Дисперсионный анализ для неполных повторных наблюдений с приложениями

Федорченко Сергей Андреевич, гр. 422

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

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

Санкт-Петербург
2018г.

Введение

Пусть x_{ijt} — значение количественного показателя у j-го индивида в i-й группе в t-й момент времени.

Пример

Численные значения анализов для пациентов, которые разделены на группы различным лечением.

Пусть ν_i — количество индивидов в i-й группе в самый представительный момент времени.

Математическая модель

$x_{ijt} = \mu + \alpha_i + e^1_{ij} + \beta_t + \gamma_{it} + e_{ijt}$, где $i = 1..I, j = 1..\nu_i, t = 1..T$.

Можно рассматривать модели двух типов:

- фиксированные эффекты $\alpha_i, \beta_t, \gamma_{it}$ — константы,
- случайные $\alpha_i, \beta_t, \gamma_{it}$ — нормальные случайные величины.

Также в некоторых данных возникает проблема пропусков — отсутствие значений в некоторые моменты времени t.

Гипотеза

$$H_0 : \gamma_{11} = ... = \gamma_{1T} = ... = \gamma_{IT} = 0.$$

В примере с пациентами данная гипотеза будет иметь смысл проверки одинаковой эффективности взаимодействия разных методов лечения со временем. Методом проверки данной гипотезы является дисперсионный анализ.

Задача

- Построить матричную статистику критерия для случайного эффекта взаимодействия при пропущенных данных.

Двухфакторная модель с фиксированными эффектами

Определение

$x_{ijt} = \mu + \alpha_i + e^1_{ij} + \beta_t + \gamma_{it} + e_{ijt}$, где $i = 1..I, j = 1..\nu_i, t = 1..T$.

- μ — генеральное среднее,
- α_i — фиксированный эффект группы,
- β_t — фиксированный эффект времени,
- $e^1_{ij} \sim N(0, \sigma^2_1)$ — ошибки, вызванные разнообразием индивидов,
- γ_{it} — фиксированный эффект взаимодействия,
- $e_{ijt} \sim N(0, \sigma^2)$ — общие ошибки наблюдений.

$$\sum_{i=1}^{I} \alpha_i = 0, \sum_{t=1}^{T} \beta_t = 0, \sum_{i=1}^{I} \gamma_{it} = 0, \sum_{t=1}^{T} \gamma_{it} = 0.$$

При наличии пропусков потребуется ввести дополнительные обозначения:
Пусть m_{it} — количество наблюдений в i группе в t момент времени, N_{ij} — множество временных точек индивида j в группе i, n_{ij} — количество наблюдений j-го индивида в группе i.
Обозначим $m_{i\cdot} = \sum\limits_{t=1}^{T} m_{it},\ m_{\cdot t} = \sum\limits_{i=1}^{I} m_{it},\ m_{\cdot\cdot} = \sum\limits_{t=1}^{T} m_{t\cdot}$,
$x_{ij\cdot} = \frac{1}{n_{ij}} \sum\limits_{t \in N_{ij}} x_{ijt}$ — усреднение по индексу t (аналогично другие усреднения).

Проблема

- $\mathbb{E} x_{ij\cdot} \neq \mu + \alpha_i$,
- $\mathbb{E}(x_{ijt} - x_{ij\cdot}) \neq \beta_t + \gamma_{it}$.

Групповая поправка

Для случая пропущенных данных в модели с фиксированными эффектами в статье [Н.П. Алексеева, 2017] введены две поправки, индивидуальная и групповая:
$L = (x_{\cdot\cdot 1} - x_{\cdots}, \dots, x_{\cdot\cdot T} - x_{\cdots})^T, \quad K = (x_{1\cdot\cdot} - x_{\cdots}, \dots, x_{I\cdot\cdot} - x_{\cdots})^T,$

$$M = \begin{bmatrix} \frac{m_{11}}{m_{1\cdot}} & \dots & \frac{m_{1T}}{m_{1\cdot}} \\ \vdots & \dots & \vdots \\ \frac{m_{I1}}{m_{I\cdot}} & \dots & \frac{m_{IT}}{m_{I\cdot}} \end{bmatrix}, \quad N = \begin{bmatrix} \frac{m_{11}}{m_{\cdot 1}} & \dots & \frac{m_{I1}}{m_{\cdot 1}} \\ \vdots & \dots & \vdots \\ \frac{m_{1T}}{m_{\cdot T}} & \dots & \frac{m_{IT}}{m_{\cdot T}} \end{bmatrix}.$$

$$P_0 = MN, P_0^\infty = \lim_{n\to\infty} P_0^n.$$

Определим вектор групповой поправки, как

$$G = \sum_{i=0}^{\infty} P_0^i (ML - P_0 K).$$

- J^i — матрицу инциденций значений в i-й группе,
- Λ_{ν_i} — матрица размера ν_i с элементами $\frac{1}{n_{ij}}$
- Λ_{iT} — диагональная матрица с элементами $\frac{1}{m_{it}}$
- $R_i = \Lambda_{\nu_i} J^i, P_i = R_i \Lambda_{iT} (J^i)^T, \pi(i) = \left(\frac{n_{i1}}{m_{i\cdot}}, \ldots, \frac{n_{i\nu_i}}{m_{i\cdot}} \right)$.
- $U_i = \{x_{i \cdot t}\}_{t=1}^{T}$, $V_i = \{x_{ij\cdot}\}_{j=1}^{\nu_i}$

