Дисперсионный анализ

Малов Сергей Васильевич

Санкт-Петербургский электротехнический университет

28 ноября 2020 г.

План

1. Однофакторный дисперсионный анализ
2. Двухфакторный анализ
3. Многофакторный анализ

Модель однофакторного анализа

Простая группировка

- Распределение наблюдаемой величины Y определяется значением фактора группировки z.
- Наблюдение (Y, z)
  - Y – наблюдаемая величина (исследуемая характеристика)
  - $z \in \{1, \dots, d\}$ – фактор группировки, имеющий d уровней
- Модель:

$$\mathbb{E}_\theta(Y|z = i) = \eta_j, \quad i = 1, \dots, d$$

  - $\boldsymbol{\eta} = (\eta_1, \dots, \eta_d)'$ – средние по группам
  - $\mathbb{D}_\theta Y = \sigma^2$ – параметр дисперсии
- Соответствующая модель линейной регрессии

$$\mathbb{E}_\theta(\boldsymbol{Y}|\boldsymbol{z}) = \mathbf{X}'\boldsymbol{\eta}$$

  - $$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 1 & 1 & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \underbrace{0 \quad 0}_{z=1} & & \underbrace{0 \quad 0 \quad 0}_{z=2} & & & \dots & \underbrace{1}_{z=d} \end{pmatrix}$$
  - $Y_1, \dots, Y_d$ – независимые величины

Модель однофакторного анализа

Запись с использованием группировки

- Y_{ij}, j-е наблюдение i-й группы, $j = 1, \dots, n_i$, $i = 1, \dots, d$
- Модель: $\mathbb{E}_\theta Y_{ij} = \eta_i$, $j = 1, \dots, n_i$, $i = 1, \dots, d$
- Величины Y_{ij} независимы
- Дополнительные предположения: $Y_{ij} \sim \mathcal{N}(\eta_i, \sigma^2)$

Оценивание

- При наличии хотя бы одного наблюдения в группе $(n_i \geq 1)$:

$$\hat{\eta}_i = \overline{Y}_{i+} = \sum_{i=1}^{n_i} Y_{ij}/n_i$$

- $\hat{\eta}_1, \dots, \hat{\eta}_d$ независимы и $\mathbb{D}\hat{\eta}_i = \sigma^2/n_i$
- При отсутствии наблюдений в группе $(n_i = 0)$ параметр η_i не может быть оценен
- При наличии пустых ячеек соответствующие параметры следует исключить из модели

Модель однофакторного анализа

Запись параметра с использованием сравнений

- Сравнение параметров $\eta_1, \dots, \eta_d$:

$$\psi = \sum_{i=1}^{d} c_i \eta_i, \quad \text{где} \quad \sum_{i=1}^{d} c_i = 0.$$

- Выбор весов $\{v_i\}_{i=1}^{d}$: $v_i \geq 0$ и $\sum_{i=1}^{d} v_i = 1$
- Параметризация

$$\eta_i = \mu + \alpha_i, \quad \alpha_* = \sum_{i=1}^{d} v_i \alpha_i = 0$$

  - $\mu = \eta_* = \sum_{i=1}^{d} v_i \eta_i$ – взвешенное среднее
  - $\alpha_i = \eta_i - \eta_*$ – главные эффекты, $i = 1, \dots, d$.
  - α_i – сравнения параметров $\boldsymbol{\eta}$
- Стандартный выбор весов
  - $v_1 = 1$, $v_1 = 0$ при $i \neq 1$ – базовый первый уровень
  - $v_d = 1$, $v_1 = 0$ при $i \neq d$ – базовый последний уровень
  - $v_i = 1/d$ – равные веса

Изучение влияния фактора на результат

Гипотеза однородности групп: отсутствие влияния фактора на результат

- Нулевая гипотеза
$$H_0 : \eta_1 = \ldots = \eta_d$$
- Эквивалентная форма записи
$$H_0 : \alpha_1 = \ldots = \alpha_{d-1} (= \alpha_d) = 0$$
- Можно записать с использованием любых $d - 1$ линейно независимых сравнений $\psi_1, \ldots, \psi_{d-1}$
$$H_0 : \psi_1 = \ldots = \psi_{d-1} = 0$$

Проверка гипотезы

- Статистика $\mathbb{F}$-критерия

$$\mathbb{F} = \frac{\overline{SS}_H}{\overline{SS}_e} = \frac{\overline{SS}_H/q}{\overline{SS}_e/(n-r)}$$

