_{ik}(y_{ik} - y_{i\cdot})^2$$

есть несмещенная оценка дисперсии σ_e^2.

Перейдем теперь к рассмотрению вопроса различения гипотез. Для этой модели факторы случайны, поэтому гипотеза об отсутствии влияния фактора на отклик может быть сформулирована только в терминах дисперсии фактора: $\mathbf{H} : \sigma_A^2 = 0$. Сумма квадратов ошибок, обусловленная отклонением от гипотезы, как и в модели I однофакторного *ДА*, равна

$$SS_A = \sum_{ik} (y_{i\cdot} - y_{\cdot\cdot})^2 = K \sum_i (y_{i\cdot} - y_{\cdot\cdot})^2.$$

Для того чтобы найти её распределение, положим $g_i = \alpha_i + \varepsilon_{i\cdot}$. Очевидно, все g_i независимы и имеют нормальное распределение со средними ноль и дисперсией $\sigma_g^2 = \mathbf{D}\, g_i = \sigma_A^2 + \sigma_e^2/K$. Поэтому распределение

$$SS_A = K \sum_i (g_i - g_\cdot)^2 \;\sim\; K\,\sigma_g^2\,\chi_{I-1}^2 = (K\sigma_A^2 + \sigma_e^2)\chi_{I-1}^2.$$

Поскольку при гипотезе $\mathbf{H}$ средние SS_e и SS_A совпадают и равны σ_e^2, то для построения F-критерия осталось показать, что статистики SS_e и SS_A независимы. Этот факт сразу вытекает из стандартного разложения полной суммы квадратов и теоремы Cochran'a (III.8). Таким образом, если верна гипотеза $\mathbf{H}$, то статистика

$$\mathcal{F} = \frac{\overline{SS}_A}{\overline{SS_e}} = \frac{\overline{SS}_A/\sigma_e^2}{\overline{SS_e}\,/\sigma_e^2} \;\sim\; F_{I-1,K-1}$$

и критическая область $\mathcal{F} \geqslant F^{\alpha}_{I-1,K-1}$ имеет уровень α.

Статистику $\mathcal{F}$ можно использовать также для проверки гипотез более общего вида: $\mathbf{H}_{\theta_0} : \sigma_A^2 \leqslant_{(\geqslant)} \theta_0\sigma_e^2$, с заданной границей θ_0. Критическая область для $\mathbf{H}_{\theta_0}$ имеет вид $\mathcal{F} \geqslant C$, где критическая константа определяется из соотношения

$$\mathbf{P}\{\mathcal{F} \geqslant C\} \leqslant \alpha \tag{IV.2}$$

для всех параметров, определяющих модель и удовлетворяющих условиям гипотезы. Статистика

$$\mathcal{F} = \frac{\overline{SS}_A}{\overline{SS_e}} \;\sim\; \frac{K\sigma_A^2+\sigma_e^2}{\sigma_e^2}\frac{\chi^2_{I-1}/(I-1)}{\chi^2_{I(K-1)}/(I(K-1)} \;\sim\; (1+K\theta)\,F_{I-1,I(K-1)}$$

с параметром $\theta = \sigma_A^2/\sigma_e^2$. Очевидно, вероятность в (IV.2) монотонно возрастает с увеличением параметра θ. Поэтому для того, чтобы выполнялось неравенство (IV.2) при всех $\theta \leqslant \theta_0$, достаточно выбрать константу C из условия $\mathbf{P}\{F_{I-1,I(K-1)} \geqslant C/(1+K\theta_0)\} = \alpha$. Т.е.

$$C = (1+K\theta_0)F^{\alpha}_{I-1,I(K-1)}.$$

Оценки компонент дисперсии

Поскольку среднее значение хи-квадрат случайной величины равно числу ее степеней свободы, то среднее

$$\mathbf{E}\,\overline{SS_e} = \mathbf{E}\,\frac{\sigma_e^2\chi^2_{\nu_e}}{\nu_e} = \sigma_e^2$$

и, аналогично, $\mathbf{E}\,\overline{SS}_A = \sigma_e^2 + K\sigma_A^2$. Поэтому можно выбрать в качестве несмещенных оценок для компонент дисперсии σ_e^2 и σ_A^2 статистики

$$\hat\sigma_e^2 = \overline{SS_e}, \qquad \hat\sigma_A^2 = \frac{1}{K}(\overline{SS}_A - \overline{SS_e}).$$

Ясно, что с положительной вероятностью оценка $\hat\sigma_A^2$ может принимать отрицательные значения, в то время как сам параметр всегда неотрицателен. Иногда эту оценку видоизменяют, полагая ее равной некоторой положительной величине, когда оценка отрицательна. Такая модификация делает оценку смещенной, и изучение ее свойств значительно усложняется. Мы остановимся здесь только на исследовании состоятельности неизмененной оценки $\hat\sigma_A^2$.

Напомним, что дисперсия хи-квадрат случайной величины с m степенями свободы равна $2m$. Поэтому

$$\mathbf{D}\,\hat\sigma_e^2 = \mathbf{D}\,\frac{\sigma_e^2\chi^2_{\nu_e}}{\nu_e} = \frac{2\sigma_e^4}{\nu_e} = \frac{2\sigma_e^4}{I(K-1)}$$

и, в силу независимости $\overline{SS}_A$ и $\overline{SS_e}$,

$$\mathbf{D}\,\hat\sigma_A^2 = \frac{1}{K^2}(\mathbf{D}\,\overline{SS}_A + \mathbf{D}\,\overline{SS_e}) = \frac{2}{I-1}(\sigma_A^2 + \frac{1}{K}\sigma_e^2)^2 + \frac{2}{I(K-1)}\frac{\sigma_e^4}{K^2}.$$

Таким образом, дисперсия оценки $\hat\sigma_e^2$ стремится к нулю (а сама оценка будет состоятельной в среднеквадратическом) при $I \to \infty$ или $K \to \infty$, даже если одна из величин K, I остается конечной. По другому себя ведет дисперсия $\hat\sigma_A^2$. Очевидно, оценка $\hat\sigma_A^2$ будет состоятельной только в случае, когда число групп (образцов) $I \to \infty$.

§ 2. Полная классификация по двум признакам

Продолжая изучение производственного эксперимента, рассмотренного в начале предыдущего параграфа, предположим, что результат измерения образца зависит кроме всего прочего от степени настройки измерительного прибора. Можно предположить, что ошибка настройки прибора оказывает также случайное влияние на отклик. В этом

случае мы приходим к двухфакторной модели *ДА* со случайными эффектами. В рамках нормальных допущений основные предположения этой модели можно записать в виде

$$\Omega: \begin{cases} y_{ik} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = \overline{1,I},\ j = \overline{1,J},\ k = \overline{1,K}, \\ \{\alpha_i\},\ \{\beta_j\},\ \{\gamma_{ij}\},\ \{\varepsilon_{ik}\} \ \text{— независимы в совокупности}, \\ \alpha_i \sim \mathcal{N}(0,\sigma_A^2),\ \beta_j \sim \mathcal{N}(0,\sigma_B^2),\ \gamma_{ij} \sim \mathcal{N}(0,\sigma_{AB}^2), \\ \varepsilon_{ik} \sim \mathcal{N}(0,\sigma_e^2). \end{cases} \tag{IV.3}$$

Поскольку вид основных предположений этой модели ничем не отличается от вида основных предположений в модели I *ДА*, то все SS здесь вычисляются по тем же формулам, что и для модели I. Различия появляются при установлении их распределений. Рассмотрим, например, $SS_A = JK\sum(y_{i\cdot\cdot} - y_{\cdot\cdot\cdot})^2$. Подставляя в это соотношение модельные предположения IV.3 и вводя обозначение $g_i = \alpha_i + \gamma_{i\cdot} + \varepsilon_{i\cdot\cdot}$, получаем

$$SS_A = JK\sum_i(\alpha_i + \gamma_{i\cdot} + \varepsilon_{\mathbf{i}\cdot\cdot} - \alpha_{\cdot} - \gamma_{\cdot\cdot} - \varepsilon_{\cdot\cdot\cdot})^2 = JK\sum_i(g_i - g_{\cdot})^2.$$

Очевидно, все g_i имеют нормальное распределение со средним нуль и дисперсией $\mathbf{D}\,\alpha_i + \mathbf{D}\,\gamma_{i\cdot} + \mathbf{D}\,\varepsilon_{i\cdot\cdot} = \sigma_A^2 + \sigma_{AB}^2/J + \sigma_e^2/JK$. Следовательно,

$$SS_A \ \sim\ (JK\sigma_A^2 + K\sigma_{AB}^2 + \sigma_e^2)\chi_{I-1}^2.$$

Аналогично получаются распределения всех остальных SS :

$$\begin{aligned} SS_B &\sim (IK\sigma_B^2 + K\sigma_{AB}^2 + \sigma_e^2)\chi_{J-1}^2, \\ SS_{AB} &\sim (K\sigma_{AB}^2 + \sigma_e^2)\chi_{(I-1)(J-1)}^2, \\ SS_e &\sim \sigma_e^2\chi_{IJ(K-1)}^2. \end{aligned} \tag{IV.4}$$

Построим с помощью этих распределений критерии проверки стандартных гипотез.

Если верна гипотеза $\mathbf{H}_A : \sigma_A^2 = 0$, отношение $\overline{SS}_A/\overline{SS_e}$ уже не будет иметь распределение Фишера, поскольку математические ожидания рассматриваемых SS не равны. Очевидно, желаемое распределение Фишера здесь будет иметь отношение $\overline{SS}_A/\overline{SS}_{AB}$. Составляя аналогичные отношения для критериев проверки гипотез $\mathbf{H}_B : \sigma_B^2 = 0$ и $\mathbf{H}_{AB} : \sigma_{AB}^2 = 0$, получаем следующую таблицу *ДА*.