Зададим последовательность $A_i(k) = P_i A_i(k-1)$, с начальным вектором $A_i(0) = R_i U_i - P_i V_i$. Определим вектор индивидуальных поправок i-й группы как $H_i = \sum\limits_{k=1}^{\infty} A_i(k)$.

Утверждение [Н.П. Алексеева, 2012,2017]

Пусть $y_{ijt} = x_{ijt} - x_{ij.} + H_{ij} + G_i$, тогда

- $y_{ijt} = \beta_t + \gamma_{it} + \delta_{ijt}$,
- $\delta_{ijt} = e_{ijt} - e_{ij.} + \varepsilon_{ij} + \epsilon_i$, где $\varepsilon_{ij}, \epsilon_i$ — случайные компоненты H_{ij}, G_i соответственно.

Матричный вид модели

Определения

- $Y = (Y_{11}, ..., Y_{1T}, ..., Y_{I1}, ..., Y_{IT})^{\mathrm{T}}$, где Y_{it} — вектор-строки с m_{it} компонентами $\{y_{ijt}\}_{j=1}^{m_{it}}$,
- $\Theta = (\beta_1, \gamma_{11}, ..., \gamma_{I1}, ..., \beta_{T-1}, \gamma_{1,T-1}, ..., \gamma_{I,T-1})$ — вектор с $I(T-1)$ компонентами,
- $\delta = (\delta_{11}, ..., \delta_{IT})^{\mathrm{T}}$, где δ_{it} — вектор-строки с m_{it} компонентами.

Таким образом можем модель записать в матричном виде:

$$Y = H\Theta + \delta,$$

где δ — вектор с ковариационной матрицей $\Sigma = \sigma^2\Lambda$ [Н.П. Алексеева, 2017].

Матрица плана H

Пусть $\iota = I - 1, \tau = T - 1$.

	β_1	γ_{11}	γ_{21}	$\dots$	$\gamma_{\iota 1}$	$\dots$	β_τ	$\gamma_{1,\tau}$	$\gamma_{2,\tau}$	$\dots$	$\gamma_{\iota,\tau}$
Y_{11}	1	1	0	$\dots$	0	$\dots$	0	0	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Y_{1\tau}$	0	0	0	$\dots$	0	$\dots$	1	1	0	$\dots$	0
Y_{1T}	-1	-1	0	$\dots$	0	$\dots$	-1	-1	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{I1}	1	-1	-1	$\dots$	-1	$\dots$	0	0	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Y_{I\tau}$	0	0	0	$\dots$	0	$\dots$	1	-1	-1	$\dots$	-1
Y_{IT}	-1	-1	-1	$\dots$	-1	$\dots$	-1	-1	-1	$\dots$	-1

Статистика критерия

- МНК оценка вектора параметров:
  $\hat{\Theta} = H(H^{\mathrm{T}}\Lambda^{-1}H)^{-1}H^{\mathrm{T}}\Lambda^{-1}Y$,
- Строим матрицу плана H_b усеченной модели $y_{ijt} = \beta_t + \delta_{ijt}$ из столбцов H, соответствующих β,
- МНК-оценка параметров $\hat{\beta} = (H_b^{\mathrm{T}}\Lambda^{-1}H_b)^{-1}(H_b)^{\mathrm{T}}\Lambda^{-1}Y$.

$$R_0 = (Y - H\hat{\Theta})^{\mathrm{T}}\Lambda^{-1}(Y - H\hat{\Theta}),$$

$$R_1 = (Y - H_b\hat{\beta})^{\mathrm{T}}\Lambda^{-1}(Y - H_b\hat{\beta}).$$

Утверждение

$$\mathbb{E}R_0 = \sigma^2(m_{..} - m_{.1} - I(T-1)).$$

$$\mathbb{E}R_1 = \sigma^2(m_{..} - m_{.1} - T + 1) + \frac{\sum\limits_{i=1}^{I}\sum\limits_{t=1}^{T} m_{it}\gamma_{it}^2}{(I-1)(T-1)}.$$

