Дисперсионный анализ многомерных неполных наблюдений с приложением в медицине

Пономаренко Артем, гр. 422

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доцент Алексеева Н.П.

Рецензент: специалист по био-статистике, Скурат Е.П.

Санкт-Петербург
2022г.

Постановка задачи

Рассматриваются зависимости между количественными и качественными признаками с помощью дисперсионного и симптомного анализа.

Задачи:

1. Редукция размерности на основе симптомного анализа,
2. Сравнение результатов симптомного и дисперсионного анализа,
3. Поиск наиболее информативных признаков,
4. Применение дисперсионного анализа для неполных повторных наблюдений.

Коэффициенты неопределенности

Формула Шеннона

$$H(\xi) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}.$$

Количество информации

$$I(\xi, \eta) = H(\xi) - H(\xi \,|\, \eta) = H(\eta) - H(\eta \,|\, \xi).$$

Односторонние коэффициенты неопределенности

$$J_{X|Y} = \frac{I(X,Y)}{H(Y)} \cdot 100\%, \quad J_{Y|X} = \frac{I(X,Y)}{H(X)} \cdot 100\%.$$

Двусторонние коэффициенты неопределенности

$$J = \frac{H(X)}{H(X)+H(Y)} J_{X|Y} + \frac{H(Y)}{H(X)+H(Y)} J_{Y|X}.$$

Симптомный анализ

Симптом

Пусть $\mathbb{X}_n = (X_0, \ldots, X_{n-1})$ — случайный вектор дихотомических признаков, α_n — вектор коэффициентов. Тогда $\mathcal{L}(\mathbb{X}_n) = \alpha_n^{\intercal}\mathbb{X}_n \pmod 2$ — симптом.

Суперсимптом

Пусть $V(\mathbb{X}_1) = X_1, V(\mathbb{X}_n) = (V(\mathbb{X}_{n-1}), X_n, V(\mathbb{X}_{n-1})\, X_n)$ — импульсный вектор произведений.
Тогда $S(\mathbb{X}_n) = \alpha_n^{\intercal} V(\mathbb{X}_n) \pmod 2$ — суперсимптом.

Алгоритм перебора суперсимптомов [Н.П. Алексеева, 2021]:

1. Составляются всевозможные комбинации из троек признаков,
2. Итеративно перебираются полученные суперсимптомы,
3. Для каждого суперсимптома вычисляется его значимость.

Трёхфакторный дисперсионный анализ

$x_{ijk} = \mu + a_i + b_j + c_k + (ab)_{ij} + (bc)_{jk} + (ac)_{ik} + (abc)_{ijk} + \epsilon_{ijkl}$, где:

- a_i, b_j, c_k — дифференциальные эффекты факторов A, B, C,
- $(ab)_{ij}, (bc)_{jk}, (ac)_{ik}$ — эффекты взаимодействия факторов первого порядка,
- $(abc)_{ijk}$ — эффекты взаимодействия факторов второго порядка,
- $\epsilon_{ijkl} \sim N(0, \sigma^2)$ — независимые случайные ошибки.

Модель повторных наблюдений

$$x_{ijt} = \mu + \alpha_i + e_{ij}^{(1)} + \beta_t + \gamma_{it} + e_{ijt}, \text{ где:}$$

- μ — генеральное среднее,
- α_i — фиксированный эффект группы,
- β_t — фиксированный эффект времени,
- γ_{it} — фиксированный эффект взаимодействия группы и времени,
- $e_{ij}^{(1)} \sim N\left(0, \sigma_1^2\right)$ — независимые ошибки, вызванные разнообразием индивидов,
- $e_{ijt} \sim N\left(0, \sigma^2\right)$ — независимые общие ошибки.

