Проверка гипотез в дисперсионном анализе повторных неполных наблюдений

Н.П. Алексеева, Е.В.Вербицкая

СПбГУ, математико-механический факультет
СПбГМУ, лаб.биомедицинской статистики

2018 г.

Структура данных с повторностями

Пусть в i-й группе, $i = 1, 2, \ldots, I$ имеется ν_i индивидов, $\nu_1 + \ldots + \nu_I = n$, которые набдюдаются в моменты времени $t = 1, 2, \ldots, T$. Выборку наблюдений обозначим через x_{ijt}.

Модель

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

- μ параметр генерального среднего,
- α_i дифф.эффекты групп,
- β_t дифф. эффекты фактора времени,
- γ_{it} дифф. эффекты взаимодействия факторов времени и разделения на группы,
- $e^1_{ij} \sim \mathcal{N}(0, \sigma_1)$, $e_{ijt} \sim \mathcal{N}(0, \sigma)$ независимые случайные ошибки.

Случай неполных данных

Для каждого индивида имеется множество наблюдений в точках $N_{ij} \subset (1, 2, \ldots, T)$, $\mathrm{card}(N_{ij}) = n_{ij}$.
В i-й группе моменты в времени t имеются данные об индивидах $M_{it} \subset (1, 2, \ldots, \nu_i)$, $\mathrm{card}(M_{it}) = m_{it}$,

$$\sum_{i=1}^{I} m_{it} = m_{i\cdot}, \quad \sum_{t=1}^{T} m_{it} = m_{\cdot t}, \quad \sum_{t=1}^{T}\sum_{i=1}^{I} m_{it} = m_{\cdot\cdot}$$

Замечание

Необходимо предположение о наличии полных данных хотя бы в одной точке, и $\forall t, i\ m_{it} > 0$.

Оценка параметров модели

В случае полных данных устраняется индивидуальное влияние, измеряемое индивидуальным средним $x_{ij.}$ и модель разделяется на две модели:

$$x_{ij.} = \mu + \alpha_i + e^1_{ij}, \quad x_{ijt} - x_{ij.} = \beta_t + \gamma_{it} + e_{ijt}$$

Лемма

В случае неполных данных имеет место смещение

$$Ex_{ij.} = \mu + \alpha_i + W_{ij}, \text{ где } W_{ij} = \frac{1}{n_{ij}} \sum_{t \in N_{ij}} (\beta_t + \gamma_{it})$$

Групповая поправка

Определим групповую поправку как
$G = \sum_{i=0}^{\infty}(MN)^i(ML - MNK)$, где вектора $L = \{x_{..t} - x_{...}\}_{t=1}^{T}$,
$K = \{x_{i..} - x_{...}\}_{i=1}^{I}$, матрицы $M = \{m_{i1}, \ldots, m_{iT}\}_{i=1}^{I}$,
$N = \{m_{1t}, \ldots, m_{it}\}_{t=1}^{T}$.

Лемма

Пусть $P_0 = MN$, $Q_0 = (I - P_0 + P_0^{\infty})^{-1}$, $A = P_0Q_0$, $B = Q_0M$.
Тогда компонента групповой поправки имеет вид

$$G_i = \frac{1}{m_{i\cdot}} \sum_{t=1}^{T} m_{it}\beta_t + \varepsilon_i, \quad \varepsilon_i = \sum_{k=1}^{I}\sum_{t=1}^{T} m_{kt}\left(\frac{b_{it}}{m_{t\cdot}} - \frac{a_{ik}}{m_{k\cdot}}\right) e_{i\cdot t}.$$

Замечание

В случае однородности пропусков групповая поправка равна нулю.

