Из кн.: Кузнецов В.М. **Основы научных исследований в животноводстве.** Киров: Зональный НИИСХ Северо-Востока, 2006.- 568 с.

20. ДИСПЕРСИОННЫЙ АНАЛИЗ: МАТРИЧНЫЙ ПОДХОД

20.1. Система нормальных уравнений

Пусть имеются результаты наблюдений за 18 животными в 3-х группах или градациях фактора А (см. также [70,122,138,152]):

Градации	Значения наблюдений								Сумма	n_i	Среднее
A_1	3	5	6	2	-	-	-	-	16	4	4
A_2	5	6	2	7	8	3	9	8	48	8	6
A_3	7	6	4	3	6	4	-	-	30	6	5
Сумма									90	18	5,2222

Каждое наблюдение можно записать в виде биометрической модели:

$$y_{ij} = \mu + a_i + e_{ij},$$

$$\text{причем } i = 1,2,...,a\,(=3);\ j = 1,2,...,18;$$

где y_{ij} - наблюдение над j-ым животным в i-ой градации фактора А (может быть возраст, год, порода, отец, и т.д.); μ - общее среднее при равных частотах градаций фактора А; a_i - эффект i-ой градации фактора А, выраженный как отклонение от общего среднего μ; e_{ij} - случайная ошибка наблюдения (=остаток, эффект неучтенных факторов) для ij-го животного (предполагают, что e_{ij} имеет нормальное распределение со средней 0 и вариансой σ_e^2).

Элементы модели, справа от знака равенства, являются параметрами или константами. Оценки этих параметров обозначают знаком «^» над символом. Например, $\hat{\mu}$ является оценкой μ; $\hat{a}_i$ является оценкой a_i и $\hat{\sigma}_e^2$ является оценкой σ_e^2. Если a_i - это рандомизированные эффекты, то вариансу a_i (σ_a^2) оценивают значением $\hat{\sigma}_a^2$.

Каждое наблюдение можно записать в виде уравнения:

$$\begin{aligned}
3 = y_{11} &= \mu + a_1 &&+ e_{11}\\
5 = y_{12} &= \mu + a_1 &&+ e_{12}\\
6 = y_{13} &= \mu + a_1 &&+ e_{13}\\
2 = y_{14} &= \mu + a_1 &&+ e_{14}\\
5 = y_{21} &= \mu &+ a_2 &+ e_{21}\\
6 = y_{22} &= \mu &+ a_2 &+ e_{22}
\end{aligned}$$

$$
\begin{aligned}
2 = y_{23} &= \mu + a_2 + e_{23}\\
7 = y_{24} &= \mu + a_2 + e_{24}\\
8 = y_{25} &= \mu + a_2 + e_{25}\\
3 = y_{26} &= \mu + a_2 + e_{26}\\
9 = y_{27} &= \mu + a_2 + e_{27}\\
8 = y_{28} &= \mu + a_2 + e_{28}\\
7 = y_{31} &= \mu + a_3 + e_{31}\\
6 = y_{32} &= \mu + a_3 + e_{32}\\
4 = y_{33} &= \mu + a_3 + e_{33}\\
3 = y_{34} &= \mu + a_3 + e_{34}\\
6 = y_{35} &= \mu + a_3 + e_{35}\\
4 = y_{36} &= \mu + a_3 + e_{36}
\end{aligned}
$$

или в матричной записи:

$$y = Xb + e,$$

где y - вектор наблюдений; X - матрица из 0 и 1; b - вектор параметров; e - вектор случайных ошибок (=остатки).

Уравнение в «развернутом» виде:

$$
\underset{y}{\begin{bmatrix}3\\5\\6\\2\\5\\6\\2\\7\\8\\3\\9\\8\\7\\6\\4\\3\\6\\4\end{bmatrix}} = \begin{bmatrix}y_{11}\\y_{12}\\y_{13}\\y_{14}\\y_{21}\\y_{22}\\y_{23}\\y_{24}\\y_{25}\\y_{26}\\y_{27}\\y_{28}\\y_{31}\\y_{32}\\y_{33}\\y_{34}\\y_{35}\\y_{36}\end{bmatrix} = \underset{X}{\begin{bmatrix}1&1&0&0\\1&1&0&0\\1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&1&0\\1&0&1&0\\1&0&1&0\\1&0&1&0\\1&0&1&0\\1&0&1&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\\1&0&0&1\\1&0&0&1\\1&0&0&1\\1&0&0&1\end{bmatrix}} \times \underset{b}{\begin{bmatrix}\mu\\a_1\\a_2\\a_3\end{bmatrix}} + \underset{e}{\begin{bmatrix}e_{11}\\e_{12}\\e_{13}\\e_{14}\\e_{21}\\e_{22}\\e_{23}\\e_{24}\\e_{25}\\e_{26}\\e_{27}\\e_{28}\\e_{31}\\e_{32}\\e_{33}\\e_{34}\\e_{35}\\e_{36}\end{bmatrix}}.
$$

Матрицу X называют *инцидентной* (матрица плана). Она отражает структуру данных - показывает к какой градации (группе) относится соответствующее наблюдение.

