Множественные сравнения для повторных наблюдений

Зенкова Наталья Валентиновна, гр. 422

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к.ф.-м.н., доцент Алексеева Н. П.
Рецензент: к.ф.-м.н., аналитик Уфлянд А. Г.

Санкт-Петербург
2018г.

Модель для повторных наблюдений с пропусками

Модель дисперсионного анализа для повторных наблюдений:

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

- μ — генеральное среднее,
- α_i, β_t, γ_{it} — фиксированные эффекты группы, времени и взаимодействия этих двух эффектов,
- $e_{ij}^{(1)} \sim \mathrm{N}(0, \sigma_1^2)$, $e_{ijt} \sim \mathrm{N}(0, \sigma^2)$ — взаимно независимые ошибки.

  Пусть

  - $N_{ij} \subset \{1, \ldots, T\}$, где T — количество временных точек.
  - $M_{it} \subset \{1, \ldots, N\}$, где N — количество индивидов.
  - I — количество групп.

Важно: $N_{ij} \neq \{1, \ldots, T\}$, $M_{it} \neq \{1, \ldots, N\}$.

Постановка задачи

Задача: Модифицировать критерии множественных сравнений для повторных наблюдений с пропусками и применить их к реальным данным.

В терминах проверки гипотез:

1. $H_0 : \beta_1 = \ldots = \beta_T$ против $H_1 : \exists\ t^{'} \neq t,\ \ \beta_{t'} \neq \beta_t\ \ \forall\ t = 1, \ldots, T$

   H_0 отвергается $\Rightarrow$ множественные сравнения.

2. $H_0^{(i,j)} : t_i \neq t_j \quad \beta_{t_i} = \beta_{t_j}$ против
   $H_1^{(i,j)} : \exists\ \ (t_i, t_j),\ \ t_i \neq t_j \quad \beta_{t_i} \neq \beta_{t_j}.$

Определение

$FWER$ — вероятность хотя бы один раз отвергнуть гипотезу $H_0^{(i_0,j_0)}$, когда верны $H_0^{(i,j)}$ для $i \neq j$.

Глобальная задача множественной проверки гипотез: $FWER \leq \alpha$.

Смещение модели

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

Разделим модель на 2 части: $x_{ijt} = z_{ij} + y_{ijt}$, где в случае полных данных получим:

$$z_{ij} = x_{ij.}, \quad y_{ijt} = x_{ijt} - x_{ij.},$$
$$\mathbb{E}(z_{ij}) = \mu + \alpha_i, \quad \mathbb{E}(y_{ijt}) = \beta_t + \gamma_{it}.$$

Утверждение

1. $\mathbb{E}(z_{ij}) = \mathbb{E}(x_{ij.})$, если $N_{ij} = \{1, \ldots, T\}$.
2. $\mathbb{E}(z_{ij}) \neq \mathbb{E}(x_{ij.})$, если $N_{ij} \subset \{1, \ldots, T\}$ и $N_{ij} \neq \{1, \ldots, T\}$.

$x_{ij.} = \frac{1}{n_{ij}} \sum\limits_{t \in \mathrm{N}_{ij}} x_{ijt}$ — индивидуальное среднее.

Смещение модели: $\mathbb{E}(x_{ij.}) - \mu - \alpha_i = \frac{1}{n_{ij}} \sum\limits_{t \in N_{ij}} (\beta_t + \gamma_{it})$.

Утверждение

В работе [Alexeyeva, 2017] введены такие G_i (групповая поправка) и H_{ij} (индивидуальная поправка), что

$$\mathbb{E}(G_i + H_{ij}) = \frac{1}{n_{ij}} \sum_{t \in N_{ij}} (\beta_t + \gamma_{it}),$$

где N_{ij} — кол-во вр. точек у индивида j в группе i, $n_{ij} = |N_{ij}|$.

Таким образом, получаем 2 модели:

$$z_{ij} = x_{ij.} - (H_{ij} + G_i) = \mu + \alpha_i + \varepsilon^1_{ij},$$
$$y_{ijt} = x_{ijt} - x_{ij.} + (H_{ij} + G_i) = \beta_t + \gamma_{it} + \varepsilon_{ijt},$$

где ε_{ijt} — ошибки с ковариационной матрицей $\mathbf{\Lambda} = \sigma^2 \mathbf{\Sigma}$.