Индивидуальная поправка

Пусть J^i матрица инцидентности i-й группы, Λ_{iT} и Λ_{ν_i} диагональные матрицы с векторами (m_{it}) и (ν_{ij}) на главной диагонали.
Вектор индивидуальной поправки $H = \{H_{ij}\}_{i=1,j=1}^{I,\ \nu_i}$, имеет вид $H = (I - C)V + DU$, где $V_i = \{x_{ij\cdot}\}_{j=1}^{\nu_i}$, $U_i = \{x_{i\cdot t}\}_{t=1}^{T}$,
$R_i = \Lambda_{\nu_i} J^i$, $P_i = R_i \Lambda_{iT} (J^i)^T$, $P_i^\infty = \mathrm{limit}_{k\to\infty} P^k$,
$Q_i = (I - P_i + P_i^\infty)^{-1}, C_i = P_i^\infty + Q_i, D_i = Q_i R_i$. .

Предложение

$$H_{ij} = W_{ij} - G_i + \mathcal{E}_{ij},$$

$$\mathcal{E}_{ij} = e_{ij\cdot} - \sum_{l=1}^{\nu_i} c_{jl}^i e_{il\cdot} + \sum_{\tau=1}^{T} d_{j\tau}^i e_{i\cdot\tau}$$

Объединенная поправка и ее следствие

Теорема

Пусть $\Delta_{ij} = e_{ij\cdot} - \mathcal{E}_{ij} - \varepsilon_i$. Имеют место несмещенные модели

$$y_{ijt} = x_{ijt} - x_{ij\cdot} + (H_{ij} + G_i) = \beta_t + \gamma_{it} + e_{ijt} - \Delta_{ij},$$

$$z_{ij} = x_{ij\cdot} - (H_{ij} + G_i) = \alpha_i + e^1_{ij} + \Delta_{ij},$$

$$\tilde{D}_{ij,kl} = \frac{1}{\sigma^2}\mathbb{E}\Delta_{ij}\Delta_{kl} = F_0(i,k) + I_{[i=k]}\left(\sum_{\iota=1}^{\nu_i}\frac{c^i_{j\iota}c^k_{l\iota}}{n_{i\iota}} + \sum_{t=1}^{T}\frac{d^i_{jt}d^k_{lt}}{m_{it}}\right) -$$

$$-I_{[i=k]}\sum_{\iota=1}^{\nu_i}\sum_{\tau\in N_{i\iota}}\left(\frac{c^i_{j\iota}d^k_{l\tau}}{n_{i\iota}m_{k\tau}} + \frac{d^i_{j\tau}c^k_{l\iota}}{n_{k\iota}m_{i\tau}}\right)),$$

$$F_0(i,k) = \sum_{t=1}^{T}\frac{b_{it}b_{kt}}{m_{\cdot t}} - \sum_{\iota=1}^{I}\frac{a_{i\iota}a_{k\iota}}{m_{\iota\cdot}}.$$

Ковариации ошибок

$$\mathbb{E}(e^1_{ij} + \Delta_{ij})(e^1_{kl} + \Delta_{kl}) = \sigma_1^2 I_{[i=k,j=l]} + \sigma^2 \tilde{D}_{ij,kl}\,.$$

Определим ковариационную матрицу ошибок
$\Sigma = \mathbb{E}(e_{ijt} - \Delta_{ij})(e_{kl\tau} - \Delta_{kl}) = \sigma^2 \Lambda.$

$$\Sigma = \sigma^2 (I_{[i=k,j=l,t=\tau]} - M_{ij,kl\tau} - M_{kl,ijt} + \tilde{D}_{ij,kl})\,,$$

$$M_{ij,kl\tau} = \frac{1}{\sigma^2}\mathbb{E}\Delta_{ij}e_{kl\tau} = I_{[i=k]}\left(\frac{c^i_{jl}}{n_{il}} - \frac{d^i_{j\tau}}{m_{i\tau}}\right) + \left(\frac{a_{ik}}{m_{k.}} - \frac{b_{i\tau}}{m_{.\tau}}\right).$$

Матрица плана H модели $\mathbb{E}y_{ijt} = \beta_t + \gamma_{it}$

$$q_{it} = -\frac{m_{it}}{m_{iT}},\quad q_{\cdot t} = -\frac{m_{\cdot t}}{m_{\cdot T}}, p_{it} = -\frac{m_{it}}{m_{It}},\quad p_{i\cdot} = -\frac{m_{i\cdot}}{m_{I\cdot}},\quad r_{it} = \frac{m_{it}}{m_{IT}}$$