Эффекты a_i оценивают LS-методом через минимизацию остаточной суммы квадратов (ошибки):

$$SS_e = \sum_{ij} e_{ij}^2 = \sum_{ij} (y_{ij} - \mu - a_i)^2$$

или в матричной записи

$$\sum_{ij} e_{ij}^2 = ee' = (y - Xb)'(y - Xb).$$

Выражение выводится из параметров μ и a_i, т.е. вектора b. Их оценки принимают за *константы*. Для минимизации ошибки частные производные приравнивают к 0, что приводит уравнение к виду

$$X'Xb = X'y.$$

Систему $X'Xb = X'y$ называют «*системой нормальных уравнений*». Она похожа на систему нормальных уравнений для множественной регрессии. Однако имеет некоторые существенные различия:

- независимые переменные не являются настоящими, а только сигнализируют о том, относится ли наблюдение к данной градации (1) или нет (0);
- параметры меняют свое содержание - они не соответствуют коэффициентам уравнения множественной регрессии.

В данном случае важны оценки, основанные на наблюдениях, по формуле:

$$X'X\hat{b} = X'y,$$

где $\hat{b}$ - вектор оценок, который получают из решения системы нормальных уравнений LS-методом.

20.2. Решение

Однозначное решение возможно только в том случае, если для матрицы $X'X$ существует обратная матрица $(X'X)^{-1}$, т.к.

$$\hat{b} = (X'X)^{-1} X'y.$$

Если $(X'X)^{-1}$ не существует, то возможно бесконечное число решений.

Система нормальных уравнений для однофакторной модели:

$$\begin{bmatrix} n & n_1 & n_2 & \dots & n_p \\ n_1 & n_1 & 0 & \dots & 0 \\ n_2 & 0 & n_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_p & 0 & 0 & \dots & n_p \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{matrix} RHM \\ \begin{bmatrix} \sum_i \sum_j y_{ij} \\ \sum_j y_{1j} \\ \sum_j y_{2j} \\ \vdots \\ \sum_j y_{pj} \end{bmatrix} \end{matrix}.$$

Каждой константе соответствует свое уравнение. Свойства нормальных уравнений:

- элементы (X'X) симметричны относительно главной диагонали;
- элементы в уравнении для $\hat{\mu}$ соответствуют диагональным элементам в уравнениях для $\hat{a}_i$;
- все недиагональные элементы как по строкам, так и по столбцам для $\hat{a}_i$ равны 0;
- сумма коэффициентов для $\hat{a}_i$ в уравнении для $\hat{\mu}$ равна коэффициенту для $\hat{\mu}$ в этом же уравнении;
- сумма членов правой части (RHM) уравнений для $\hat{a}_i$ равна члену правой части уравнения для $\hat{\mu}$.

Для числового примера имеем:

$$\begin{bmatrix} 18 & 4 & 8 & 6 \\ 4 & 4 & 0 & 0 \\ 8 & 0 & 8 & 0 \\ 6 & 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \end{bmatrix} = \begin{bmatrix} 94 \\ 16 \\ 48 \\ 30 \end{bmatrix}.$$

Значения коэффициентов первого столбца матрицы $X'X$ равно сумме значений коэффициентов по остальным столбцам. Аналогично по строкам. Таким образом, колонки и строки взаимозависимы. Поэтому матрица $X'X$ неполного ранга. Для нее не существует обратной матрицы $(X'X)^{-1}$. В данном случае ранг матрицы равен числу уравнений минус 1 (=p), т.е. существует только «p» независимых уравнений.

Так как число неизвестных параметров равно числу уравнений (=p+1), то чтобы решить систему нормальных уравнений необходимо ввести *допущение* (constraint). Обычное допущение при получении оценок a_i, как отклонение от μ, это $\sum \hat{a}_i = 0$, но возможны и $\hat{\mu} = 0$ или $\hat{a}_p = 0$.

Независимо от выбора допущения, в однофакторной модели суммы квадратов будут аналогичны. Решение не влияет на *остаточную* сумму квадратов. Так как в большинстве случаев остальные суммы квадратов получают через разность остаточных сумм квадратов двух моделей, то они также являются схожими (исключения возможны).

