Эргодический метод компенсации пропусков в дисперсионном анализе повторных наблюдений с приложением в фармакологии

Уфлянд Анна Григорьевна, гр. 522

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

Научный руководитель: к.ф.-м.н., доц. Алексеева Н.П.
Рецензент: мл. научн. сотр. Ананьевская П.В.

Санкт-Петербург
2013г.

Лонгитюдные данные. Пример из фармакологии

- Эксперимент – программа отказа от употребления наркотиков
- Признак: индекс депрессии Бека
- Индивиды – больные героиновой наркоманией ($I = 4$)
- Временные точки ($T = 8$) – из наблюдений с 2-х недельным интервалом в течение 6 месяцев
- Пропусков 33%, полные данные в начале исследования

group	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7	Time 8
1	19	7	3	0	0	0	1	0
1	24	18	15	4	NA	5	7	13
4	17	3	1	0	3	0	NA	NA
4	15	NA	14	NA	NA	NA	NA	NA
1	14	4	1	0	0	0	0	0
1	22	5	NA	3	NA	2	0	0
2	0	0	0	0	NA	NA	NA	NA
2	21	NA	NA	NA	NA	NA	NA	NA
3	18	8	13	3	3	1	1	4
1	14	NA	NA	NA	NA	NA	NA	NA
1	23	15	15	10	6	NA	9	NA
3	18	5	15	8	11	8	10	4
3	0	0	0	NA	NA	NA	NA	NA
1	9	2	2	1	2	1	1	NA
1	20	NA	NA	NA	NA	NA	NA	NA
3	26	27	23	18	15	10	10	5
3	20	18	NA	NA	NA	NA	NA	NA
3	29	29	11	12	20	7	0	NA
4	10	7	2	0	3	NA	NA	NA

Рис. : Изменение индекса депрессии Бека во времени

Дисперсионный анализ для повторных наблюдений

$$x_{ijt} = \mu + \alpha_i + \varepsilon^1_{ij} + \beta_t + \gamma_{it} + \varepsilon_{ijt}, \text{ где}$$

μ генеральное среднее,
α_i фиксированный эффект группы,
β_t фиксированный эффект времени,
γ_{it} эффект взаимодействия группы и времени,

$$i = 1, \ldots, I; \quad j = 1, \ldots, \nu_i; \quad \sum_{i=1}^{I} \nu_i = N; \quad t = 1, \ldots, T;$$

$\varepsilon^1_{ij} \sim N(0, \sigma_1^2), \quad \varepsilon_{ijt} \sim N(0, \sigma^2)$ – независимые ошибки .

[Афифи А., Эйзен С., 1982]

Методы для анализа данных с пропусками

Метод	**Особенности**
Исключение индивидов с пропусками в данных	Мало данных Потеря информации
LOCF (метод протягивания последнего имеющегося наблюдения) [R.M.Hamer, 2009]	Ложное увеличение количества степеней свободы для статистик критериев Смещения в оценках эффектов
Эргодический метод	Без заполнения пропусков Не привносится искусственная информация

Модель эргодического метода

Напомним, что модель имеет вид $x_{ijt} = \mu + \alpha_i + \varepsilon^1_{ij} + \beta_t + \gamma_{it} + \varepsilon_{ijt}$.
Для оценки параметров её разделяют: $x_{ijt} = x_{ij.} + (x_{ijt} - x_{ij.})$,

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

Пусть N_{ij} и $M_{it}-$ множества полных наблюдений в группе i для индивида j и в временной точке t соответственно,
$n_{ij} = \#N_{ij},\ m_{it} = \#M_{it}$

- Смещённость модели
  $\mathbb{E}x_{ij.} \neq \mu + \alpha_i$, где $x_{ij.} = \frac{1}{n_{ij}} \sum\limits_{t \in \mathbf{N}_{ij}} x_{ijt}$
- Решение проблемы
  введение индивидуальных смещений H_{ij}

Модель эргодического метода

Определение

Следующее равенство задаёт ***смещение*** *для индивида j из группы i,*

$$H_{ij} = \sum_{k=1}^{\infty} A_{ij}(k),$$
$$A_{ij}(1) = \frac{1}{n_{ij}} \sum_{t \in N_{ij}} \frac{1}{m_{it}} \sum_{l \in M_{it}} (x_{ilt} - x_{il.}),$$
$$A_{ij}(k+1) = \frac{1}{n_{ij}} \sum_{t \in N_{ij}} \frac{1}{m_{it}} \sum_{l \in M_{it}} A_{il}(k).$$