  - $SS_H = SS(\hat{\boldsymbol{\eta}}_H) - SS_e = \sum_i k_i(\overline{Y}_{i+} - \overline{Y}_{++})^2$
  - $SS_e = \sum_{i=1}^{d} \sum_{i=1}^{n_i} (Y_{ij} - \overline{Y}_{++})^2$
  - числа степеней свободы: $q = d - 1$, $n - r = n_+ - d$
- Распределение статистики $\mathbb{F}$-критерия при нулевой гипотезе $F_{d-1,n-d}$
- Распределение статистики $\mathbb{F}$-критерия при альтернативе $F_{\nu,d-1,n-d}$
  - параметр нецентральности $\nu = \sum_{i=1}^{d} n_i(\eta_i - \bar{\eta})^2$.

Множественные сравнения

Уточнение результатов проверки $H_0 : \psi_1 = \ldots = \psi_{d-1} = 0$

- Метод Шеффе позволяет получить совместные доверительные интервалы всех ψ_i и их линейных комбинаций $\psi = \alpha_1\psi_1 + \ldots + \alpha_q\psi_q$, $q \leq d-1$

$$[\hat{\psi} - \sqrt{x_\alpha q}\hat{\sigma}_\psi, \hat{\psi} + \sqrt{x_\alpha q}\hat{\sigma}_\psi]$$

  - x_α: $F_{q,n-d}(x_\alpha) = \alpha$ – квантиль распределения Фишера-Снедекора
  - $\hat{\sigma}^2_\psi$ – оценка дисперсии МНК оценки $\hat{\psi}$ параметра ψ
- $\mathbb{F}$-критерий принимает гипотезу в том, и только в том случае, если доверительный интервал для каждого ψ содержит 0.
- Метод Шеффе позволяет выявить сравнения, ответственные за отвержение гипотезы в случае ее отвержения

Множественные сравнения

С использованием метода Шеффе можно проверять односторонние гипотезы

- Например, для проврки гипотезы

$$H_0 : \eta_1 < \ldots < \eta_d$$

следует:

- построить совместные доверительные интервалы для сравнений $\psi_i = \eta_{i+1} - \eta_i$, $i = 1, \ldots, d - 1$
- если все доверительные интервалы полностью окажутся в положительной области, то гипотезу можно принять.

План

1. Однофакторный дисперсионный анализ
2. Двухфакторный анализ
3. Многофакторный анализ

Формулировка в духе однофакторного анализа

- Распределение наблюдаемой величины Y определяется значением двух факторов группировки (z_1, z_2) .
- Наблюдение $(Y, \boldsymbol{z})$, $\boldsymbol{z} = (z_1, z_2)$
  - Y – наблюдаемая величина (исследуемая характеристика)
  - $z_l \in \{1, \ldots, d_l\}$ – фактор группировки, имеющий d_l уровней
  - $\boldsymbol{z}$ – фактор простой группировки, имеющий $d_1 d_2$ уровней
- Модель

$$\mathbb{E}_\theta(Y|\boldsymbol{z} = (i,j)) = \eta_{ij}, \quad i = 1, \ldots, d_1,\ j = 1, \ldots, d_2$$

  - η_{ij}, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_2$ – средние по группам
  - $\mathbb{D}_\theta Y = \sigma^2$ – параметр дисперсии
- Статистические данные $(\boldsymbol{Y}, \boldsymbol{z})$
  - $\boldsymbol{Y} = (Y_1, \ldots, Y_d)'$ – независимые величины
  - При $\boldsymbol{z}_s = (i,j)$ наблюдение Y_s – имеет нормальное распределение $\mathcal{N}(\eta_{ij}, \sigma^2)$
- Запись с группировкой

$$\mathbb{E}(Y_{ijk}) = \eta_{ij},\ i = 1, \ldots, d_1,\ j = 1, \ldots, d_2,\ k = 1, \ldots, n_{ij}$$