После выбора допущения, система нормальных уравнений *дополняется* еще одним уравнением или, что практичнее, нормальные уравнения *редуцируются* до независимого уровня.

Допущение $\sum \hat{a}_i = 0$ означает, что отдельные a_i - это отклонения от невзвешенного $\hat{\mu}$. Если при p=3 известно 2 элемента, то третий может быть рассчитан. Так, при известных $\hat{a}_1$ и $\hat{a}_2$,

$$\hat{a}_3 = -(\hat{a}_1 + \hat{a}_2).$$

Поэтому $\hat{a}_3$ и соответствующее уравнение можно исключить путем вычитания коэффициентов для $\hat{a}_3$ из остальных коэффициентов в рядах и столбцах (кроме ряда и колонки для μ). Ниже даны этапы решения *системы нормальных уравнений*.

Этап 1. Из коэффициентов всех рядов, кроме первого, вычитают коэффициенты для $\hat{a}_3$ ([6 0 0 6] см. стр. 340):

	$\hat{\mu}$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
$\hat{\mu}$	18	4	8	6
$\hat{a}_1$	-2	4	0	-6
$\hat{a}_2$	2	0	8	-6

Этап 2. Из всех столбцов, кроме первого, вычитают коэффициенты последнего столбца ([6 -6 -6]’ см. этап 1):

	$\hat{\mu}$	$\hat{a}_1$	$\hat{a}_2$
$\hat{\mu}$	18	-2	2
$\hat{a}_1$	-2	10	6
$\hat{a}_2$	2	6	14

В результате этих операций формируется *редуцированная* матрица X′X (субиндекс «r» означает «редуцированная»):

$$(X'X)_r = \begin{bmatrix} 18 & -2 & 2 \\ -2 & 10 & 6 \\ 2 & 6 & 14 \end{bmatrix}.$$

Подобным образом вычитают соответствующий элемент в векторе $X'y$:

$$(X'y)_r = \begin{bmatrix} 94 \\ -14 \\ 18 \end{bmatrix}.$$

Редуцированную матрицу $X'X$ можно получить также путем редукции матрицы X с последующим умножением на X'. При допущении $\sum \hat{a}_i = 0$, значения столбца a_3 в матрице X вычитают из столбцов a_1 и a_2. (Как известно, каждое наблюдение можно записать с помощью оценок параметров и остатка, например, $y_{31} = \hat{\mu} + \hat{a}_3 + \hat{e}_{31}$. При $\hat{a}_3 = -(\hat{a}_1 + \hat{a}_2)$ после редукции получим $y_{31} = \hat{\mu} - \hat{a}_1 - \hat{a}_2 + \hat{e}_{31}$.) Эта процедура легче, чем редуцировать матрицу $X'X$. Редуцированная матрица X:

$$X'_r = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}.$$

После принятия одного из допущения можно, таким образом, получить единственное решение:

$$\hat{b}_r = (X'X)_r^{-1} (X'y)_r.$$

Инверсия матрицы $(X'X)_r$:

$$\begin{bmatrix} 18 & -2 & 2 \\ -2 & 10 & 6 \\ 2 & 6 & 14 \end{bmatrix}^{-1} = \begin{bmatrix} 0{,}0602 & 0{,}0231 & -0{,}0185 \\ 0{,}0231 & 0{,}1435 & -0{,}0648 \\ -0{,}0185 & -0{,}0648 & 0{,}1019 \end{bmatrix}.$$

При допущении $\sum \hat{a}_i = 0$ решение будет (в рамочке даны элементы субматрицы Z; см. ниже):

$$\hat{b}_r = \begin{bmatrix} \hat{\mu} \\ \hat{a}_1 \\ \hat{a}_2 \end{bmatrix} = \begin{bmatrix} 0{,}0602 & 0{,}0231 & -0{,}0185 \\ 0{,}0231 & \boxed{\begin{matrix} 0{,}1435 & -0{,}0648 \\ -0{,}0648 & 0{,}1019 \end{matrix}} \\ -0{,}0185 & \end{bmatrix} \begin{bmatrix} 94 \\ -14 \\ 18 \end{bmatrix},$$

что дает:

$$\hat{\mu} = (0{,}0602)(94) + (0{,}0231)(-14) + (-0{,}0185)(18) = 5;$$
$$\hat{a}_1 = (0{,}0231)(94) + (0{,}1435)(-14) + (-0{,}0648)(18) = -1;$$
$$\hat{a}_2 = (-0{,}0185)(94) + (-0{,}0648)(-14) + (0{,}1019)(18) = +1.$$

Так как допускалось, что $\sum \hat{a}_i = \hat{a}_1 + \hat{a}_2 + \hat{a}_3 = 0$,

то $$\hat{a}_3 = -(\hat{a}_1 + \hat{a}_2) = -(-1+1) = 0.$$

20.3. Разложение суммы квадратов

В однофакторном дисперсионном анализе общую сумму квадратов (SS_y) раскладывают на сумму квадратов модели, т.е. между градациями фактора (SS_a) и на остаточную сумму квадратов (SS_e).