Тогда введение поправок в $x_{ijt} = x_{ij.} + (x_{ijt} - x_{ij.})$ в виде:

$$x_{ij.} - H_{ij} = \mu + \alpha_i + \Theta^1_{ij}$$
$$x_{ijt} - x_{ij.} + H_{ij} = \beta_t + \gamma_{it} + \Theta_{ijt}$$

приводит к тому, что ошибки $\Theta^1_{ij}, \Theta_{ijt}$ — коррелированные с нулевыми мат.ожиданиями. [Н.П.Алексеева, 2012]

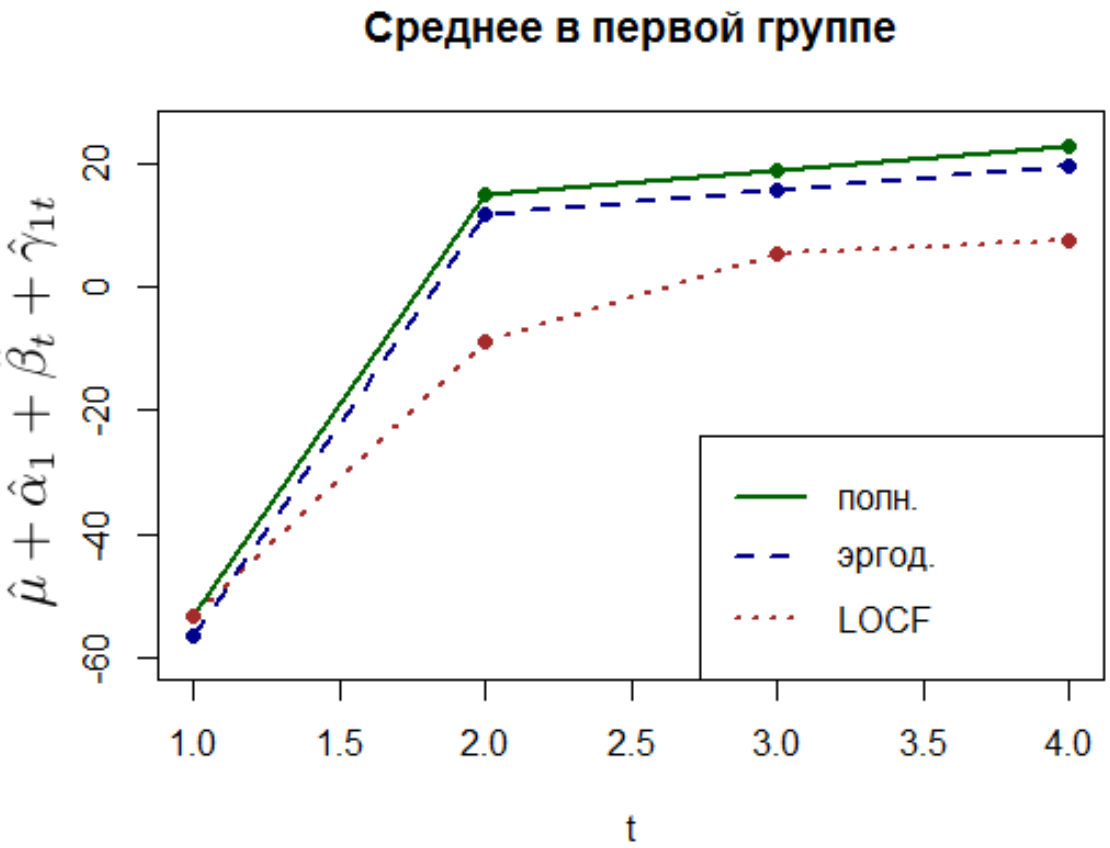
Постановка задачи

1. Сравнение с наиболее распостранёнными методами на конкретных примерах.
2. Исследование свойств метода: несмещённости оценок, инвариантности выборочных характеристик, эргодичности.
3. Получение условий, при которых улучшается эргодичность матрицы наблюдений.

Сравнение методов на модельных данных

Несмещённость эффектов взаимодействия

- Две группы по 75 индивидов, 4 временные точки.
- Пропуски во временных точках $\sim Bin(p),\ p = 0.3, 0.5, 0.7$.