Распределения квадратичных форм

Утверждение

Рассмотрим

$$R_0 = (Y - H\hat{\Theta})^{\mathrm{T}}\Lambda^{-1}(Y - H\hat{\Theta}),$$

$$R_1 = (Y - H_b\hat{\beta})^{\mathrm{T}}\Lambda^{-1}(Y - H_b\hat{\beta}),$$

$$f = \frac{(R_1 - R_0)/((T-1)(I-1))}{R_0/(m_{..} - m_{.1} - I(T-1))}.$$

Тогда

- $\frac{1}{\sigma^2}R_0 \sim \chi^2(m_{..} - m_{.1} - I(T-1))$,
- $R_1 - R_0$ и R_0 независимы.
  При верной $H_0 : \gamma_{11} = ... = \gamma_{IT} = 0$
- $\frac{1}{\sigma^2}(R_1 - R_0) \sim \chi^2((I-1)(T-1))$,
- $f \sim F((T-1)(I-1), m_{..} - m_{.1} - I(T-1))$.

Двухфакторная модель со случайными эффектами

Определение

$x_{ijt} = \mu + a_i + e^1_{ij} + b_t + g_{it} + e_{ijt}$, где $i = 1..I, j = 1..\nu_i, t = 1..T$.

- μ — генеральное среднее,
- $a_i \sim N(0, \sigma^2_a)$ — случайный эффект группы,
- $b_t \sim N(0, \sigma^2_b)$ — случайный эффект времени,
- $e^1_{ij} \sim N(0, \sigma^2_1)$ — ошибки, вызванные разнообразием индивидов,
- $g_{it} \sim N(0, \sigma^2_g)$ — случайный эффект взаимодействия,
- $e_{ijt} \sim N(0, \sigma^2)$ — общие ошибки наблюдений.

Отличия в проверке гипотез для случайных эффектов

Гипотеза

$$H_0 : \sigma_g^2 = 0.$$

Утверждение

Рассмотрим

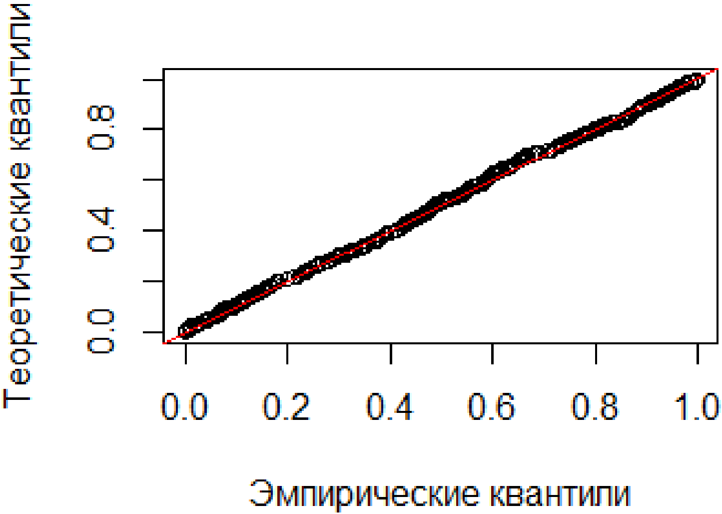
$$Q_e = (Y - H\hat{\Theta})^{\mathrm{T}}\Lambda^{-1}(Y - H\hat{\Theta}), Q_g = (Y - H_b\hat{\beta})^{\mathrm{T}}\Lambda^{-1}(Y - H_b\hat{\beta}).$$

Тогда

$$\mathbb{E}Q_e = \sigma^2(m_{..} - m_{.1} - I(T-1)),$$

$$\mathbb{E}Q_g = \sigma^2(m_{..} - m_{.1} - T + 1) + \frac{m_{..}}{(I-1)(T-1)}\sigma_g^2,$$

При верной $H_0 : \sigma_g^2 = 0$ имеет место
$\frac{(Q_g - Q_e)(m_{..} - m_{.1} - I(T-1))}{Q_e((I-1)(T-1))} \sim F((I-1)(T-1), m_{..} - m_{.1} - I(T-1)).$


Теоретические квантили
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
Эмпирические квантили

- 280 наркозависимых пациентов,
- Индекс тяжести в двух временных точках,
- Группировка по методам лечения: препарат Налтрексон, плацебо и антидепрессанты.

Оценки эффектов взаимодействия:

		t_1	t_2
Naltrexone	*Placebo*	-0.007	0.019
Placebo	*Placebo*	-0.012	0.046
Naltrexone	*Antidepressant*	0.014	-0.067
Placebo	*Antidepressant*	0.005	-0.057

При проверке гипотезы $\sigma_g^2 = 0$ получаем $p - value = 0.059$, таким образом нет оснований отвергать нулевую гипотезу со стандартным уровнем значимости.

Итоги

В рамках работы было выполнено следующее :

- Построена статистика для эффекта взаимодействия в случае пропущенных данных в модели со случайными эффектами,
- Произведена проверка статистик моделированием.