  - n_{ij} – число наблюдений в группе $\boldsymbol{z} = (i,j)$

Двухфакторный подход

- Выбор весов $\{v_i\} : \sum_{i=1}^{d_1} v_i = 1$, $\{w_j\} : \sum_{i=1}^{d_2} w_j = 1$
- Чтобы разделить влияние факторов используют параметризацию
$$\eta_{ij} = \mu + \alpha_i^{(1)} + \alpha_j^{(2)} + \alpha_{ij}^{(12)}$$
    - μ – взвешенное среднее
    - $\alpha_i^{(1)}$, $\alpha_j^{(2)}$ – главные эффекты
    - $\alpha_{ij}^{(12)}$ – взаимодействия
- Явные формулы для параметров модели
    - $\mu = \eta_{**} = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} v_i w_j \eta_{ij}$
    - $\alpha_i^{(1)} = \eta_{i*} - \eta_{**} = \sum_{j=1}^{d_2} w_j \eta_{ij} - \eta_{**}$
    - $\alpha_j^{(2)} = \eta_{*j} - \eta_{**} = \sum_{i=1}^{d_1} v_i \eta_{ij} - \eta_{**}$
    - $\alpha_{ij}^{(12)} = \eta_{ij} - \eta_{i*} - \eta_{*j} + \eta_{**}$
- Ограничения
    - $\alpha_*^{(1)} = \sum_{i=1}^{d_1} v_i \alpha_i^{(1)} = 0$; $\alpha_*^{(2)} = \sum_{j=1}^{d_2} w_j \alpha_j^{(2)} = 0$
    - $\alpha_{i*}^{(12)} = 0$ при всех i; $\alpha_{*j}^{(12)} = 0$ при всех j.

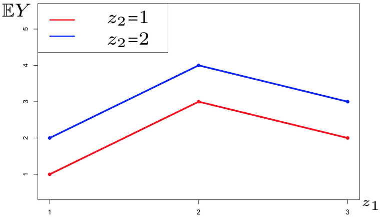
Взаимодействия и главные эффекты

- Гипотеза отсутствия взаимодействий
$$H_{(12)} : \alpha_{ij}^{(12)} = 0, i = 1, \ldots, d_1, j = 1, \ldots, d_2$$
  - $\alpha_{ij}^{(12)}$ – сравнения параметров η_{ij}
  - число степеней свободы (размерность параметра $\boldsymbol{\alpha}^{(12)}$): $(d_1 - 1)(d_2 - 1)$
- При выполнении $H_{(12)}$ получаем аддитивную модель
$$\eta_{ij} = \mu + \alpha_i^{(1)} + \alpha_j^{(2)}$$
  - факторы действуют независимо
  - размерность параметра аддитивной модели $d_1 + d_2 - 1$
  - аддитивная модель может использоваться для некоторых неполных планов ($n_{ij} = 0$ при некоторых i, j)
- Справедливость гипотезы не зависит от выбора весов

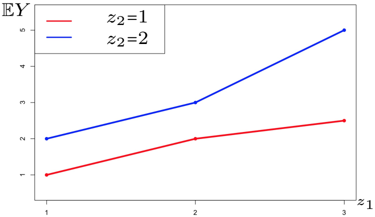
Модель $\mathbb{E}(Y|z_1 = i, z_2 = j) = \mu + \alpha_i^{(1)} + \alpha_j^{(2)} + \alpha_{ij}^{(12)}$

- $z_1 \in \{1, 2, 3\}$; $z_2 \in \{1, 2\}$

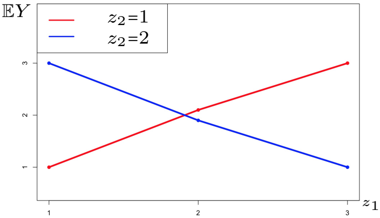
Аддитивная модель



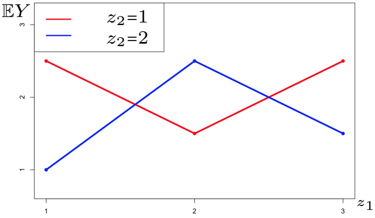
Согласованное действие



Пересечение 1



Пересечение 2



Проверка значимости взаимодействий

Проверка гипотезы отсутствия взаимодействий

- Нулевая гипотеза
$$H_{(12)} : \gamma_{ij} = 0, i = 1, \ldots, d_1, j = 1, \ldots, d_2$$
  - может быть переписана с использованием $(d_1 - 1)(d_2 - 1)$ линейных комбинаций γ_{ij}
  - при выборе весов с использованием для каждого фактора базового уровня, обращение в нуль $(d_1 - 1)(d_2 - 1)$ значений γ_{ij}, соответствующих нулевым значениям весов, задает гипотезу $H_{(12)}$
- Для проверки гипотезы используют $\mathbb{F}$-критерий
- В случае одного наблюдения $n_{ij} = 1$ при всех (i, j) проверка гипотезы $H_{(12)}$ невозможна
- Статситика критерия имеет $F_{(d_1-1)(d_2-1),n-d_1 d_2}$-распределение
- При альтернативе $\mathbb{F}$-статисика имеет нецентральное $F_{\nu,(d_1-1)(d_2-1),n-d_1 d_2}$-распределение
- Методы множественного сравнения позволяют делать более точные выводы о характере взаимодействий