	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
$Y_{1,T}$	$q_{\cdot 1}$	q_{11}	0	$\dots$	0	$\dots$	$q_{\cdot\tau}$	$q_{1,\tau}$	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Y_{\iota,1}$	1	0	0	$\dots$	1	$\dots$	0	0	0	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Y_{\iota,\tau}$	0	0	0	$\dots$	0	$\dots$	1	0	0	$\dots$	1
$Y_{\iota,T}$	$q_{\cdot 1}$	0	0	$\dots$	$q_{\iota,1}$	$\dots$	$q_{\cdot\tau}$	0	0	$\dots$	$q_{\iota,\tau}$
$Y_{I,1}$	1	p_{11}	p_{21}	$\dots$	$p_{\iota,1}$	$\dots$	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	$p_{1,\tau}$	$p_{2,\tau}$	$\dots$	$p_{\iota,\tau}$
$Y_{I,T}$	$q_{\cdot 1}$	r_{11}	r_{21}	$\dots$	$r_{\iota,1}$	$\dots$	$q_{\cdot\tau}v$	$r_{1,\tau}$	$r_{2,\tau}$	$\dots$	$r_{\iota,\tau}$

Проверка значимости эффектов взаимодействия

$\Theta = (\beta_1, \gamma_{11}, \ldots, \gamma_{\iota,1}, \ldots, \beta_{\mathrm{T}-1}, \gamma_{1,\mathrm{T}-1}, \ldots, \gamma_{\iota,\mathrm{T}-1})$
Оценка $\hat{\Theta} = (\mathrm{H}^{\mathrm{T}}\Lambda^{-1}\mathrm{H})^{-}\mathrm{H}^{\mathrm{T}}\Lambda^{-1}\mathrm{Y}$ при минимизации

$$\mathrm{Q}_{2\mathrm{e}}(\Theta) = (\mathrm{Y} - \mathrm{H}\Theta^{\mathrm{T}})^{\mathrm{T}}\Lambda^{-1}(\mathrm{Y} - \mathrm{H}\Theta^{\mathrm{T}})\,.$$

Число степеней свободы $\mu_{2\mathrm{e}} = \mathrm{m}_{\cdot\cdot} - \mathrm{n} - \mathrm{I}(\mathrm{T}-1)$.
Вектор усеченной модели параметров $\Theta_* = (\beta_1, \ldots, \beta_{\mathrm{T}-1})^{\mathrm{T}}$
Оценка $\hat{\Theta}_* = (\mathrm{H}_*^{\mathrm{T}}\Lambda^{-1}\mathrm{H}_*)^{-}\mathrm{H}_*^{\mathrm{T}}\Lambda^{-1}\mathrm{Y}$ при минимизации

$$\mathrm{Q}_{2\mathrm{e}*}(\Theta_*) = (\mathrm{Y} - \mathrm{H}_*\Theta_*^{\mathrm{T}})^{\mathrm{T}}\Lambda^{-1}(\mathrm{Y} - \mathrm{H}_*\Theta_*^{\mathrm{T}}),$$

число степеней свободы равно
$\mu_{2\mathrm{e}*} - \mu_{2\mathrm{e}} = \mathrm{m}_{\cdot\cdot} - \mathrm{n} - \mathrm{T} + 1 - (\mathrm{m}_{\cdot\cdot} - \mathrm{n} - \mathrm{I}(\mathrm{T}-1)) = \mathrm{IT} - \mathrm{T} - \mathrm{I} + 1.$

$$\mathrm{F} = \frac{(\mathrm{Q}_{2\mathrm{e}*}(\hat{\Theta}_*) - \mathrm{Q}_{2\mathrm{e}}(\hat{\Theta}))(\mathrm{m}_{\cdot\cdot} - \mathrm{n} - \mathrm{I}(\mathrm{T}-1))}{\mathrm{Q}_{2\mathrm{e}}(\hat{\Theta})(\mathrm{IT} - \mathrm{T} - \mathrm{I} + 1)} \sim \mathcal{F}(\mu_{2\mathrm{e}*} - \mu_{2\mathrm{e}}, \mu_{2\mathrm{e}})$$