Дисперсионный анализ для повторных наблюдений с пропусками

В матричном виде:

$$Y = \mathbf{H}\Theta + \mathcal{E}, \text{ где}$$

- Y — вектор наблюдений, Θ — вектор пар-ов длины $I(T-1)$,
- $\mathbf{H}$ — матрица плана размерности $m_{..}$ на $I(T-1)$,
- $\mathcal{E}$ — вектор ошибок с ковариационной матрицей $\mathbf{\Lambda} = \sigma^2\mathbf{\Sigma}$.

Утверждение [Alexeyeva, 2017]

Несмещённая оценка дисперсии σ^2 имеет вид:

$$\hat{\sigma}^2 = R_0^2/(m_{..} - N - I(T-1)),$$

где $R_0^2 = (Y - \mathbf{H}\hat{\Theta})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(Y - \mathbf{H}\hat{\Theta})$, $\hat{\Theta} = (\mathbf{H}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{\Sigma}^{-1}Y$ — оценка Θ.

Дисперсионный анализ для повторных наблюдений с пропусками

Для проверки гипотезы об отсутствии эффекта времени β_t рассматривается усечённая модель с матрицей плана $\mathbf{H}_*$:

$$Y = \mathbf{H}_* \Theta_* + \mathcal{E}.$$

Модифицированный критерий Фишера [Alexeyeva, 2017]

Статистика критерия Фишера имеет вид:

$$F = \frac{(R_{0*}^2 - R_0^2)/((T-1)(I-1))}{R_0^2/(m_{..} - N - I(T-1))} \sim F((T-1)(I-1), m_{..} - N - I(T-1)),$$

где $\hat{\Theta}_* = (\mathbf{H}_*^{\mathrm{T}} \mathbf{\Sigma}^{-1} \mathbf{H}_*)^{-1} \mathbf{H}_*^{\mathrm{T}} \mathbf{\Sigma}^{-1} Y$, $R_{0*} = (Y - \mathbf{H}_* \hat{\Theta}_*)^{\mathrm{T}} \mathbf{\Sigma}^{-1} (Y - \mathbf{H}_* \hat{\Theta}_*)$.

Множественные сравнения для повторных наблюдений с пропусками

Определение

Сравнениями параметров модели $\beta_1, \dots, \beta_T$ называются линейные комбинации $\sum\limits_{t=1}^{T} c_t \beta_t$, где $\sum\limits_{t=1}^{T} c_t = 0$.

Обозначим $\hat{\psi} = \sum\limits_{t=1}^{T} c_t \hat{\beta}_t$ — оценка сравнений.

В матричном виде:

$$\hat{\psi} = C\hat{\Theta} = C(\mathbf{H}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{\Sigma}^{-1}Y = \mathbf{A}Y,$$

где $\mathrm{rank}(C) := q$, $\hat{\psi} = \sum\limits_{t=1}^{T} c_t \hat{\Theta}_t$ и $\sum\limits_{t=1}^{T} c_t = 0$.

Утверждение

$\hat{\psi}$ — несмещённая оценка сравнений: $\mathbb{E}(\hat{\psi}) = C\Theta$.

Множественные сравнения для повторных наблюдений с пропусками

Утверждение

Пусть $\hat{\psi} = \mathbf{A}Y$, $\mathbf{B} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. Тогда

$$\sigma^{-2}(\hat{\psi} - \mathbb{E}(\hat{\psi}))^{\mathrm{T}}\mathbf{B}^{-1}(\hat{\psi} - \mathbb{E}(\hat{\psi})) \sim \chi(q).$$

Модифицированный критерий Шеффе

Если $Y \sim N(\mathbf{H}\Theta, \sigma^2\mathbf{\Sigma})$, $\mathrm{rank}(\mathbf{H}) = I(T-1)$, то случайная величина $\hat{\psi} \sim N(\mathbb{E}(\hat{\psi}), \sigma^2\mathbf{B})$ и не зависит от

$R_0^2/\sigma^2 = (Y - \mathbf{H}\hat{\Theta})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(Y - \mathbf{H}\hat{\Theta})/\sigma^2 \sim \chi(m_{..} - N - I(T-1))$.