Для сбалансированных данных существует только один способ разложения суммы квадратов. Для *несбалансированных* данных возможны различные способы, большинство из которых не аддитивны, т.е. $SS_y \neq SS_a + SS_e$.

В любом случае, общую некорректированную сумму квадратов ($SS_y = y'y = \sum_{ij} y_{ij}^2$) можно разложить на некорректированную сумму квадратов модели (факторную), $SS_a'' = \hat{b}'X'y = \hat{b}'X'X\hat{b}$ и остаточную сумму квадратов $SS_e = y'y - \hat{b}'X'y$.

Факторную сумму квадратов, *скорректированную на среднее*, получают из выражения

$$SS_a = \hat{b}X'y - n\hat{\mu}^2,$$

где $n\hat{\mu}^2$ - сумма квадратов для среднего (SS_μ).

Расчет SS_a в компьютерной программе Harvey'я осуществляется следующим образом.

Пусть Z - это субматрица в редуцированной матрице $(X'X)_r$, соответствующая строкам и столбцам оцениваемых эффектов (на стр. 342 эта субматрица обведена)

$$Z = \begin{bmatrix} 0{,}1435 & -0{,}0648 \\ -0{,}0648 & 0{,}1019 \end{bmatrix}.$$

Тогда

$$SS_a = \hat{b}'_z Z^{-1} \hat{b}_z,$$

где $\hat{b}_z$ - часть вектора оценок $\hat{b}_r$, которая соответствует субматрице Z.

Для числового примера имеем:

$$SS_a = \hat{b}'_z Z^{-1} \hat{b}_z = [-1\ 1]\begin{bmatrix} 0{,}1435 & -0{,}0648 \\ -0{,}0648 & 0{,}1019 \end{bmatrix}^{-1} \begin{bmatrix} -1 \\ 1 \end{bmatrix} =$$

$$= [-1\ 1]\begin{bmatrix} 9{,}7778 & 6{,}2222 \\ 6{,}2222 & 13{,}7778 \end{bmatrix}^{-1} \begin{bmatrix} -1 \\ 1 \end{bmatrix} =$$

$$= [-3{,}5556\ \ 7{,}5556]\begin{bmatrix} -1 \\ 1 \end{bmatrix} = 11{,}1112.$$

Некорректированная сумма квадратов для модели после редукции:

$$SS''_a = \hat{b}'_r (X'y)_r = [\,5\ \ -1\ \ 1\,]\begin{bmatrix} 94 \\ -14 \\ 18 \end{bmatrix} = 502.$$

Следует отметить, что аналогичное значение будет и без редукции:

$$\hat{b}'X'y = [5\ -1\ 1\ 0]\begin{bmatrix} 94 \\ 16 \\ 48 \\ 30 \end{bmatrix} = 502.$$

Общая не скорректированная на среднее сумма квадратов (SS_y):

$$SS_y = y'y = 568.$$

Остаточная сумма квадратов:

$$SS_e = y'y - \hat{b}'X'y = 568 - 502 = 66.$$

Сумма квадратов для общего среднего (корректировка на среднее)

$$SS_\mu = 502 - 11{,}1112 = 490{,}8888.$$

Результаты разложения суммы квадратов сведены в таблицу:

Источник	df	SS	MS	F-критерий
Общая, в т.ч.	18	568,0000		
Среднее	1	490,8888		
Фактор А	2	11,1112	5,5556	1,263 n.s.
Остаток	15	66,0000	4,4000	

Критерий Фишера $F = MS_a / MS_e = 5{,}5556 / 4{,}4000 = 1{,}263$.

Расчеты по стандартной процедуре:

- общая некорректированная сумма квадратов:

$$SS_y = \sum_i \sum_j y_{ij}^2 = 3^2 + 5^2 + \ldots + 4^2 = 568;$$

- сумма квадратов для средней:

$$SS_\mu = (\sum_i \sum_j y_{ij})^2 / n = 490{,}8888;$$

- факторная сумма квадратов:

$$\begin{aligned} SS_a &= \sum_i [(\sum_j y_{ij})^2 / n_i] - SS_\mu = \\ &= 16^2/4 + 48^2/8 + 30^2/6 - 490{,}8888 = \\ &= 502 - 490{,}8888 = 11{,}1112; \end{aligned}$$

- остаточная сумма квадратов:

$$SS_e = SS_y - \sum_i [(\sum_j y_{ij})^2 / n_i] = 568 - 502 = 66.$$

Таким образом, оба способа расчета дают идентичные результаты.