Сравнение методов на реальных данных с модельными пропусками

Эффект фактора группы

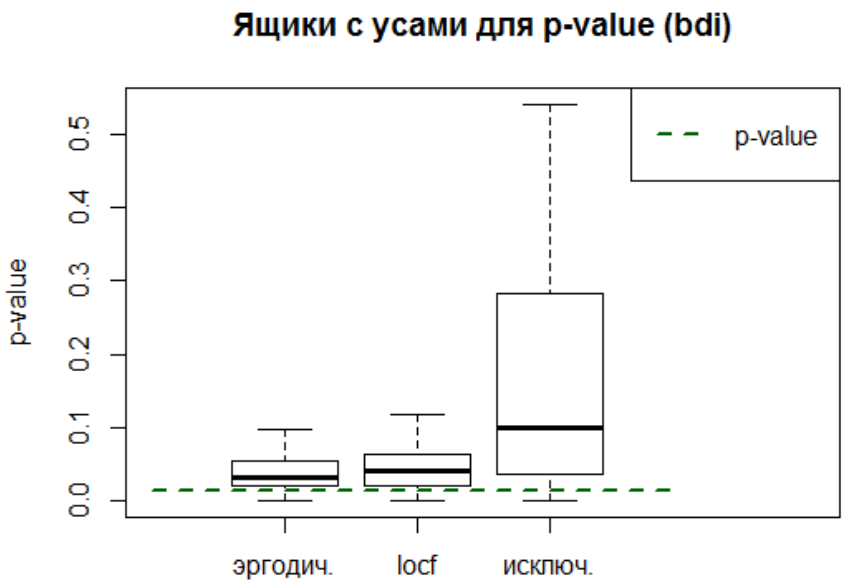


Рис. : Ящики с усами для p-value. Признак bdi

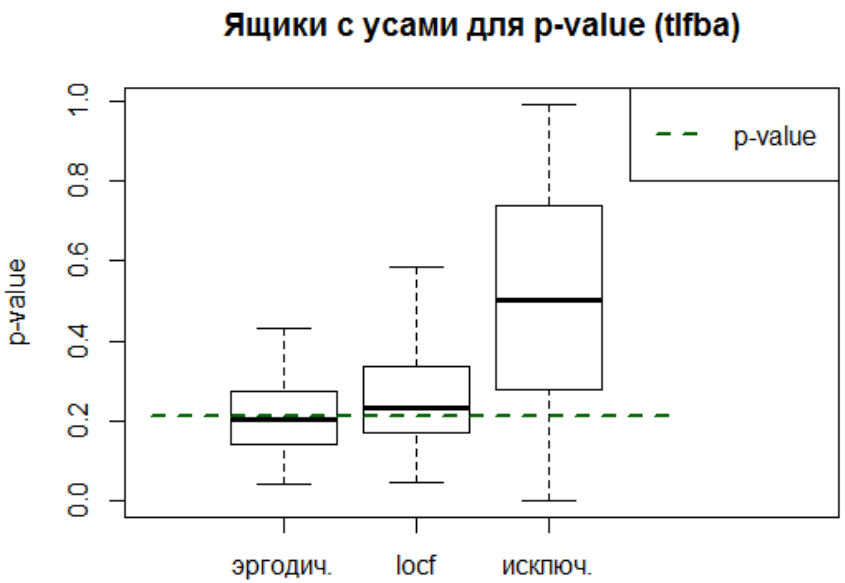
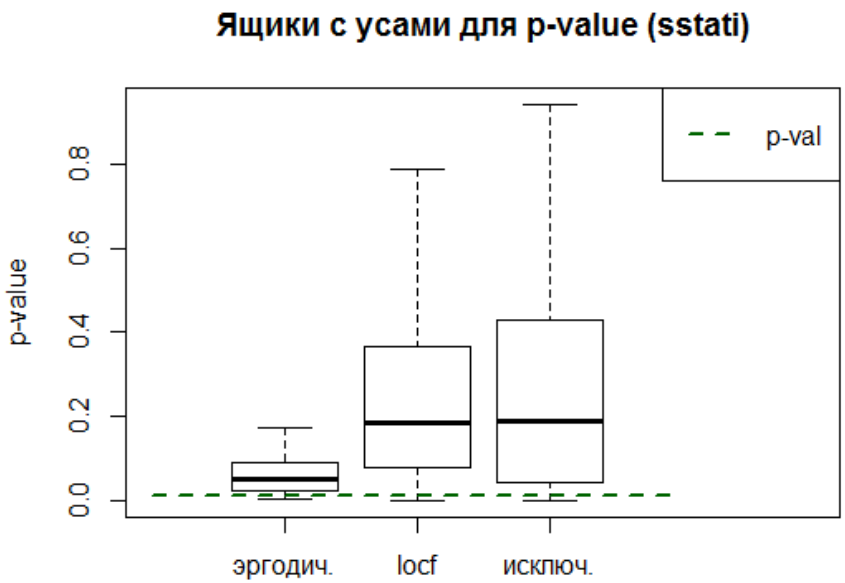