Поэтому

$$\frac{(\hat{\psi} - \mathbb{E}(\hat{\psi}))^{\mathrm{T}}\mathbf{B}^{-1}(\hat{\psi} - \mathbb{E}(\hat{\psi}))}{qs^2} \sim F(q, m_{..} - N - I(T-1)),$$

где $s^2 = R_0^2/(m_{..} - N - I(T-1))$.

Реализация дисперсионного анализа и множественных сравнений

Пусть $I = 2$, $T = 3$, $N = 29$. **Параметры модели:** $\sigma^2 = 1$, $\sigma_1^2 = 4$ и

Таблица: Параметры модели

μ	α_1	α_2	β_1	β_2	β_3	γ_{11}	γ_{21}	γ_{12}	γ_{22}	γ_{13}	γ_{23}
0	1	2	0	0	0	1.6	2.6	1	2	1	2

Данные имеют вид:

Таблица: Повторные наблюдений с пропусками

group	X1	X2	X3
1	-1.08	NA	-0.14
1	1.89	-1.41	NA
...	...	...	...
2	4.50	4.66	4.99
2	6.90	4.86	7.34
...	...	...	...

Распределение модифицированных критериев Фишера и Шеффе

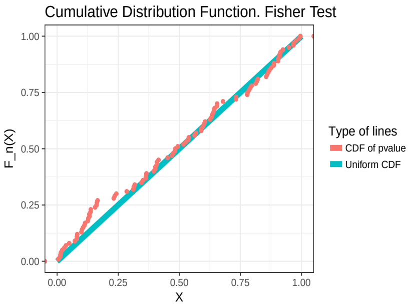

Рис.: Эмпирические функции распределения полученных p-значений и равномерного распределения

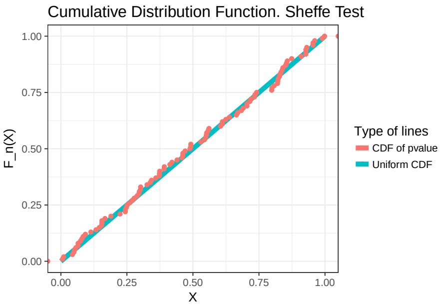

Рис.: Эмпирические функции распределения полученных p-значений и равномерного распределения

Дано:

- Данные о систолическом давлении за $T = 10$ лет после операции по замене митрального клапана;
- Количество групп $I = 2$ — мужчины и женщины;
- Количество индивидов $N = 102$.

Проверим гипотезу:

$$H_0 : \beta_1 = \ldots = \beta_T.$$

Таблица: Результат применения модифицированного критерия Фишера к реальным данным с уровнем значимости $\alpha = 0.05$

F	$pvalue$
2.08	0.0292

$pvalue < \alpha \Rightarrow H_0$ отвергается $\Rightarrow$ множественные сравнения.

Практическое применение множественных сравнений

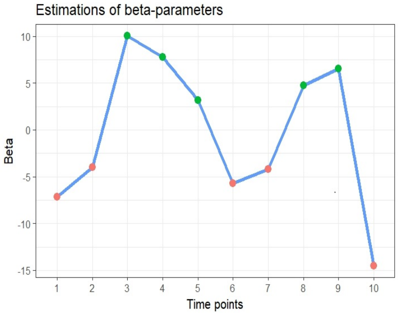

Рис.: Оценки временной компоненты β_t

Сравнение параметров:

$$\frac{\beta_1 + \beta_2 + \beta_6 + \beta_7 + \beta_{10}}{5} - \frac{\beta_3 + \beta_4 + \beta_5 + \beta_8 + \beta_9}{5}.$$

Таблица: Результат применения модифицированного критерия Шеффе к реальным данным с уровнем значимости $\alpha = 0.05$

Значение статистики критерия	$pvalue$
92.59	0.0000

Заключение

Результаты:

- Структурирована теория для повторных наблюдений с полными данными, реализована соответствующая программа на R.
- Доказана теорема о распределении статистики модифицированного критерия Фишера в случае неполных данных.
- Построен критерий множественных сравнений на основе известного критерия Шеффе для полных данных, корректность которого подтверждена моделированием.
- Разработано необходимое программное обеспечение на R и применены множественные сравнения к реальным данным.