Гипотезы о главных эффектах

Рассмотрим гипотезы

$$H_{(1)}: \alpha_i^{(1)} = 0, i = 1, \dots, d_1 \qquad H_{(2)}: \alpha_i^{(1)} = 0, j = 1, \dots, d_2$$

- Гипотеза $$H_{(1)}: \alpha_i^{(1)} = 0, i = 1, \dots, d_1$$
  - размерность параметра $d_1 - 1$
  - не влечет отсутствия влияния фактора z_1 на результат
  - при отсутствии взаимодействий сравнения главных эффектов не зависят от выбора весов
  - при наличии взаимодействий выполнение $H_{(1)}$ зависит от выбора весов
- Отсутствие влияния фактора z_1 на результат равносильно одновременному выполнению $H_{(12)}$ и $H_{(1)}$
- Отсутствие влияния двух факторов на результат определяется одновременным выполнением $H_{(12)}$, $H_{(1)}$ и $H_{(2)}$

План

1. Однофакторный дисперсионный анализ
2. Двухфакторный анализ
3. Многофакторный анализ

Однофакторный подход

- Распределение наблюдаемой величины Y определяется значением k факторов группировки $(z_1, \dots, z_k)$.
- Наблюдение $(Y, \boldsymbol{z})$, $\boldsymbol{z} = (z_1, \dots, z_k)$
  - Y – наблюдаемая величина (исследуемая характеристика)
  - $z_l \in \{1, \dots, d_l\}$ – фактор группировки, имеющий d_l уровней
  - $\boldsymbol{z}$ – фактор простой группировки, имеющий $d_1 \cdot \ldots \cdot d_k$ уровней
- Модель

$$\mathbb{E}_\theta(Y|\boldsymbol{z} = (i_1, \dots, i_k)) = \eta_{i_1 \dots i_k}, \quad i_l = 1, \dots, d_l, \ l = 1, \dots, k$$

  - $\eta_{i_1 \dots i_k}$, $i_l = 1, \dots, d_l$, $l = 1, \dots, k$ – средние по группам
  - $\mathbb{D}_\theta Y = \sigma^2$ – параметр дисперсии
- Статистические данные $(\boldsymbol{Y}, \boldsymbol{z})$
  - $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$ – независимые величины
  - При $\boldsymbol{z}_s = (i_1, \dots, i_k)$ наблюдение Y_s – имеет нормальное распределение $\mathcal{N}(\eta_{i_1 \dots i_k}, \sigma^2)$
- План полный, если каждому набору значениий факторов соответствует хоть одно наблюдение

Влияние факторов и их комбинаций

- Веса выбираются для каждого из факторов
- Помимо взаимодействий двух факторов появляются взаимодействия трех (и более) факторов – взаимодействия 2-го, 3-го и.т.д. порядков
- Главные эффекты и взаимодействия определяются рекурсивно для каждой комбинации факторов
- Взаимодействия и главные эффекты – сравнения параметров $\boldsymbol{\eta}$
- Взвешенные суммы взаимодействий по каждому индексу при любом фиксированном наборе остальных индексов равны нулю
- Взаимодействия высоких порядков трудно поддаются интерпретации

Выдвижение и проверка гипотез

- Объективными считаем гипотезы об отсутствии взаимодействий определенных факторов или главных эффектов
- Существует прямой и обратный подходы изучения модели
  - Прямой подход подразумевает последовательность выдвижения гипотез о взаимодействиях, начиная с более высоких порядков к гипотезам об отсутствии взаимодействий долее низких порядков
  - Обратный подход подразумевает введение параметров, начиная с главных эффектов к взаимодействиям высоких порядков
- Если не ставить задачу выбора наилучшей модели, то обычно достаточно ограничиться аддитивной моделью или моделью с взаимодействиями только 1-го порядка