20.4. Допущения и оценки эффектов

Для решения системы нормальных уравнений было принято допущение, что $\sum_i \hat{a}_i = 0$. В результате были получены следующие оценки:

$$\hat{\mu} = 5, \quad \hat{a}_1 = -1, \quad \hat{a}_2 = +1 \quad \text{и} \quad \hat{a}_3 = 0$$

или

$$\hat{b}' = [\, 5 \quad -1 \quad +1 \quad 0 \,].$$

Другие допущения приводят к иным оценкам. Например, при $\hat{\mu}=0$ вектор оценок будет

$$\hat{b}' = [\ 0 \quad +4 \ +6 \ +5].$$

Выше отмечалось, что допущения не влияют на суммы квадратов. Но какие допущения следует выбрать, если необходимо оценить вектор параметров b?

Ответа на этот вопрос нет, т.к. параметры определяются неоднозначно. Имеется p+1 параметров, но только «p» градаций фактора. Для того чтобы выяснить содержание параметров, необходимо ввести *ограничение* (restriktion). Разница между ограничением и допущением состоит в том, что допущения вводят только для того, чтобы решить систему нормальных уравнений. А ограничения вводят на параметры, которые входят в модель, для того чтобы придать им однозначность.

Несмещенные оценки параметров получают только тогда, когда ограничения соответствуют допущениям. При ограничении $\sum_i \hat{a}_i = 0$, μ приобретает содержание *невзвешенной средней* из средних по каждой градации фактора $(\mu + a_i)$. При этом $a_i = (\mu + a_i) - \mu$. В этом случае допущение $\sum_i \hat{a}_i = 0$ верно, в том смысле, что помогает вычислить оценки вектора b.

Общее среднее μ оценивается как *невзвешенное среднее всех средних оценок градаций* фактора. Общее среднее $\sum_i \sum_j y_{ij}/n$ является несмещенной оценкой только при ограничении $\sum_i n_i \hat{a}_i = 0$. В программе Harvey,я используется допущение $\sum_i \hat{a}_i = 0$. Значения $\hat{a}_i$ называют «*оценки констант*» (constant estimate), а $\hat{\mu}$ и $\hat{\mu} + \hat{a}_i$ - как «*средние наименьших квадратов*» (least-squares means).

Для числового примера имеем:

Общее среднее с ограничением $\sum_i n_i \hat{a}_i = 0$

$$= \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}/n = 94/18 = 5{,}2222.$$

Среднее наименьших квадратов с ограничением $\sum_i \hat{a}_i = 0$

$$= \sum_i (\sum_j y_{ij}/n_i)/p = (16/4 + 48/8 + 30/6)/3 = 5{,}0000.$$

Таким образом, средние градаций с небольшим числом наблюдений имеют тот же «вес» в *общей* средней наименьших квадратов, что и средние градаций с большим числом наблюдений. Поэтому среднее наименьших квадратов считается *невзвешанным*. Оно зависит как от оценок эффектов, включенных в модель, так и от степени сбалансированности данных.

Для рассматриваемого числового примера R( )-запись некорректированной суммы квадратов есть:

$$R(\mu,a) = SS''_a = \hat{b}'_r (X'y)_r = \hat{b}' X'y = 502,$$

причем модель имела вид

$$y_{ij} = \mu + a_i + e_{ij}.$$

При использовании модели

$$y_{ij} = \mu + e_{ij}$$

сумма квадратов по модели есть:

$$R(\mu) = SS_\mu = 490{,}8888.$$

Скорректированную на среднее сумму квадратов по фактору A (SS_a) рассчитывают из выражения

$$R(\mu,a) - R(\mu) = R(a \mid \mu) = 11{,}1112.$$

$R(a \mid \mu)$ соответствует той редукции, которая происходит при введении в модель, в которой уже имеется μ, фактора A.

Вышескорректированная сумма квадратов между градациями фактора A выражалась как

$$b'_z Z^{-1} b_z.$$

Используя R( )-запись, имеем следующие суммы квадратов для таблицы однофакторного дисперсионного анализа:

- общая $y'y$
- среднее $R(\mu)$
- факторная $R(a \mid \mu)$
- остаточная $y'y - R(\mu,a)$.