Таблица двухфакторного $Д_A$ модель II

Источник разброса	SS	Ст.свободы	$\overline{SS}$	$\mathcal{F}$	α
Фактор A (строки)	SS_A	$\nu_A = I-1$	$\overline{SS}_A$	$\dfrac{\overline{SS}_A}{\overline{SS}_{AB}}$	$1-F_{\nu_A,\nu_e}(\mathcal{F})$
Фактор B (столбцы)	SS_B	$\nu_B = J-1$	$\overline{SS}_B$	$\dfrac{\overline{SS}_B}{\overline{SS}_{AB}}$	$1-F_{\nu_B,\nu_e}(\mathcal{F})$
Взаимо-действия	SS_{AB}	$\nu_{AB} = (I-1)(J-1)$	$\overline{SS}_{AB}$	$\dfrac{\overline{SS}_{AB}}{\overline{SS_e}}$	$1-F_{\nu_{AB},\nu_e}(\mathcal{F})$
Ошибки	SS_e	$\nu_e = IJ(K-1)$	$\overline{SS_e}$	–	–
Полная сумма квадратов	$SS_{\text{П}}$	$KIJ-1$	–	–	–

Если $K=1$, то $Д_A$ может быть проведен только при дополнительных предположениях относительно уравнения регрессии. Стандартное предположение об отсутствии взаимодействий между факторами приводит здесь к равенству нулю дисперсии σ^2_{AB}. В этом случае в таблице $Д_A$ строка «ошибки» совпадает со строкой «взаимодействия» и $SS_{AB} = SS_e$.

Несмещенные оценки компонент дисперсии легко получаются из представленных выше распределений для SS :

$$\hat\sigma^2_A = \frac{1}{IK}(\overline{SS}_A - \overline{SS}_{AB}), \quad \hat\sigma^2_B = \frac{1}{JK}(\overline{SS}_B - \overline{SS}_{AB}),$$
$$\hat\sigma^2_{AB} = \frac{1}{K}(\overline{SS}_{AB} - \overline{SS_e}).$$

Так же, как и в модели II однофакторного $Д_A$, состоятельность этих оценок имеет место, только если $I\to\infty$ (для $\hat\sigma^2_A$), $J\to\infty$ (для $\hat\sigma^2_B$) и $J,I\to\infty$ (для $\hat\sigma^2_{AB}$).

Заметим, что если $K=1$, то компонента σ^2_{AB} отдельно не оценивается. В этом случае, однако, $Д_A$ проводится обычно в предположении, что $\sigma^2_{AB}=0$.

Полная классификация по трем признакам

Трехфакторный $Д_A$ в модели II (так же, как четырех, пяти и выше) уже существенно отличается от двухфакторного $Д_A$. Это различие обусловлено тем обстоятельством, что с ростом числа факторов увеличивается количество слагаемых, входящих в среднее значение для $\overline{SS}$ (IV.4). Например, для I групп фактора A, J групп фактора B, K групп фактора C и M наблюдений при каждом сочетании факторов

средние значения ошибок равны

$$\mathbf{E}\,\overline{SS}_A = \sigma_e^2 + JKM\sigma_A^2 + KM\sigma_{AB}^2 + JM\sigma_{AC}^2 + M\sigma_{ABC}^2\ ,$$

$$\mathbf{E}\,\overline{SS}_{AC} = \sigma_e^2 + JM\sigma_{AC}^2 + M\sigma_{ABC}^2\ .$$

Поэтому построение критерия проверки гипотезы $\mathbf{H}_A : \sigma_A^2 = 0$ возможно только при дополнительных предположениях относительно параметров модели. Например, если положить $\sigma_{AB}^2 = 0$, то такой критерий может быть основан на отношении $\overline{SS}_A/\overline{SS}_{AC}$, которое будет иметь в этом случае распределение Фишера. (Если предположить, что $\sigma_{AC}^2 = 0$, тогда распределение Фишера будет иметь статистика $\overline{SS}_A/\overline{SS}_{AB}$.) Однако, если мы не может предположить равенство нулю σ_{AB}^2 или σ_{AC}^2, тогда построение точного критерия становится невозможным. В этом случае применяют приближенный F-критерий, формулировку которого мы здесь опускаем.

Глава V

Факторные планы

Перейдем теперь к рассмотрению способов построения матриц плана $\mathbf{X}$, используемых при проведении факторных экспериментов. Разработка этих способов связана с необходимостью построения экономичных планов с достаточно простой процедурой вычислений.

§ 1. Дробные реплики полного факторного плана

Рассмотрим схему регрессионного эксперимента, в которой результаты измерений зависят от значений $m\,(>1)$ факторов $x_1,\ldots,x_m$. Фактор x_i может принимать конечное число значений (- уровней) $s_i\,(\geqslant 2)$, причем выбор сочетаний уровней для различных факторов в каждом из проводимых экспериментов полностью зависит от исследователя. Здесь мы ограничимся рассмотрением случая, когда все $s_i=2$. Такие *двухуровневые планы* имеют широкое распространение на практике в силу ряда обстоятельств. Во-первых, весьма типична ситуация, когда при проведении эксперимента рассматриваемые факторы могут либо присутствовать, либо отсутствовать, т.е. имеют только два значения. Во-вторых, если какой-либо фактор принимает значения на s уровнях, то можно вместо этого одного фактора рассмотреть s новых двухуровневых факторов, соответствующих уровням исходного фактора. В-третьих, как будет показано ниже, для некоторых типов регрессий наилучший план соответствует выбору из всех возможных уровней фактора только двух крайних.

В моделях с двухуровневыми факторами обычно входные переменные x_i принимают только два значения: 1, если i-ый фактор участвует в эксперименте, и -1, если фактор отсутствует. Функция регрессии

$$\begin{aligned}\eta(x_1,\ldots,x_m) \;=\; \beta_0+\sum_{i=1}^{m}\beta_i x_i+\sum_{i<j}\beta_{ij}x_ix_j+\sum_{i<j<l}\beta_{ijl}x_ix_jx_l+\ldots\\ +\sum_{i_1<i_2<\ldots<i_m}\beta_{i_1\ldots i_m}x_{i_1}\ldots x_{i_m}.\end{aligned}$$

Здесь, как обычно, параметр β_0 называется генеральным средним, параметры $\beta_1\,(i=\overline{1,m})$ главными эффектами факторов, β_{ij} – эффектами взаимодействий первого порядка или эффектами двухфакторных взаимодействий и, аналогично, $\beta_{i_1\ldots i_k}$ – эффектами взаимодействий $(k-1)$-ого порядка или эффектами k-факторных взаимодействий. Общее число неизвестных параметров, входящих в уравнение регрессии, равно, очевидно, $p=2^m$. Поэтому для нахождения всех оценок необходимо провести по крайней мере $n=2^m$ наблюдений (если предполагается проведение дисперсионного анализа относительно параметров модели, то число наблюдений должно быть больше 2^m). Понятно, что уже при $m>7$ такое требование не часто осуществимо на практике. С целью сокращения числа неизвестных параметров выдвигаются некоторые предположения относительно этих параметров. Обычно начинают со старших взаимодействий, которые отбрасываются, т.е. считаются равными нулю. Чаще всего используется два вида функций регрессии:

$$\eta(x_1,\ldots,x_m)=\beta_0+\beta_1x_1+\ldots+\beta_mx_m \tag{V.1}$$

и

$$\eta(x_1,\ldots,x_m)=\beta_0+\sum_{i=1}^{m}\beta_ix_i+\sum_{i<j}\beta_{ij}x_ix_j.$$

Определение 1. Факторный план, в соответствии с которым проводится по одному наблюдению для каждой комбинации уровней факторов, называется *полным* и обозначается 2^m.

Обозначение полного факторного плана связано с тем, что при этом плане необходимо провести ровно 2^m наблюдений. Для модели регрессии вида (V.1), которая зависит всего от $(m+1)$-ого параметра, такое число наблюдений избыточно. Однако также понятно, что большое число наблюдений позволяет оценить более богатую функцию регрессии. Эти соображения лежат в основе методики построения так называемых дробных реплик полного факторного плана.

Идею этого метода проиллюстрируем на примере трехфакторной регрессии. Сначала построим полный 2^2 план для первых двух факторов, а уровни третьего фактора будем выбирать, исходя из соотношения $x_3=x_1x_2$ (варианты – $x_3=-x_1x_2, x_3=x_1$ и т.д.). В результате получим план, задаваемый следующей матрицей

x_0	x_1	x_2	x_3
1	1	1	1
1	−1	1	−1
1	1	−1	−1
1	−1	−1	1

(V.2)

где первый столбец содержит значения аддитивного фактора (- коэффициент при β_0), который присутствует во всех экспериментах. Заметим, кроме того, что второй и третий столбцы образованы путем перебора всех 2^2 возможных сочетаний цифр -1 и 1.

Проведя формальный дисперсионный анализ по данным измерений с матрицей плана (V.2), можно найти оценки всех параметров линейной модели (V.1). Проанализируем теперь план (V.2) применительно к полной модели со всеми взаимодействиями. Матрица плана для полной модели выглядит следующим образом:

x_0	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
1	1	1	1	1	1	1	1
1	−1	1	−1	−1	1	−1	1
1	1	−1	−1	−1	−1	1	1
1	−1	−1	1	1	−1	−1	1

В этой матрице пятый столбец совпадает с четвертым, шестой – с третьим, седьмой – со вторым и восьмой – с первым. Таким образом, при анализе данных, основанных на наблюдениях с представленной матрицей плана, нельзя различить x_0 и $x_1x_2x_3$, x_1 и x_2x_3, x_2 и x_1x_3, x_3 и x_1x_2. Следовательно, если нет дополнительных предположений о параметрах полной модели, то могут быть найдены оценки не всех восьми неизвестных коэффициентов, а только четырех смешанных

$$\beta_0 + \beta_{123},\ \beta_1 + \beta_{23},\ \beta_2 + \beta_{12},\ \beta_3 + \beta_{12}.$$

Рассмотренный план называется *полурепликой* или *планом* 2^{3-1}. Для модели с общим числом факторов m дробные реплики строятся аналогичным образом. Сначала для выбранных (например, первых) $m-k$ факторов строится полный 2^{m-k} план, причем k выбирается так, чтобы $2^{m-k} \geqslant m+1$. Затем с помощью некоторых *генерирующих соотношений* выбираются уровни оставшихся k факторов.