Рис. : Ящики с усами для p-value. Признак tlfba

Сравнение методов на реальных данных с модельными пропусками

Эффект взаимодействия факторов группы и времени



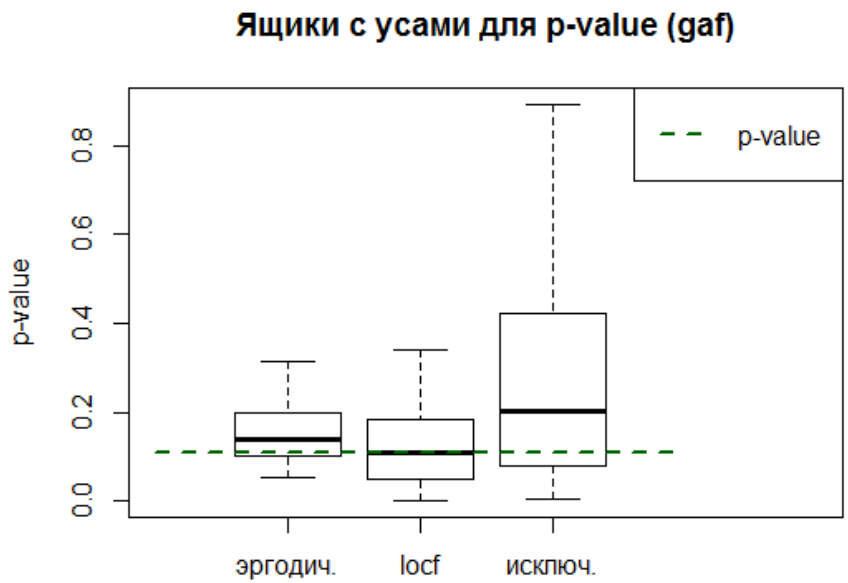


Рис. : Ящики с усами для p-value. Признак sstati

Рис. : Ящики с усами для p-value. Признак gaf

Сравнение методов на реальных данных с модельными пропусками

Таблица : Доли ложно не выявленных (ЛН) и ложно обнаруженных (ЛО) значимостей для эффектов α_i, γ_{it}

	эффект	**эргодический**	**LOCF**	**исключение**
ЛН	α_i	0.289	0.351	0.701
	γ_{it}	0.536	0.818	0.736
ЛО	α_i	0.018	0.009	0.061
	γ_{it}	0.000	0.255	0.182

Векторное представление для смещения

Лемма

Пусть вектор $H = (H_1, \ldots, H_N)$, $\bar{H} = \frac{1}{N} \sum_{j=1}^{N} H_j$

матрица $P = \Lambda_N J \Lambda_T J^{\mathbf{T}}$,
$\Lambda_N -$ *диагональная матрица* $N \times N$ *с элементами* $\frac{1}{n_j}$ *на главной диагонали,*
$\Lambda_T -$ $T \times T$ *с элементами* $\frac{1}{m_t}$, $J -$ *матрица инцидентности.*
Тогда

$$H = \sum_{k=0}^{\infty} P^k A(1),$$

P *– стохастическая,* $\lim_{k \to \infty} P^k = P^{\infty}$,
P^{∞} *состоит из строк* $\pi = \left(\frac{n_1}{m_.}, \cdots, \frac{n_N}{m_.} \right)$.

[Н.П.Алексеева, 2012]

Свойства баланса и инвариантности

Лемма

Свойство баланса для перекрёстных усреднений:

$$\forall k : \frac{1}{N}\sum_{j=1}^{N} n_j A_j(k) = \frac{1}{N}\sum_{j=1}^{N} n_j A_j(1) = 0.$$

Лемма

Свойство баланса для смещения: $\frac{1}{N}\sum_{j=1}^{N} n_j H_j = 0.$

Теорема

Пусть общее среднее в одной группе $x_{..} = \frac{1}{m_.}\sum_{j=1}^{N}\sum_{t\in N_j} x_{jt}$,

тогда оно инвариантно относительно вычитания смещения

$$x'_{..} = \frac{1}{m_.}\sum_{j=1}^{N}\sum_{t\in N_j}(x_{jt} - H_j) = x_{..}.$$

Эргодическое свойство в одной группе ($I = 1$)

У эргодических систем среднее по пространству совпадает со средним по времени

- Средние по пространству и по времени

$$x_* = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{t\in N_j} x_{jt}, \qquad x^* = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m_t}\sum_{j\in M_t} x_{jt}.$$

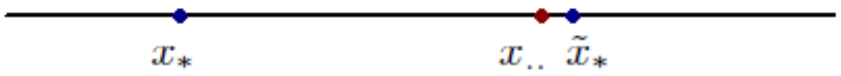
- Общее среднее $x_{..} = \frac{1}{m_.}\sum_{j=1}^{N}\sum_{t\in N_j} x_{jt}$.