Поправки в случае неполных данных

В работе [Н.П. Алексеева, 2017] были предложены индивидуальная H_{ij} и групповая G_i поправки такие, что:

$$X_{ijt} = x_{ij.} - H_{ij} - G_i$$

Тогда выполняется

$$\mathbb{E}x_{ij} = \mu + \alpha_i, \quad \mathbb{E}\left(x_{ijk} - x_{ij}\right) = \beta_k + \gamma_{ik}.$$

Построение нового признака

Пусть имеется m признаков $X_1, \ldots, X_m$, измеренные в T временных точках. Тогда

$$Y_t = \sum_{i=1}^{m} a_t X_{ti}, \quad t = 1, \ldots, T.$$

Оптимальные коэффициенты для эффектов группы, времени и взаимодействия можно численно найти из:

$$F_K(a_1, \ldots, a_m) \to \max_{a_1, \ldots, a_m}, \text{где } K \in \{A, B, AB\}$$

Результаты. Описание данных

Данные о лечении больных от COVID-19, всего 143 индивида.

- X_1 = COVID — подтверждённый COVID,
- X_2 = ИМ — инфаркт миокарда,
- X_3 = ПЖ — перегрузка желудочка.

Независимые признаки:

- исход — выписка или летальный исход,
- креатинин — количество креатинина в крови.

Мера значимости:

- дихотомические — коэффициенты неопределённости,
- метрические — p-значения критериев однородности (например, Манна–Уитни–Вилкоксона).

Результаты. Значения симптомов

$$\mathbb{X} = X_3$$

$\mathbb{X}$	j	0	1	p
0	1	24	20	0.45
0	3	3	3	0.50
0	0	0	9	1.00
0	2	0	1	1.00
1	**4**	0	1	1.00
1	**5**	0	4	1.00
1	**7**	0	1	1.00
1	**6**	0	0	—

Таблица: Значения симптома относительно смертности

$$\mathbb{X} = X_3 + X_1X_2 + X_2X_3 = \\ = X_1X_2 + \overline{X_2}X_3$$

$\mathbb{X}$	j	Среднее	Кол-во
0	0	96.09	37
0	6	104.00	1
0	2	107.25	12
0	1	147.42	7
1	**7**	148.00	1
1	**4**	185.00	3
1	**3**	204.66	3
1	**5**	—	0

Таблица: Средние значения креатинина по группам

$$j = X_1 + 2X_2 + 4X_3$$

Результаты. Сравнение методов

$$\mathbb{X} = X_3 + X_1X_2 + X_2X_3 = X_1X_2 + \overline{X_2}X_3$$

На уровне значимости $\alpha = 0.05$:

	p
X_1	**0.001**
X_2	0.9169
X_3	**0.048**
$X_1 : X_2$	**0.034**
$X_1 : X_3$	0.233
$X_2 : X_3$	0.317

	j				
$\mathbb{X}$	X_1	X_3	$X_1 : X_2$	Среднее	Кол-во
0	0	0	0	96.09	37
0	6	6	6	104.00	1
0	2	2	2	107.25	12
0	1	1	1	147.42	7
1	7	7	7	148.00	1
1	4	4	4	185.00	3
1	3	3	3	204.66	3
1	5	5	5	—	0

X_1 — COVID, X_2 — ИМ, X_3 — ПЖ; $\quad j = X_1 + 2X_2 + 4X_3$.

Результаты. Наиболее информативные признаки

Таблица: Односторонние коэффициенты неопределённости

Признак	$J(\%)$
тяжесть	19.91
Од. ЖА	9.08
ХСН	9.06
стадия по КТ	8.88
ИБС	7.77

Таблица: Коэффициенты неопределённости относительно исхода

1 признак	2 признак	$J(\%)$
НСР	тяжесть	50.31
ИБС	тяжесть	45.20
ХСН	тяжесть	43.58
ритм	тяжесть	43.01
стадия по КТ	тяжесть	41.53

Результаты. Построение нового признака

Данные с пропусками о послеоперационных показателях пациентов, измеряемые в различные моменты времени.

$\mathbf{X_1}$ = AD — диастолическое давление, $\mathbf{X_2}$ = WP — размер стенки желудочка, $\mathbf{Y}$ = исход лечения.