Анализ построенного плана применительно к полной модели проводится с помощью *определяющих соотношений*, получаемых путем перемножения генерирующих соотношений так, чтобы в левой части равенства находилась единица, а в правой – какие-либо произведения факторов. Так для дробной реплики 2^{m-k} первые k определяющих соотношений получаются из генерирующих соотношений путем перемножения последних на их левые части и последующей замены величины x_i^2 на 1. Другие определяющие соотношения получаются перемножением ранее полученных и выделением из них новых. Знание всех определяющих соотношений позволяет найти всю систему совместных оценок без изучения матрицы плана. Для того чтобы определить, с

какими взаимодействиями смешано данное, нужно на него умножить обе части всех определяющих соотношений.

Рассмотрим пример с $m = 6$ факторами. Построим дробную реплику 2^{6-3} исходя из полного 2^3 факторного плана для первых трех факторов. Уровни четвертого, пятого и шестого факторов зададим с помощью генерирующих соотношений

$$x_4 = x_1x_2,\ x_5 = x_1x_3,\ x_6 = x_2x_3.$$

Первые три определяющих соотношения получаются умножением генерирующих соотношений на их левые части:

$$1 = x_1x_2x_4,\ 1 = x_1x_3x_5,\ 1 = x_2x_3x_6.$$

Попарное перемножение этих соотношений дает соотношения

$$1 = x_2x_3x_4x_5,\ 1 = x_1x_3x_4x_6,\ 1 = x_1x_2x_5x_6.$$

Перемножая по три равенства первые, получаем

$$1 = x_4x_5x_6.$$

Для того чтобы теперь, исходя из этих семи определяющих соотношений, определить, с какими эффектами взаимодействий смешан данный эффект, нужно умножить обе части всех определяющих соотношений на этот эффект. Например, умножая все определяющие соотношения на x_1, получаем

$$\begin{aligned} x_1 &= x_2x_4 = x_3x_5 = x_3x_4x_6 = x_2x_5x_6 = \\ &= x_1x_2x_3x_6 = x_1x_4x_5x_6 = x_1x_2x_3x_4x_5. \end{aligned}$$

Таким образом, главный эффект первого фактора β_1 смешан с двухфакторными взаимодействиями β_{24}, β_{35}, трехфакторными взаимодействиями β_{346}, β_{256}, четырехфакторными взаимодействиями β_{1236}, β_{1456}, и, наконец, с пятифакторным взаимодействием β_{12345}. Символически это записывается как

$$\beta_1^* \to \beta_1 + \beta_{24} + \beta_{35} + \beta_{346} + \beta_{256} + \beta_{1236} + \beta_{1456} + \beta_{12345} \qquad \text{(V.3)}$$

и означает, что оценка главного эффекта первого фактора, построенная по дробной реплике 2^{6-3} полного плана для модели регрессии без взаимодействий, является оценкой суммы параметров из (V.3), если на самом деле верна модель регрессии с взаимодействиями.

Удобство использования дробных реплик обусловлено простотой формул для ОМНК. Не вдаваясь в подробности доказательства, заметим, что информационная матрица такого плана равна $\mathcal{S} = n\mathbf{I}$ с

общим числом наблюдений $n = 2^{m-k}$. Оценки эффектов равны взвешенным средним

$$\beta_j = \frac{1}{n}\sum_{i=1}^{n} x_{ji} y_i,$$

где $(y_1, \ldots, y_n)'$ – вектор наблюдений, $(x_{j1}, \ldots, x_{jn})$ – строка матрицы плана $\mathbf{X}$, соответствующая выбранному фактору β_j.

§ 2. Латинские планы

Определение 2. *Латинским квадратом порядка* M называется такое расположение символов в виде квадратной таблицы $M \cdot M$, при котором каждый символ появляется по одному разу в каждой строке и в каждом столбце.

Существование латинских квадратов и один из способов их построения демонстрирует следующий пример

A	B	C	D
D	A	B	C
C	D	A	B
B	C	D	A

Латинские квадраты предназначены для проведения трехфакторного эксперимента, в котором каждый из факторов разбит на M уровней. Общее числом измерений $n = M^2$, при этом сочетание уровней факторов в каждом эксперименте соответствует сочетанию номера строки (уровень первого фактора) с номером столбца (уровень второго фактора) и символом (уровень третьего фактора).

В сравнении с полным трехфакторным экспериментом латинские планы требуют проведения значительно меньшего числа наблюдений (M^2 против M^3) при сохранении удобных формул для вычисления оценок и сумм квадратов. Естественно, за такое сокращение приходится платить уменьшением количества параметров модели. Латинские планы используются лишь для дисперсионного анализа моделей без взаимодействий

$$y_{ijk} = \mu + \alpha_i^A + \alpha_j^B + \alpha_k^C + \varepsilon_{ijk}.$$

В этом случае оценки μ и α_i^A равны

$$\hat{\mu} = y_{...} = \frac{1}{M^2}\sum_{(i,j,k)\in\Lambda} y_{ijk},$$

$$\hat{\alpha}_i^A = y_{i..} - y_{...} = \frac{1}{M}\sum_{(j,k)\in\Lambda_i} y_{ijk} - y_{...},$$

где суммирование производится по множеству Λ – всех M^2 упорядоченных троек (i,j,k), и множеству Λ_i – всех пар (j,k), для которых $(i,j,k) \in \Lambda$. Оценки главных эффектов двух других факторов получаются аналогично.

Выпишем теперь суммы квадратов ошибок. Остаточная сумма квадратов

$$SS_e = \sum_{(i,j,k)\in\Lambda} (y_{ijk} - \hat{\mu} - \hat{\alpha}_i^A - \hat{\alpha}_i^B - \hat{\alpha}_i^C)^2.$$

Сумма квадратов ошибок, обусловленная влиянием фактора A, равна

$$SS_A = M \sum_i y_{i\cdot\cdot}^2 - M^2 y_{\cdot\cdot\cdot}^2.$$

Как видим, формулы вполне стандартные, с единственным отличием, что суммирование при усреднениях производится по суженным множествам индексов.

Замечание i. Выбор того или иного латинского квадрата привносит элемент субъективности в схему проведения факторного эксперимента. Чтобы этого избежать, рекомендуют строить латинский план на основе квадрата, выбранного случайным образом из множества всех возможных латинских квадратов.

Латинские прямоугольники

Если исследуемые факторы имеют разное число уровней, то при планировании факторного эксперимента применяются *латинские прямоугольники*, которые можно определить как подмножество строк (или столбцов) латинского квадрата. Например,

E	D	G	B	C	A	F
D	C	F	A	B	G	E
A	G	C	E	F	D	B
F	E	A	C	D	B	G

.

Приведенный прямоугольник можно использовать для проведения трехфакторного эксперимента с факторами, разбитыми на 7,4 и 7 уровней. Кстати, этот $(M \cdot k)$-прямоугольник с $M = 7, k = 4$, обладает тем свойством, что каждая пара символов встречается в нем ровно в $\lambda = k(k-1)/(M-1) = 2$ столбцах. Такие прямоугольники называются *квадратами Юдена*.

Греко-латинские планы

Два латинских квадрата одинакового порядка называются ортогональными, если при наложении одного на другой каждая пара символов встречается только один раз. Например, два латинских квадрата

$$\begin{array}{ccc} A & B & C \\ C & A & B \\ B & C & A \end{array} \qquad \begin{array}{ccc} \alpha & \beta & \gamma \\ \beta & \gamma & \alpha \\ \gamma & \alpha & \beta \end{array}$$

при наложении образуют квадрат

$$\begin{array}{ccc} A\alpha & B\beta & C\gamma \\ C\beta & A\gamma & B\alpha \\ B\gamma & C\alpha & A\beta \end{array},$$

в котором каждая пара латинских и греческих букв встречается по одному разу.

Квадраты, получаемые после наложения друг на друга двух ортогональных латинских квадратов, называются *греко-латинскими*. Известно, что ортогональные латинские квадраты существуют для всех $M \neq 2, \neq 6$.

Эти квадраты используются для построения планов более чем трехфакторного *ДА*. Так при проведении l-факторного *ДА* необходимо построить $l-2$ попарно ортогональных латинских квадрата. При этом номер строки в каждом квадрате будет указывать на уровень первого фактора, номер столбца на уровень второго фактора, символы квадратов, стоящие на пересечении выбранной строки и выбранного столбца, на уровни третьего, четвертого и т.д. факторов. Следует заметить, что количество попарно ортогональных латинских квадрата не может превышать $M-1$.

§ 3. Блочные схемы

В этом разделе на примере двухфакторного *ДА* мы приведем общую схему построения планов, охватывающую уже рассмотренные дробные реплики и латинские планы.

Пусть 1-ый фактор имеет I уровней, 2-ой J уровней. Поставим в соответствие уровням 1-ого фактора элементы $a_1, \ldots, a_I$. Тогда любой план проведения экспериментов может быть описан набором множеств

$$\{B_1, \ldots, B_J\},$$

состоящих из элементов $a_1, \ldots, a_I$ и указывающих на сочетания уровней факторов в различных экспериментах. Например, если первый фактор имеет 3 уровня, а второй 2 уровня, то план, соответствующий схеме с $B_1 = (a_1, a_2, a_3)$ и $B_2 = (a_1, a_3)$, подразумевает проведение пяти экспериментов со следующими сочетаниями уровней первого и второго факторов – $(1;1), (2;1), (3;1), (1;2), (3,2)$, соответственно.

Определение 3. *Двумерной блок-схемой* называется структура, порожденная множеством элементов $A = \{a_1, \ldots, a_I\}$ и

множеством блоков $\mathbf{B} = \{B_1, \ldots, B_J\}$.