Пример из фармакологии

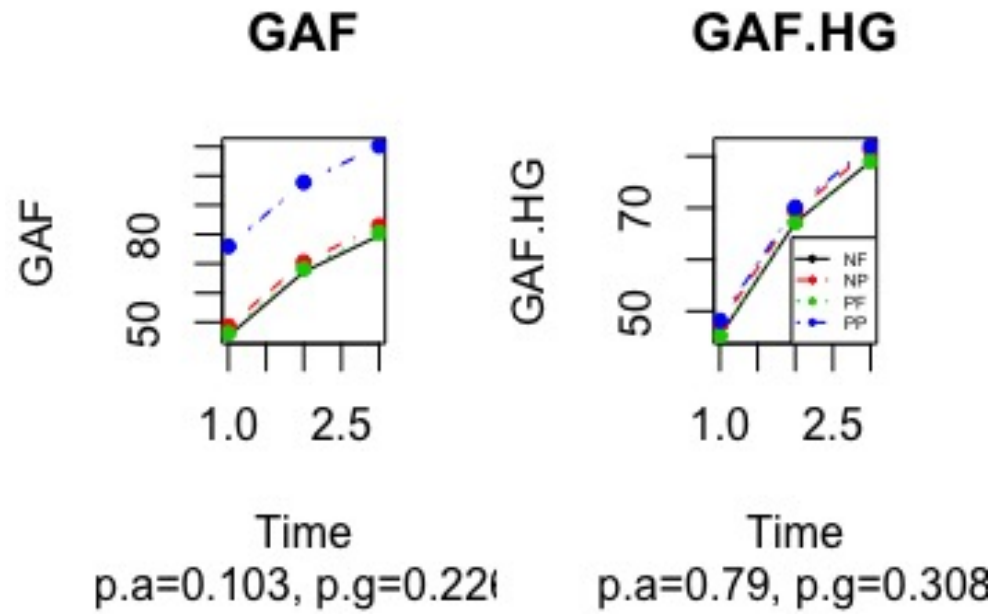

Figure: Динамика индекса глобального функционирования (ИГФ) при разной терапии при удалении наблюдений с пропусками (n = 63) и с учетом всей информации (n = 332) .

Различия по ИГФ не было, но у больных с изначально высоким ИГФ было больше шансов остаться в программе даже при отсутствии терапии.

Алгоритм построения линейного предиктора по неполным данным $\mathrm{F}(\mathcal{Y}, \mathcal{X})$

$\Omega = (\omega_1, \ldots, \omega_n)$ наблюдения по неполным данным:
зависимой переменной $\mathcal{Y}(\Omega) = (y(\omega_1), \ldots, y(\omega_n))$,
независимых переменных $\mathcal{X} = (X_1(\Omega), \ldots, X_p(\Omega))$.

Модель линейной регрессии $Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$

- $Y = \mathcal{Y}(\Omega_1)$, $\Omega_1 \subseteq \Omega = (\omega_1, \ldots, \omega_n)$;
- $X = (X_{t_1}(\Omega_1), \ldots, X_{t_k}(\Omega_1))$, $(t_1, \ldots, t_k) \subseteq (1, \ldots, p)$
- $\hat{\beta}, \hat{\sigma^2}$ оценки МНК

Предиктор $\hat{Y}(\Omega) = X_* \hat{\beta}$, где $X_* = (X_{t_1}(\Omega), \ldots, X_{t_k}(\Omega))$

Метод заполнения пропусков Predictive Mean Matching

Алгоритм замены пропущенных значений наблюдаемыми, наиболее близкими к предсказанным

Актуальность темы
для статистического анализа многомерных данных с повторностями нужна редукция размерности
для этого используется факторный анализ