Зададим **эргодическое свойство** как $x_* = x_{..} = x^*$.

Определение

Будем говорить, что эргодичность улучшилась, если

$|x_* - x_{..}| > |\tilde{x}_* - x_{..}|$, *где* $\tilde{x}_* = x_* - \bar{H}$ *и* $\bar{H} = \frac{1}{N}\sum_{j=1}^{N} H_j$.



Влияние тренда на улучшение эргодичности

Теорема

Пусть модель имеет вид: $x_{jt} = \mu + \beta_t + \delta_j + \varepsilon_{jt}$,

$Q = \{q_{jk}\}_{j=1,k=1}^{N} = (I - P + P^{\infty})^{-1}$, $\bar{\beta} = \frac{1}{N}\sum\limits_{j=1}^{N}\frac{1}{n_j}\sum\limits_{t\in N_j}\beta_t$.

Имеет место разделение $\bar{H} = \xi - \eta$ *на зависящую и не зависящую от времени компоненты, где*

$$\xi = x_* - x_{..} = \bar{\beta} + \delta_{.} + \varepsilon_{..} - \frac{1}{m_{.}}\sum_{j=1}^{N} n_j\delta_j - \frac{1}{m_{.}}\sum_{j=1}^{N} n_j\varepsilon_{j.},$$

$$\eta = \tilde{x}_* - x_{..} = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N} q_{jk}(\delta_k - \delta_{.}(k) + \varepsilon_{k.} - \varepsilon_{..}(k)).$$

Моменты компонент среднего смещения

Лемма

Математические ожидания $\bar{H}$ *и его составляющих равны*

$$\mathbb{E}\bar{H} = \bar{\beta}, \quad \mathbb{E}\xi = \bar{\beta}, \quad \mathbb{E}\eta = 0\,.$$

Лемма

Пусть $Q_k = \sum\limits_{j=1}^{N} q_{jk}$; σ^2 *и* σ_1^2 *– дисперсии ошибок* $\varepsilon_{jt}, \delta_j$, *тогда*

$$\mathbb{D}\xi = \sum_{j=1}^{N} \left(\frac{\sigma^2}{n_j} + \sigma_1^2\right) \left(\frac{1}{N} - \frac{n_j}{m_\cdot}\right)^2$$

$$\mathbb{D}\eta = \frac{\sigma_1^2}{N^2} \sum_{k=1}^{N} Q_k^2 \left(\sum_{t \in N_k} \sum_{l \in M_t} \frac{1}{m_t n_l} - 1\right)^2 +$$

$$+ \frac{\sigma^2}{N^2} \sum_{k=1}^{N} \sum_{t \in N_k} \left(\frac{1}{m_t} \sum_{l \in M_t} \frac{Q_l}{n_l} - \frac{Q_k}{n_k}\right)^2 .$$

Влияние тренда на улучшение эргодичности

Теорема

Пусть $\tilde{\sigma}_1^2 = \mathbb{D}\xi, \quad \tilde{\sigma}_2^2 = \mathbb{D}\eta.$

$$\textit{Тогда } \mathbb{P}\{|\xi| \le |\eta|\} \le \frac{2}{k^2} \textit{ при } k \in \left(\sqrt{2}; \frac{|\bar{\beta}|}{\tilde{\sigma}_1 + \tilde{\sigma}_2}\right).$$

Таблица : $n = 100,\ T = 4,\ \sigma = 0.1,\ \sigma_1 = 0.2,\ N = 1000.$

$(\beta_2, \beta_3, \beta_4)$	$\hat{\mathbb{P}}\{\lvert\xi\rvert \le \lvert\eta\rvert\}$
(0.01, 0.02, 0.03)	0.433
(0.05, 0.06, 0.07)	0.251
(0.07, 0.08, 0.09)	0.195
(0.10, 0.15, 0.20)	0.044
(0.15, 0.20, 0.25)	0.024
(0.10, 0.20, 0.30)	0.006
(0.20, 0.30, 0.40)	0.001
(0.30, 0.40, 0.50)	0.000

Основные результаты

- Получены и доказаны свойства метода: свойства баланса для перекрёстных усреднений и смещения, инвариантность общего среднего, а также ряд вспомогательных утверждений.
- Получены условия, при которых улучшается эргодичность матрицы наблюдений.
- Обнаружен более короткий способ вычисления дисперсий компонент смещения.
- Алгоритм метода реализован в виде программного кода на языке R.

Спасибо за внимание!