Введем некоторые обозначения:

k_j - число элементов в блоке B_j;

r_i - число блоков, содержащих элемент a_i;

λ_{lm} - число блоков, содержащих пару элементов (a_l, a_m).

Определение 4. Блок-схема называется

правильной, если все блоки имеют одинаковый размер: $k_j = k,\ \forall j$;

полной, если она правильная и $k = I$, то есть каждый блок содержит все элементы;

равноповторной, если каждый элемент встречается в равном числе блоков $r_i = r$;

симметричной, если она правильная, равноповторная и $I = J$ (то есть оба фактора имеют одинаковое число уровней);

сбалансированной, если она правильная, равноповторная и любая пара элементов принадлежит одному и тому же числу блоков: $\lambda_{lm} = \lambda$.

Основной интерес представляют неполные сбалансированные блок-схемы, так называемые BIB-планы. Эти планы обладают рядом оптимальных свойств и простотой проведения $Д4$. Дисперсии оценок главных эффектов факторов, построенных по наблюдениям, основанным на BIB-планах, равны между собой.

В следующем примере приводится симметричный BIB-план с параметрами $I = J = 4,\ k = r = 3,\ \lambda = 2$.

элементы	B_1	B_2	B_3	B_4
1	+		+	+
2	+	+		+
3	+	+	+	
4		+	+	+

Замечание ii. Не для всех сочетаний параметров существуют BIB-планы. Можно доказать, что параметры BIB-плана должны удовлетворять следующим соотношениям:

а) $kJ = rI$;

б) $r(k-1) = \lambda(I-1)$;

в) $I \leqslant J$;

г) $k \leqslant r$;

д) если $I = J$ – четное число, то $k - \lambda$ есть точный квадрат.

Замечание iii. Понятие двумерной блок-схемы обобщается на многомерный случай для построения планов многофакторного эксперимента. Примером трехмерной блок-схемы является латинский квадрат, а четырехмерной блок-схемы – греко-латинский квадрат.

Глава VI

Оптимальные планы регрессионных экспериментов

В этой главе мы рассмотрим задачу построения планов, оптимизирующих оценки параметров линейных регрессионных моделей. В предлагаемой ниже теории результат статистического эксперимента y зависит от некоторого k-мерного вектора входных переменных $x \in \mathcal{R}^k$:

$$y = \sum_{j=1}^{p} \beta_j f_j(x) + \varepsilon\,.$$

Множество значений $\mathcal{X}$, которые может принимать вектор x, называется *пространством планирования* и предполагается ограниченным и замкнутым. Функции $f_1, \ldots, f_p$ суть известные непрерывные функции на $\mathcal{X}$, называемые базисными функциями. Задача состоит в том, чтобы по результатам измерений $y_1, \ldots, y_n$ в n различных точках $x_1, \ldots, x_n$ оценить неизвестные параметры $\beta_1, \ldots, \beta_p$.

Примеры. 1) Ранее рассматривавшееся стандартное уравнение регрессии $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$ для задач дисперсионного анализа многофакторного эксперимента также может быть записано в представленном виде, если положить вектор входных переменных $x = (x_1, \ldots, x_p)$, а функции $f_i(x) = x_i$.

2) Часто возникает задача выявления зависимости отклика от некоторого «непрерывного» аргумента (например, продолжительность жизни микроорганизмов в зависимости от концентрации раствора в среде их обитания). Среди возможных форм функциональной зависимости чаще всего выбирают полиномиальную или тригонометрическую. Это обусловлено тем, что все практически полезные функции связи могут быть разложены в ряд Тейлора или в ряд Фурье. Отрезок такого ряда и выбирается в качестве приближенного описания реально существующего явления.

3) Рассмотрим задачу определения массы трех предметов на весах с двумя чашками. Если истинные веса этих предметов равны β_1, β_2, β_3, и отклонение стрелки весов в правую сторону интерпретируется как

вес со знаком плюс, а в левую сторону как вес со знаком минус, то показания весов можно представить в виде

$$y = \beta_1 x_2 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

где ε – ошибка измерения, обусловленная неучтенными факторами, а переменная x_i принимает значение -1, если при взвешивании i-ый предмет находился на левой чашке весов, $+1$, если на правой, и 0, если предмет отсутствовал в данном эксперименте по взвешиванию.

§ 1. Планы эксперимента и их информационная матрица

Запишем уравнение регрессии в матричной форме

$$\eta(x) = \sum_{j=1}^{p} \beta_j f_j(x) = \overrightarrow{\beta}' \overrightarrow{f}(x),$$

где $\overrightarrow{\beta} = (\beta_1, \ldots, \beta_p)'$ и $\overrightarrow{f}(x) = (f_1(x), \ldots, f_p(x))'$. Тогда вектор результатов статистического эксперимента представим в виде

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

где матрица плана

$$\mathbf{X} = \begin{pmatrix} f_1(x_1) & \ldots & f_1(x_n) \\ \vdots & \cdots & \vdots \\ f_p(x_1) & \ldots & f_p(x_n) \end{pmatrix} = \left(\overrightarrow{f}(x_1), \ldots, \overrightarrow{f}(x_n) \right).$$

Как было показано ранее, качество ОМНК зависит от матрицы $\mathbf{X}$:

$$\mathrm{Cov}(\overrightarrow{\beta^*}) = \sigma^2 (\mathcal{S}')^{-1} = \sigma^2 (\mathbf{X}\,\mathbf{X}')^{-1}, \qquad \text{(VI.1)}$$

то есть, в конечном счете, от выбора точек $x_1, \ldots, x_n$.

Определение 1. Совокупность точек $x_1, \ldots, x_n$ называется *планом эксперимента.*

Введем некоторые обозначения для планов эксперимента, задаваемых различными способами.

План эксперимента, в котором среди заданных точек $x_1, \ldots, x_n$ имеется только N различных $x_1, \ldots, x_N$, называется *дискретным* и обозначается

$$\xi = \begin{Bmatrix} x_1, & \ldots, & x_N \\ p_1, & \ldots, & p_N \end{Bmatrix}, \qquad \text{(VI.2)}$$

где $p_i = r_i/n$ и r_i есть число повторений i-ой точки. Очевидно, $\sum p_i = 1$. Для этого плана информационная матрица

$$\mathcal{S}(\xi) = \mathbf{X}\,\mathbf{X}' = \left(\overrightarrow{f}(x_1), \ldots, \overrightarrow{f}(x_n)\right) \begin{pmatrix} \overrightarrow{f}'(x_1) \\ \vdots \\ \overrightarrow{f}'(x_n) \end{pmatrix} = \sum_{i=1}^{n} \overrightarrow{f}(x_i)\overrightarrow{f}'(x_i).$$

С учетом повторяющихся точек x_i получаем

$$\mathcal{S}(\xi) = \sum_{i=1}^{N} r_i \overrightarrow{f}(x_i)\overrightarrow{f}'(x_i) = n\sum_{i=1}^{N} p_i \overrightarrow{f}(x_i)\overrightarrow{f}'(x_i)\ . \qquad \text{(VI.3)}$$

Поскольку информационная матрица выступает в роли обобщенной меры качества плана, то более удобно рассматривать вместо самой матрицы $\mathcal{S}$ нормированную информационную матрицу $\mathcal{S}\,/n$. В этом случае сравнение планов автоматически производится при одинаковом числе наблюдений. В дальнейшем всегда под информационной матрицей будет подразумеваться нормированная информационная матрица.

Обобщением плана (VI.2) является план, в котором точки $x_1, \ldots, x_n$ выбираются случайным образом из совокупности точек $x_1, \ldots,\ x_N$ в соответствии с распределением $\{p_1, \ldots, p_N\}$. Иными словами, для определения точки x, в которой будет производиться наблюдение y, проводится дополнительный случайный эксперимент, с вероятностью p_i принимающий значение $x = x_i$. Такой план в литературе называется «непрерывным» или «приближенным» .

Обозначение непрерывного плана совпадает с (VI.2). Информационная матрица непрерывного плана вычисляется аналогично матрице детерминированного плана по формуле (VI.3). Если ввести случайную величину X с распределением, задаваемым таблицей (VI.2), то информационная матрица может быть записана через среднее значение X :

$$\mathcal{S}(\xi) = \mathbf{E}\,\overrightarrow{f}(X)\overrightarrow{f}'(X)\ . \qquad \text{(VI.4)}$$

Замечание i. Подчеркнем существенное различие между дискретными и непрерывными планами. Для дискретного плана вида (VI.2) с $p_i = r_i/n$ общее количество наблюдений ограничено числом n или кратным ему – $2n, 3n, \ldots$. Однако по тому же самому плану, трактуемому как непрерывный, общее число наблюдений может принимать любое значение.

Дальнейшее обобщение способа задания плана эксперимента состоит в том, что совокупность точек $x_1, \ldots, x_n$ не фиксируется, а выбор точки x для проведения наблюдения производится случайно из всего

пространства планирования $\mathcal{X}$ в соответствии с некоторым вероятностным распределением (вероятностной мерой) ξ на $\mathcal{X}$. Информационная матрица этого плана вычисляется аналогично (VI.4).

Как уже отмечалось, в качестве характеристики сравнения планов естественно выбрать матрицу, обратную информационной

$$\mathcal{D}(\xi) = (\mathcal{S}(\xi))^{-1}$$

(см. (VI.1)). Эту матрицу будем называть *дисперсионной*. Заметим сразу, что дисперсионная матрица определена только для невырожденных информационных матриц, то есть для планов полного ранга p. Следующее утверждение показывает, что такие планы обязаны иметь не менее p различных входных точек $x_1, \ldots, x_N$.