Ищем коэффициенты a_1, a_2 линейной комбинации:

$$Z = a_1 X_1 + a_2 X_2.$$

Таблица: Коэффициенты линейной комбинации и p-значения

Эффект	a_1	a_2	p-значение
Группа	1.295	0.223	0.231
Время	0.271	-2.000	**0.002**
Взаимодействие	0.696	1.125	0.553

Результаты. График взаимодействия

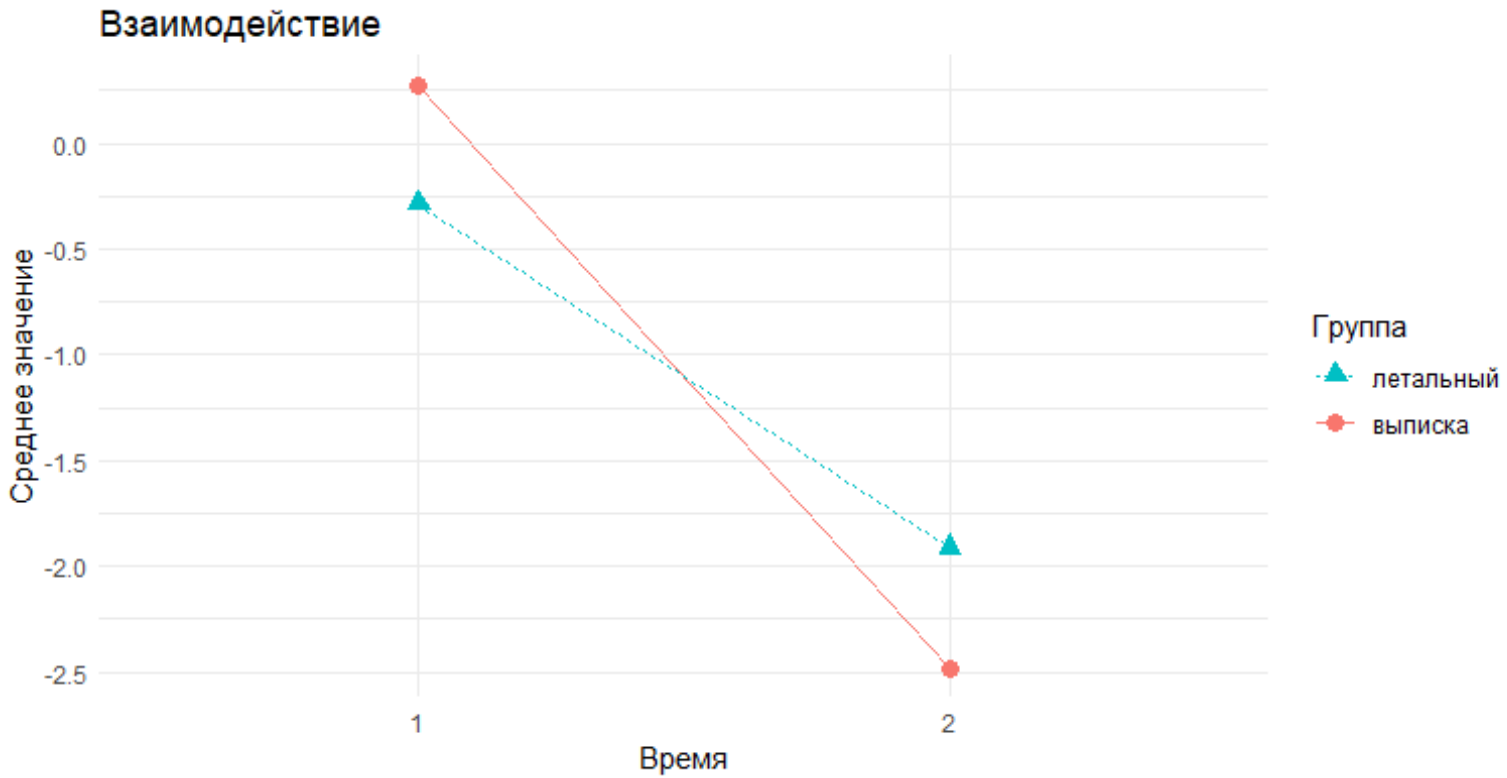

Рис.: График средних нового признака в зависимости от группы и времени

Итоги

Результаты:

- Написана программа для нахождения оптимального суперсимптома с помощью коэффициеинтов неопределённости и критерия Вилкоксона,
- Написана программа на R для построения наиболее значимых новых признаков в случае неполных данных с повторениями,
- Найдены наиболее значимые факторы, влияющие на исход лечения,
- Изучен и примененён симптомный анализ для редукции размерности.