Лемма VI.1. *Если для плана ξ вида (VI.2) число точек $N < p$, то этот план вырожденный.*

Доказательство. Так как ранг суммы матриц не больше суммы рангов, то в силу (VI.3) ранг информационной матрицы

$$\operatorname{rang} \mathcal{S}(\xi) \leqslant \sum_{i=1}^{N} \operatorname{rang} \overrightarrow{f}(x_i) \overrightarrow{f}'(x_i).$$

Столбцы матрицы $\overrightarrow{f}(x) \overrightarrow{f}'(x)$ равны $(f_l(x)f_1(x), \ldots, f_l(x)f_p(x))'$, $l = \overline{1,p}$. Поэтому, если все функции $f_l(x) = 0$, то ранг этой матрицы равен 0. Если же $f_l(x) \neq 0$ хотя бы для одного l, например, $f_1(x) \neq 0$, то все столбцы матрицы $\overrightarrow{f}(x) \overrightarrow{f}'(x)$ будут пропорциональны первому. Следовательно, ранг матрицы $\overrightarrow{f}(x) \overrightarrow{f}'(x)$ не больше 1, а ранг информационной матрицы $\operatorname{rang} \mathcal{S}(\xi) \leqslant N < p$, то есть информационная матрица вырождена. $\otimes$

Примеры построения и сравнения планов

Пример 1. Линейная регрессия на отрезке:

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad x \in [-1; 1].$$

Для задачи определения функциональной зависимости одной переменной от другой область изменения управляемой переменной обычно представляет собой ограниченный замкнутый интервал $\mathcal{R}^1$. Любой такой интервал может быть линейно преобразован в интервал $[-1; 1]$.

Для этой задачи вектор-функция $\overrightarrow{f}(x) = (1, x)'$ и произведение

$$\overrightarrow{f}(x) \overrightarrow{f}'(x) = \begin{bmatrix} 1 & x \\ x & x^2 \end{bmatrix}.$$

Следовательно, информационная матрица любого плана равна

$$\mathcal{S}(\xi) = \begin{bmatrix} 1 & \mathbf{E}\,X \\ \mathbf{E}\,X & \mathbf{E}\,X^2 \end{bmatrix}.$$

Рассмотрим сначала план

$$\xi_1 = \begin{Bmatrix} -1 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{Bmatrix}.$$

Для этого плана $\mathbf{E}\,X = 0,\ \mathbf{E}\,X^2 = 2/3,$ следовательно,

$$\mathcal{S}(\xi_1) = \begin{bmatrix} 1 & 0 \\ 0 & 2/3 \end{bmatrix} \quad \text{и} \quad \mathcal{D}(\xi_1) = \begin{bmatrix} 1 & 0 \\ 0 & 3/2 \end{bmatrix}.$$

Второй план сосредоточим в точках –1 и 1:

$$\xi_2 = \begin{Bmatrix} -1 & 1 \\ 1/2 & 1/2 \end{Bmatrix}.$$

Для этого плана $\mathbf{E}\,X = 0,\ \mathbf{E}\,X^2 = 1,$ поэтому

$$\mathcal{S}(\xi_2) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{и} \quad \mathcal{D}(\xi_2) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \leqslant \mathcal{D}(\xi_1).$$

Таким образом, можно утверждать, что если по обоим планам проводится одинаковое число наблюдений (например, 6,12 и т.д., если они дискретные, и произвольное число, если непрерывные), то план ξ_2 предпочтительнее плана ξ_1.

Пример 2. Квадратическая регрессия на отрезке $[-1; 1]$.

Рассмотрим уравнение регрессии вида $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ и связанные с ним два плана

$$\xi_1 = \begin{Bmatrix} -1 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{Bmatrix},\ \xi_2 = \begin{Bmatrix} -1 & 0 & 1 \\ 1/4 & 2/4 & 1/4 \end{Bmatrix}.$$

Для этой регрессии вектор-функция $\overrightarrow{f}\,(x) = (1, x, x^2)'$ и информационная матрица любого плана равна

$$\mathcal{S}(\xi) = \mathbf{E}\,\overrightarrow{f}\,(X)\,\overrightarrow{f}\,'(X) = \begin{bmatrix} 1 & \mathbf{E}\,X & \mathbf{E}\,X^2 \\ \mathbf{E}\,X & \mathbf{E}\,X^2 & \mathbf{E}\,X^3 \\ \mathbf{E}\,X^2 & \mathbf{E}\,X^3 & \mathbf{E}\,X^4 \end{bmatrix}.$$

Для плана ξ_1 средние $\mathbf{E}\,X = 0,\ \mathbf{E}\,X^2 = 2/3,\ \mathbf{E}\,X^3 = 0,\ \mathbf{E}\,X^4 = 2/3,$ следовательно, информационная и дисперсионная матрицы равны

$$\mathcal{S}(\xi_1) = \begin{bmatrix} 1 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 2/3 \end{bmatrix} \quad \text{и} \quad \mathcal{D}(\xi_1) = \begin{bmatrix} 3 & 0 & -3 \\ 0 & 3/2 & 0 \\ -3 & 0 & 9/2 \end{bmatrix}.$$

Аналогично, для плана ξ_2 $\mathbf{E}\,X = 0$, $\mathbf{E}\,X^2 = 1/2$, $\mathbf{E}\,X^3 = 0$, $\mathbf{E}\,X^4 = 1/2$, и

$$\mathcal{S}(\xi_2) = \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix} \qquad \text{и} \qquad \mathcal{D}(\xi_2) = \begin{bmatrix} 2 & 0 & -2 \\ 0 & 2 & 0 \\ -2 & 0 & 4 \end{bmatrix}.$$

Разность матриц

$$\mathcal{D}(\xi_1) - \mathcal{D}(\xi_2) = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -1/2 & 0 \\ -1 & 0 & 1/2 \end{bmatrix}$$

не будет, очевидно, ни положительной, ни отрицательной. Иными словами, планы ξ_1 и ξ_2 не сравнимы.

П р и м е р 3. Взвешивание трех предметов.

Стандартная методика состоит в поочередном взвешивании каждого из предметов на одной из чашек весов. Мы сейчас покажем, что такой «естественный» план дает очень плохие оценки искомых весов.

Уравнение регрессии, описывающее данный эксперимент, задается вектор-функцией $\overrightarrow{f}\,(x) = (x_1, x_2, x_3)'$ и, следовательно, произведение $\overrightarrow{f}\;\overrightarrow{f}\,'$ представляет собой матрицу, состоящую из всевозможных произведений $x_l x_m$. Для «естественного» плана, сосредоточенного в трех точках $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, отличны от нуля будут только диагональные элементы: $x_1x_1 = 1$ для точки $(1,0,0)$, $x_2x_2 = 1$ для точки $(0,1,0)$ и $x_3x_3 = 1$ для точки $(0,0,1)$. Поэтому информационная матрица этого плана

$$\mathcal{S}(\xi_1) = \frac{1}{3}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \frac{1}{3}\mathbf{I} \quad \text{и} \quad \mathcal{D}(\xi_1) = 3\mathbf{I}.$$

Рассмотрим теперь другой план ξ_2, сосредоточенный в восьми различных точках вида $(\pm1, \pm1, \pm1)$ с равными весами. Таким образом, по этому плану при общем числе наблюдений $n = 8$ будет проведено по одному взвешиванию сразу всех предметов при различных их сочетаниях на чашках весов. Для этого плана средние значения $\mathbf{E}\,X_l X_m = 0$ при различных l, m, так как произведение $x_l x_m$ в половине случаев принимает значение 1 и в половине случаев значение -1. Очевидно, среднее значение $\mathbf{E}\,X_l X_l = 1$. Следовательно, для плана ξ_2

$$\mathcal{S}(\xi_2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{I} \quad \text{и} \quad \mathcal{D}(\xi_1) = \mathbf{I},$$

то есть этот план в 3 раза эффективней «естественного» плана.

§ 2. Критерии оптимальности

Как видно из примера 2, не каждая пара планов будет иметь сравнимые дисперсионные матрицы. Следовательно, для определения понятия оптимального плана необходимо введение некоторой числовой характеристики, зависящей от дисперсионной матрицы. Общий подход здесь состоит в рассмотрении некоторого функционала Φ, заданного на множестве возможных дисперсионных матриц, и определении оптимального плана через соотношение

$$\xi^* = \arg\min_{\xi} \Phi[\mathcal{D}(\xi)].$$

Различные функционалы Φ определяют различные *критерии оптимальности.*

Выбор Φ зависит от целей проводимого эксперимента. Так, если по результатам статистического эксперимента необходимо оценить все параметры регрессии, которые представляют самостоятельный интерес (например, вес предметов), то естественно стараться уменьшить возможный разброс всего вектора получаемых оценок. Числовой характеристикой разброса векторной случайной величины является обобщенная дисперсия – определитель ковариационной матрицы.

Определение 2. План

$$\xi^* = \arg\min_{\xi} \| \mathcal{D}(\xi)\|$$

называется *D-оптимальным* (от английского «determinant»).

Определение 3. План

$$\xi^* = \arg\min_{\xi} \|A' \mathcal{D}(\xi) A\|$$

называется *обобщенно D-оптимальным.*

Здесь матрица A размера $p \cdot s\,(s \leqslant p)$ и полного ранга s служит для установления «предпочтения» одних характеристик вектора неизвестных параметров над другими. Например, если эксперимент управляется p факторами, а необходимо оценить влияние только первых s из них, то следует выбрать матрицу $A' = (\mathbf{I}, 0)$, где $\mathbf{I}$ – единичная матрица $s \cdot s$. В этом случае произведение $A'\,\mathcal{D}\,A$ равно главному минору порядка s матрицы $\mathcal{D}$ и в качестве оптимизируемой характеристики выступает разброс первых s компонент вектора $\overrightarrow{\beta^*}$.

Другой характеристикой разброса является сумма дисперсий оценок всех параметров регрессии – след дисперсионной матрицы, что приводит к следующему критерию оптимальности.

Определение 4. План

$$\xi^* = \arg\min_{\xi} \operatorname{tr} \mathcal{D}(\xi)$$

называется *A-оптимальным* (от английского «average» – среднее).

Следующий критерий оптимальности связан с задачей прогноза возможного значения отклика y при различных значениях входной переменной x (например, прогноз ожидаемого дохода от вложенных средств). Наилучший прогноз равен, очевидно, значению функции регрессии при выбранном значении входной переменной. По теореме Гаусса-Маркова оценка с минимальной дисперсией для функции регрессии получается при подстановке в нее ОМНК параметров регрессии: $\eta^*(x) = \overrightarrow{\beta^*}' \overrightarrow{f}(x)$. Дисперсия этой оценки равна

$$\mathbf{D}\, \overrightarrow{\beta^*}' \overrightarrow{f}(x) = \overrightarrow{f}'(x) \operatorname{Cov}(\overrightarrow{\beta^*}) \overrightarrow{f}(x) = \frac{\sigma^2}{n} \overrightarrow{f}'(x)\, \mathcal{D}(\xi) \overrightarrow{f}(x). \qquad \text{(VI.5)}$$

Определение 5. Функция $d(x;\xi) = \overrightarrow{f}'(x)\, \mathcal{D}(\xi) \overrightarrow{f}(x)$ называется дисперсией оценки поверхности отклика.

Приведем одно важное свойство этой функции.

Теорема VI.1. *Пусть X - случайная величина, определяемая планом ξ с невырожденной информационной матрицей. Тогда справедливы соотношения*

$$\mathbf{E}\, d(X;\xi) = p, \qquad \text{(VI.6)}$$

$$\max_{x \in \mathcal{X}} d(x;\xi) \geqslant p, \qquad \text{(VI.7)}$$

причем, если во втором соотношении достигается знак равенства, то план ξ сосредоточен в точках максимума функции $d(x;\xi)$.

Доказательство. Заметим, что $\overrightarrow{f}'(x)\, \mathcal{D}(\xi) \overrightarrow{f}(x)$ есть скалярная величина, поэтому она равняется своему следу:

$$\overrightarrow{f}'(x)\, \mathcal{D}(\xi) \overrightarrow{f}(x) = \operatorname{tr}\{ \overrightarrow{f}'(x)\, \mathcal{D}(\xi) \overrightarrow{f}(x)\} = \operatorname{tr}\{ \mathcal{D}(\xi) \overrightarrow{f}(x) \overrightarrow{f}'(x)\}$$

(по свойству следа, внутри которого можно переставлять любые пары матриц, если такая перестановка алгебраически допустима). Следовательно, справедлива цепочка равенств

$$\begin{aligned} \mathbf{E}\, d(x;\xi) &= \mathbf{E} \operatorname{tr}\left\{ \mathcal{D}(\xi) \overrightarrow{f}(X) \overrightarrow{f}'(X) \right\} = \\ &= \operatorname{tr}\left\{ \mathcal{D}(\xi)\, \mathbf{E}\, \overrightarrow{f}(X) \overrightarrow{f}'(X) \right\} = \operatorname{tr}\{\mathcal{D}(\xi)\, \mathcal{S}(\xi)\} = \operatorname{tr}\{\mathbf{I}\} = p. \end{aligned}$$

Второе утверждение теоремы вытекает из первого. $\otimes$

Если исследователя интересует прогноз в конкретной точке $x = x_0$, то естественно постараться уменьшить дисперсию прогноза именно в этой точке. В связи с этим вводится следующий критерий оптимальности.

Определение 6. План

$$\xi^* = \arg\min_{\xi} d(x_0; \xi)$$

называется оптимальным для экстраполяции в точку x_0.

В том случае, когда планируется использование уравнения регрессии для прогноза во всех возможных точках $x \in \mathcal{X}$, предлагается несколько критериев оптимальности. Один из них основан на идее минимаксности.

Определение 7. План

$$\xi^* = \arg\min_{\xi}[\max_{x \in \mathcal{X}} d(x; \xi)]$$

называется G-оптимальным.

Другой критерий использует в качестве оптимизируемой характеристики взвешенное среднее значение дисперсии поверхности отклика.

Определение 8. План

$$\xi^* = \arg\min_{\xi} \int_{\mathcal{Z}} d(x; \xi)\omega(x)\mathrm{d}x$$

называется Q-оптимальным. Здесь $\omega(x)$ – неотрицательная функция на множестве $\mathcal{Z}$, задающая относительную важность точек $z \in \mathcal{Z}$. Множество $\mathcal{Z}$ состоит из точек, для которых предполагается проведение прогноза отклика.

Последние два критерия, а также критерий A-оптимальности являются частными случаями критерия L-оптимальности.

Определение 9. План

$$\xi^* = \arg\min_{\xi} \mathrm{tr}[L\, \mathcal{D}(\xi)]$$

называется L-оптимальным. Здесь L – фиксированная, неотрицательно определенная матрица $p \cdot p$.

Возможны, конечно, и другие определения критериев оптимальности, однако, как говорили классики, «нельзя объять необъятное».

§ 3. Теоремы эквивалентности

Не существует универсального способа построения оптимальных критериев, однако имеет место ряд утверждений, позволяющих проверить «подозрительные» планы на их возможную оптимальность. Эти утверждения называются теоремами эквивалентности.

Первая теорема принадлежит Киферу и Вольфовицу.

Теорема VI.2. *В классе всех непрерывных планов следующие утверждения эквивалентны:*

а) план ξ^* D*-оптимален;*

б) план ξ^* G*-оптимален;*

в) $\max\limits_{x\in\mathcal{X}} d(x;\xi^*) = p$.

Информационные матрицы всех планов, удовлетворяющих одному из трех указанных утверждений, совпадают между собой. Все эти планы сосредоточены в точках x_i, *для которых* $d(x_i;\xi^*) = p$.

Доказательство. Рассмотрим два плана ξ^* и ξ и положим план $\xi_a = (1-a)\xi^* + a\xi$ с некоторым $a \in (0;1)$. Справедлива

Лемма VI.2. *Обозначим через* X *случайную величину, определяемую планом* ξ. *Тогда, если план* ξ^* *невырожден, то*

$$\left.\frac{\partial \ln \| \mathcal{S}(\xi_a)\|}{\partial a}\right|_{a=0+} = \mathbf{E}\, d(X;\xi^*) - p. \tag{VI.8}$$

Сначала воспользуемся этой леммой для доказательства теоремы, а доказательство леммы оставим напоследок.

Пусть ξ^* – D-оптимальный план. Тогда $\| \mathcal{S}(\xi_a)\| \leqslant \| \mathcal{S}(\xi^*)\|$, то есть в точке $a = 0$ достигается максимум функции $\| \mathcal{S}(\xi_a)\|$, и поэтому

$$\left.\frac{\partial \ln \| \mathcal{S}(\xi_a)\|}{\partial a}\right|_{a=0+} \leqslant 0.$$

Выберем в качестве ξ план вида $\xi = \left\{ \begin{matrix} x \\ 1 \end{matrix} \right\}$, сосредоточенный в произвольной точке x пространства планирования $\mathcal{X}$. Для этого плана $\mathbf{E}\, d(X;\xi) = d(x;\xi)$ и, следовательно,

$$\left.\frac{\partial \ln \| \mathcal{S}(\xi_a)\|}{\partial a}\right|_{a=0+} = d(x;\xi^*) - p \leqslant 0.$$

Отсюда, с учетом неравенства (VI.7), получаем, что $\max d(x;\xi^*) = p$ и план ξ^* G-оптимален. Таким образом, доказано, что из D-оптимальности плана ξ^* следует его G-оптимальность и выполняется равенство в) теоремы. То, что план ξ^* сосредоточен в точках максимума функции дисперсии отклика, следует из теоремы VI.1.

Предположим теперь, что план ξ^* G-оптимален, но не является D-оптимальным. Так как, в силу только что доказанного свойства, для D-оптимального плана выполняется свойство в), то для G-оптимального плана ξ^* также имеет место равенство

$$\max_{x\in\mathcal{X}} d(x;\xi^*) = p. \tag{VI.9}$$

Выберем в определении плана ξ_a в качестве плана ξ какой-либо D-оптимальный план. Позже мы докажем, что функционал $-\ln\|\mathcal{S}(\xi)\|$ строго выпукл, то есть справедливо неравенство

$$\ln\|\mathcal{S}(\xi_a)\| \geqslant (1-a)\ln\|\mathcal{S}(\xi^*)\| + a\ln\|\mathcal{S}(\xi)\|.$$

Отсюда следует, что дифференциальное отношение

$$\frac{\ln\|\mathcal{S}(\xi_a)\| - \ln\|\mathcal{S}(\xi^*)\|}{a} \geqslant \ln\|\mathcal{S}(\xi)\| - \ln\|\mathcal{S}(\xi^*)\| > 0.$$

Устремляя a к нулю, получаем (с учетом равенства VI.8 леммы VI.2) строгое неравенство $\mathbf{E}\,d(X;\xi^*) > p$, которое противоречит VI.9. Следовательно, G-оптимальный план будет и D-оптимальным.

Покажем, что информационные матрицы всех D-оптимальных планов совпадают. Предположим от противного, что найдутся два D-оптимальных плана ξ_1 и ξ_2 с разными информационными матрицами. Образуем из них новый план посредством выпуклой комбинации: $\xi = (1-a)\xi_1 + a\xi_2$. Для этого плана в силу строгой выпуклости функционала $-\ln\|\mathcal{S}(\xi)\|$ логарифм определителя дисперсионной матрицы

$$\begin{aligned}
\ln\|\mathcal{D}(\xi)\| &= -\ln\|\mathcal{S}(\xi)\| < \\
&< -(1-a)\ln\|\mathcal{S}(\xi_1)\| - a\ln\|\mathcal{S}(\xi_2)\| = \\
&= (1-a)\ln\|\mathcal{D}(\xi_1)\| + a\ln\|\mathcal{D}(\xi_2)\| = \min_{\xi}\ln\|\mathcal{D}(\xi)\|.
\end{aligned}$$

То есть определитель дисперсионной матрицы плана ξ меньше наименьшего возможного значения, что, естественно, невозможно.

Для доказательства теоремы осталось показать строгую выпуклость выше рассмотренного функционала и доказать лемму VI.2. Известно, что для плотности многомерного нормального распределения

$\mathcal{N}_p(\overrightarrow{0}, A^{-1})$ с ковариационной матрицей A^{-1}

$$\frac{\|A\|^{1/2}}{\pi^{p/2}} \int\limits_{\mathcal{R}^p} \exp\{-\overrightarrow{x}' A \overrightarrow{x}\} d\overrightarrow{x} = 1.$$

Следовательно, для любых двух различных положительно определенных матриц A и B имеем

$$\frac{\pi^{p/2}}{\|(1-a)A + aB\|^{1/2}} = \int\limits_{\mathcal{R}^p} \exp\{-\overrightarrow{x}'((1-a)A + aB)\overrightarrow{x}\} d\overrightarrow{x} =$$

$$= \int\limits_{\mathcal{R}^p} (\exp\{-\overrightarrow{x}' A \overrightarrow{x}\})^{1-a} (\exp\{-\overrightarrow{x}' B \overrightarrow{x}\})^a d\overrightarrow{x} <$$

$$< \left(\int\limits_{\mathcal{R}^p} \exp\{-\overrightarrow{x}' A \overrightarrow{x}\} d\overrightarrow{x} \right)^{1-a} \times \left(\int\limits_{\mathcal{R}^p} \exp\{-\overrightarrow{x}' B \overrightarrow{x}\} d\overrightarrow{x} \right)^a =$$

$$= \frac{\pi^{p(1-a)/2} \pi^{pa/2}}{(\|A\|)^{(1-a)/2} (\|B\|)^{a/2}}$$

в силу известного неравенства Гёльдера. Отсюда, произведя сокращения и прологарифмировав обе части неравенства, получаем требуемую выпуклость.

Докажем теперь лемму VI.2. Обозначим через $s_{ij}(\xi)$ элементы информационной матрицы $\mathcal{S}(\xi)$, а через $s^{ij}(\xi)$ элементы матрицы, обратной к $\mathcal{S}(\xi)$. Тогда, как известно,

$$\| \mathcal{S}(\xi)\| = \sum_{j=1}^{p} (-1)^{i+j} \| \mathcal{S}_{ij}(\xi)\| s_{ij}(\xi),$$

$$s^{ij}(\xi) = (-1)^{i+j} \frac{\| \mathcal{S}_{ij}(\xi)\|}{\| \mathcal{S}(\xi)\|}, \ i, j = 1, \ldots, p,$$

где $\mathcal{S}_{ij}(\xi)$ – определитель матрицы, полученной из матрицы $\mathcal{S}(\xi)$ вычеркиванием i-ой строки и j-ого столбца. Итак, производная

$$\frac{\partial \| \mathcal{S}(\xi)\|}{\partial s_{ij}} = (-1)^{i+j} \| \mathcal{S}_{ij}(\xi)\|$$

и

$$\frac{1}{\| \mathcal{S}(\xi)\|} \frac{\partial \| \mathcal{S}(\xi)\|}{\partial s_{ij}} = s^{ij}(\xi).$$

Кроме того, так как $s_{ij}(\xi_a) = (1-a)s_{ij}(\xi^*) + a s_{ij}(\xi)$, то

$$\frac{\partial s_{ij}(\xi_a)}{\partial a} = s_{ij}(\xi) - s_{ij}(\xi^*).$$

Воспользовавшись теперь теоремой о дифференцировании сложной функции, получаем

$$\left.\frac{\partial \ln \| \mathcal{S}(\xi_a)\|}{\partial a}\right|_{a=0} = \frac{1}{\| \mathcal{S}(\xi_a)\|}\frac{\partial \| \mathcal{S}(\xi_a)\|}{\partial a} =$$

$$= \frac{1}{\| \mathcal{S}(\xi_a)\|}\sum_{ij}\frac{\partial \| \mathcal{S}(\xi_a)\|}{\partial s_{ij}}\frac{\partial s_{ij}(\xi_a)}{\partial a} = \sum_{ij} s^{ij}(\xi^*)[s_{ij}(\xi) - s_{ij}(\xi^*)] =$$

$$= \operatorname{tr} \mathcal{S}^{-1}(\xi^*)[\mathcal{S}(\xi) - \mathcal{S}(\xi^*)] = \operatorname{tr} \mathcal{S}^{-1}(\xi^*)\,\mathcal{S}(\xi) - \operatorname{tr} \mathbf{I}_p =$$

$$= \operatorname{tr} \mathcal{S}^{-1}(\xi^*)\,\mathbf{E}\,\overrightarrow{f}(X)\overrightarrow{f}'(X) - p = \mathbf{E}\,\overrightarrow{f}'(X)\,\mathcal{S}^{-1}(\xi^*)\overrightarrow{f}(X) - p =$$

$$= \mathbf{E}\, d(X;\xi^*) - p.$$

Лемма доказана. $\otimes$

Примеры применения теоремы Кифера-Вольфовица

Пример 1. Для случая линейной регрессии на отрезке $[-1;1]$ D-оптимальным, а соответственно и G-оптимальным является план

$$\xi^* = \left\{ \begin{array}{cc} -1 & 1 \\ 1/2 & 1/2 \end{array} \right\}.$$

Действительно, как мы уже видели, для этого плана дисперсионная матрица равна единичной $\mathcal{D}(\xi^*) = \mathbf{I}$. Поэтому дисперсия поверхности отклика равна $d(x;\xi^*) = \overrightarrow{f}'(x)\mathbf{I}\overrightarrow{f}(x) = (1,x)(1,x)' = 1 + x^2$.

Максимум этой функции равен 2 (= числу неизвестных параметров) и достигается в точках $-1, 1$. Следовательно, по теореме Кифера-Вольфовица план ξ^* D- и G-оптимален.

По плану ξ^* можно проводить любое число наблюдений, только если он интерпретируется как непрерывный. Для дискретного плана ξ^* число наблюдений может быть только четным. Покажем, что для дискретных планов с нечетным числом наблюдений, кратным 3, критерии D-оптимальности и G-оптимальности не совпадают. Рассмотрим план общего вида

$$\xi = \left\{ \begin{array}{ccc} x_1 & x_2 & x_3 \\ 1/3 & 1/3 & 1/3 \end{array} \right\},$$

где по крайней мере две точки x_1, x_2, x_3 различны (это необходимо, чтобы план был невырожденным). Информационная матрица этого плана равна

$$\mathcal{S}(\xi) = \left[\begin{array}{cc} 1 & \mathbf{E}\,x \\ \mathbf{E}\,x & \mathbf{E}\,x^2 \end{array} \right] = \frac{1}{3}\left[\begin{array}{cc} 3 & x_1 + x_2 + x_3 \\ x_1 + x_2 + x_3 & x_1^2 + x_2^2 + x_3^2 \end{array} \right],$$

а ее определитель $|\mathcal{S}(\xi)| = \frac{2}{9}(x_1^2 + x_2^2 + x_3^2 - x_1 - x_2 - x_3)$.

Если здесь существует D-оптимальный план, тогда на этом плане достигается максимум определителя информационной матрицы, то есть все частные производные $\mathcal{S}(\xi)$ по переменным x_i равны нулю, либо этот максимум достигается на крайних точках пространства планирования: $x_i = \pm 1$. Система уравнений для частных производных имеет вид

$$\begin{aligned} 2x_1 - x_2 - x_3 &= 0, \\ -x_1 + 2x_2 - x_3 &= 0, \\ -x_1 - x_2 + 2x_3 &= 0. \end{aligned}$$

Эта система вырождена и имеет бесконечно много решений $x_1 = x_2 = x_3 = c$, где c – произвольная константа. Для любого из этих решений план ξ вырожден, так как сосредоточен всего в одной точке при двух неизвестных параметрах. На крайних точках ± 1 существует всего два невырожденных плана

$$\xi_1 = \left\{ \begin{array}{cc} -1 & 1 \\ 1/3 & 2/3 \end{array} \right\} \qquad \text{и} \qquad \xi_2 = \left\{ \begin{array}{cc} -1 & 1 \\ 2/3 & 1/3 \end{array} \right\},$$

которые и будут D-оптимальными. Информационная и дисперсионная матрицы для плана ξ_1 (также как и для плана ξ_2) равны

$$\mathcal{S}(\xi_1) = \left[\begin{array}{cc} 1 & 1/3 \\ 1/3 & 1 \end{array} \right] \qquad \text{и} \qquad \mathcal{D}(\xi_1) = \frac{3}{8} \left[\begin{array}{cc} 3 & -1 \\ -1 & 3 \end{array} \right].$$

Легко видеть, что дисперсия поверхности отклика $d(x;\xi_1) = \frac{3}{8}(3 - 2x + 3x^2) = 1 + \frac{1}{8}(3x - 1)^2$. Максимум ее достигается при $x = -1$ и равен $\max d(x;\xi_1) = 3$, то есть планы ξ_1 и ξ_2 D-оптимальны.

Рассмотрим дискретный план ξ_3, сосредоточенный с равными весами в трех точках $-1, 0, 1$. Для этого плана дисперсия отклика $d(x;\xi_3) = 1 + 3x^2/2$ (докажите!), а ее максимум равен, очевидно, $5/2 < 3$. То есть план ξ_3 по критерию G-оптимальности лучше D-оптимальных планов ξ_1 и ξ_2.

Пример 3. Взвешивание трех предметов.

Для второго из рассматриваемых в этом примере планов дисперсионная матрица равна единичной $\mathcal{D}(\xi_2) = \mathbf{I}$. Следовательно, дисперсия поверхности отклика равна $d(x;\xi_2) = (x_1, x_2, x_3)\mathbf{I}(x_1, x_2, x_3)' = x_1^2 + x_2^2 + x_3^2$. Максимум дисперсии отклика равен, очевидно, 3 (– числу неизвестных параметров) и достигается на точках вида $(\pm 1, \pm 1, \pm 1)$. По теореме Кифера-Вольфовица план ξ_2 D-оптимален.

Теоремы эквивалентности для L-оптимальных планов

Имеет место следующая

Теорема VI.3. *Если существует L-оптимальный план, тогда следующие утверждения эквивалентны:*

а) план ξ^* L-оптимален;

б) максимум функции

$$\varphi(x;\xi^*) = \overrightarrow{f}'(x)\,\mathcal{D}(\xi)\,L\,\mathcal{D}(\xi)\overrightarrow{f}(x) \qquad \text{(VI.10)}$$

равен $\max\limits_{x\in\mathcal{X}} \varphi(x;\xi^*) = tr[L\,\mathcal{D}(\xi^*)]$.

План ξ^* сосредоточен в точках максимума функции $\varphi(x;\xi^*)$.

Замечание ii. Легко видеть, что для критерия A-оптимальности матрица $L = \mathbf{I}$, а функция (VI.10) равна

$$\varphi(x;\xi) = \overrightarrow{f}'(x)\,\mathcal{D}^2(\xi)\overrightarrow{f}(x).$$

Для критерия Q-оптимальности $L = \int\limits_Z \overrightarrow{f}(x)\,\overrightarrow{f}'(x)\omega(x)\mathrm{d}x$ и функция

$$\varphi(x;\xi) = \int\limits_Z (\overrightarrow{f}'(x)\,\mathcal{D}(\xi)\overrightarrow{f}(z))^2\omega(z)\mathrm{d}z.$$

Для критерия оптимальности при экстраполяции в точку функция

$$\varphi(x;\xi) = (\overrightarrow{f}'(x)\,\mathcal{D}(\xi)\overrightarrow{f}(x_0))^2.$$

В качестве примера покажем, что для квадратической регрессии A-оптимальным является план

$$\xi_2 = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 1/4 & 2/4 & 1/4 \end{array} \right\}.$$

Как мы уже видели, дисперсионная матрица этого плана равна

$$\mathcal{D}(\xi_2) = \left[\begin{array}{ccc} 2 & 0 & -2 \\ 0 & 2 & 0 \\ -2 & 0 & 4 \end{array} \right] \qquad \text{и} \qquad \mathcal{D}^2(\xi_2) = \left[\begin{array}{ccc} 8 & 0 & -12 \\ 0 & 4 & 0 \\ -12 & 0 & 20 \end{array} \right].$$

Следовательно, функция (VI.10) равна

$$\varphi(x;\xi) = (1, x, x^2)\,\mathcal{D}^2 \left(\begin{array}{c} 1 \\ x \\ x^2 \end{array} \right) = 8 - 20x^2(1-x^2).$$

Очевидно, что $\max \varphi(x;\xi_2) = 8 = \mathrm{tr}\,\mathcal{D}(\xi_2)$ и достигается в точках $-1, 0, 1$, откуда следует A-оптимальность плана ξ_2.

§ 4. D-оптимальные планы для полиномиальной и тригонометрической регрессии

Полиномиальная регрессия на отрезке $[-1;1]$

В следующей теореме дается способ построения D-оптимальных планов для полиномиальной регрессии вида $\eta(x) = \beta_0 + \beta_1 x + \ldots + \beta_m x^m$, $x \in [-1;1]$. Предварительно введем одно вспомогательное

Определение 10. Полиномом Лежандра степени n называется полином $P_n(x)$, определяемый соотношением

$$P_n(x) = \frac{1}{n!2^n}\frac{d^n}{dx^n}(x^2-1)^n.$$

Для первых трех степеней полином Лежандра равен

$$P_0(x) = 1,\ P_1(x) = x,\ P_2(x) = \frac{1}{2}(3x^2-1),\ P_3(x) = \frac{1}{2}(5x^3-3x).$$

Теорема VI.4. *Для полиномиальной регрессии степени m на отрезке $[-1;1]$ существует единственный непрерывный D-оптимальный план. Этот план сосредоточен с равными весами $1/(m+1)$ в $(m+1)$-ой точке, которые являются решениями уравнения*

$$(1-x^2)\frac{d}{dx}P_m(x) = 0,$$

где $P_m(x)$ – полином Лежандра степени m.

Пример 2. Для квадратичной регрессии степень $m = 2$, производная $P_2'(x) = 3x$, а уравнение $(1-x^2)3x = 0$ имеет три решения $x = -1, x = 0, x = 1$. Итак, для квадратичной регрессии непрерывный D-оптимальный план сосредоточен с весами $1/3$ в точках $-1, 0, 1$.

Тригонометрическая регрессия на отрезке $[0;2\pi]$

Другой важный частный случай уравнения регрессии возникает при аппроксимации отклика отрезком ряда Фурье, т.е.

$$\eta(x;\beta) = \beta_1 + \sum_{j=1}^{k}[\beta_{2j}\cos jx + \beta_{2j+1}\sin jx] \qquad \text{(VI.11)}$$

с $p = 2k+1$ неизвестными параметрами и пространством планирования $\mathcal{X} = [0, 2\pi]$. В следующей теореме приводится вид D-оптимальных планов для тригонометрической регрессии; в ней, в частности, фигурирует непрерывный план $\xi^* = \frac{1}{2\pi}dx$, выбирающий точки в соответствии с равномерным распределением на отрезке $[0, 2\pi]$.

Теорема VI.5. *Если уравнение регрессии имеет вид (VI.11), тогда равномерный непрерывный план $\xi^* = (1/2\pi)dx$ является D-оптимальным непрерывным планом. План ξ_N, сосредоточенный в $N \geqslant 2k+1$ точках*

$$x_i = \frac{i-1}{N}2\pi, \qquad i = \overline{1, N},$$

также будет D-оптимальным непрерывным планом.

§ 5. Оптимальные планы первого порядка

Предположим теперь, что отклик зависит от m переменных (факторов) и уравнение регрессии имеет вид

$$\eta(\overrightarrow{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m \tag{VI.12}$$

с входной точкой $\overrightarrow{x} = (x_1, \ldots, x_m)'$. Совокупность входных точек образует стандартную матрицу плана $\mathbf{X}'$, строки которой равны $(1, x_{1i}, \ldots, x_{mi}), i = \overline{1, n}$. Важную роль для задач с подобной регрессией играют ортогональные планы, для которых информационная, а, следовательно, и дисперсионная матрицы диагональны и невырождены. Для ортогональных планов выполнены соотношения

$$\sum\nolimits_{i=1}^{n} x_{ji}x_{li} = 0, \quad j, l = 1, \ldots, m,\ l \neq j,$$
$$\sum\nolimits_{i=1}^{n} x_{ji} = 0, \quad i = 1, \ldots, m.$$

Пространство планирования, как обычно, предполагается ограниченным замкнутым подмножеством пространства $\mathcal{R}^m$. Здесь удобно будет записать это пространство в виде шаров

$$\sum\nolimits_{i=1}^{n} x_{ji}^2 \leqslant C_j^2,\ j = 1, \ldots, m, \tag{VI.13}$$

где C_j заданы.

Теорема VI.6. *Пусть (VI.12) — модель регрессии с дисперсией ошибки σ^2. Тогда для дискретных планов с пространством планирования вида (VI.13) дисперсии ОМНК всех неизвестных параметров*

$$\mathbf{D}(\beta_j^*) \geqslant \frac{\sigma^2}{C_j^2},\ j = 0, \ldots, m, \tag{VI.14}$$

с $C_0^2 = n$. *Причем, если план ортогональный и в (VI.13) выполняется равенство, тогда и в (VI.14) достигается знак равенства, то есть этот план является* A*-оптимальным в классе дискретных планов с пространством планирования (VI.13).*

Пример 3. Для задачи взвешивания трех предметов x_j принимает три значения $-1, 0, 1$, поэтому точки планирования удовлетворяют соотношениям (VI.13) с $C_j^2 = n$. Следовательно, A-оптимальным будет любой ортогональный план, при котором в каждом взвешивании участвуют все три предмета. В частности, таковым будет план, сосредоточенный в 8-ми точках вида $(\pm 1, \pm 1, \pm 1)$. Также A-оптимальным будет план, построенный как дробная реплика 2^{3-1} полного факторного плана и требующий 4 взвешивания с расположением предметов $(-1, -1, 1)$, $(1, -1, -1)$, $(-1, 1, -1)$, $(1, 1, 1)$.

Построение D-оптимальных планов для регрессии (VI.12) возможно, если предположить, что пространство планирования совпадает с m-мерным кубом $\mathcal{X} = [-1; 1]^m$, т.е. все x_j принимают значения из интервала $[-1; 1]$. Имеет место следующая

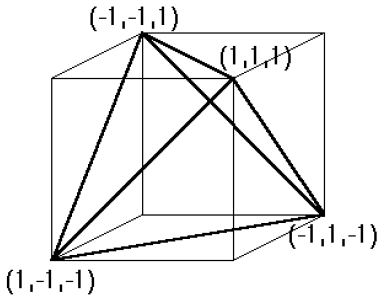
Теорема VI.7. *Пусть уравнение регрессии задается (VI.12) с пространством планирования* $\mathcal{X} = [-1; 1]^m$. *Тогда*

1) полный факторный план, сосредоточенный во всех 2^m *вершинах куба* $\mathcal{X}$, D*-оптимален;*

2) D*-оптимальным является также план, точки сосредоточения которого лежат в вершинах куба* $\mathcal{X}$ *и образуют правильный* m*-мерный* $(m+1)$*-угольник (симплекс-план).*

Замечание iii.

На плоскости при размерности $m = 2$ симплекс-планы не существуют. При $m = 3$ любой правильный тетраэдр, например, $(1, 1, 1), (1, -1, -1), (-1, 1, -1)$ и $(-1, -1, 1)$ образует 3-мерный симплекс.



Симплекс-планы, если они существуют, также A-оптимальны.

Литература

[1] Шеффе Г. Дисперсионный анализ. – М.: Наука, 1980. – 512 с.

[2] Ермаков С.М., Жиглявский А.А. Математическая теория оптимального эксперимента. – М.: Наука, 1987. – 320 с.

[3] Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.:ИНФРА-М, Финансы и статистика, 1995. – 384 с.

[4] Себер Дж. Линейный регрессионный анализ. – М.: Мир, 1980. – 456 с.