# О проверке гипотез в дисперсионном анализе повторяющихся неполных наблюдений [1]

Алексеева Н.П., доцент кафедры статистического моделирования СПбГУ, nina.alekseeva@spbu.ru

Федорченко С.А., консультант в Insilico Medicine, fedorchenko.ser.eja@yandex.ru

## Аннотация

В модели дисперсионного анализа для повторяющихся наблюдений (ANOVA Repeated Measures ) при условии неполных данных во временных точках кроме первой решается задача централизации модели, приводящей к коррелированности ошибок. В соответствии со структурой инцидентности наблюдений вычисляется ковариационная матрица ошибок. Статистическая модель с коррелированными ошибками используется для проверки значимости отклонения от нуля дифференциальных эффектов. В аналогичной модели со случайными эффектами для факторов повторения и их взаимодействия вычисляются математические ожидания статистик, необходимых для построения отношения Фишера. Показано, что для проверки значимости эффекта ваимодействия используется одна и та же статистика вне зависимости от дифференциального или случайного характера данного эффекта. Для проверки значимости случайного фактора повторения получен дополнительный критерий в виде равенства, справедливого, в частности, при условии одинакового числа наблюдений в ячейке.

## Введение

Дисперсионный анализ повторяющихся наблюдений, в литературе известный под названием ANOVA Repeated Measures (AVRM) – один из самых востребованных методов анализа лонгитюдных данных [1], позволяющий в рамках одной статистической модели осуществлять проверку значимости динамики наблюдений, главных эффектов и эффек-

[1] Работа поддержана грантом РФФИ: 20-01-00096.

тов их взаимодействия с фактором времени. Проблема возникает, если данные неполные. Не умаляя достоинств методов, предлагающих заполнение пропущенных данных, остановимся на эргодическом методе, предложенном автором в [2], примененном для анализа неполных кардиологических наблюдений в [3], представленном на международной конференции [4]. Идея метода заключается в централизации модели через введение индивидуальной поправки и дальнейшем пересчете ковариационной матрицы ошибок. В [5] было предложено усовершенствование модели за счет введения групповой поправки, необходимой в случае неравномерности распределения пропущенных данных в отдельных группах. Несмотря на то, что удалось довести исследование до числа и получить адекватные результаты, осталось несколько нераскрытых тем. Так ранее не обсуждалось то, на чем основана идея построения поправок. Кроме того не рассмотрены возможности применения случайных эффектов, и актуальным остается тестирование работы полученных критериев на модельных данных.

## Модель AVRM с дифференциальными эффектами

Модель дисперсионного анализа для повторяемых наблюдений AVRM с дифференциальными эффектами можно представить в виде

$$x_{ijk} = \mu + \alpha_i + \delta_{ij} + \beta_k + \gamma_{ik} + e_{ijk}, \tag{1}$$

где $x_{ijk}$ наблюдение $j$- го индивида из группы $i$ в момент времени $t$, $\mu$ параметр генерального среднего, $\alpha_i$, $\beta_k, \gamma_{ik}$ дифференциальные эффекты, соответствующие группе $i = 1, 2, \ldots, I$, временной точке $k = 1, 2, \ldots, T$ и эффектам взаимодействия факторов времени и группы. Ошибки $\delta_{ij}$, $e_{ijk}$ предполагаются независимыми нормально распределенными величинами с нулевыми средними и соответственно дисперсиями $\sigma_1^2$ и $\sigma^2$. Обозначим через $M_{it}$ номера индивидов из группы $i$, имеющие наблюдения в момент времени $t$, а через $m_{it}$ их количество, $m_{i\cdot} = \sum_{t=1}^{T} m_{it}$, $m_{\cdot t} = \sum_{i=1}^{I} m_{it}$, $m_{\cdot\cdot} = \sum_{t=1}^{T} m_{t\cdot}$. Пусть $N_{ij}$ множество временных точек индивида $j$ из группы $i$, и $n_{ij}$ их количество. Ограничения на параметры выбираем в соответствии с частичным планом $\sum_{i=1}^{I} \frac{\alpha_i m_{i\cdot}}{m_{\cdot\cdot}} = 0$, $\sum_{t=1}^{T} \frac{\beta_t m_{\cdot t}}{m_{\cdot\cdot}} = 0$, $\sum_{i=1}^{I} \frac{\gamma_{it} m_{it}}{m_{\cdot\cdot}} = 0$, $\sum_{t=1}^{T} \frac{\gamma_{it} m_{it}}{m_{\cdot\cdot}} = 0$ [6]. Оценки параметров модели (1) с учетом этих ограничений $\hat{\alpha}_i = x_{i\cdot\cdot} - x_{\cdots}$, $\hat{\beta}_t = x_{\cdot\cdot t} - x_{\cdots}$, $\hat{\gamma}_{it} = x_{i\cdot t} - x_{i\cdot\cdot} - x_{\cdot\cdot t} + x_{\cdots}$ выражаются через разнообразные формы усреднения наблюдений: индивидуальное

среднее для $j$-го индивида из $i$-й группы $x_{ij\cdot}$, среднее $x_{\cdot\cdot t}$ в заданный момент времени $t$, внутригрупповое среднее $x_{i\cdot\cdot}$, общее среднее $x_{\cdot\cdot\cdot\cdot}$. В случае полных данных проверку гипотез для параметров модели (1) можно свести к проверке гипотез для двух моделей с математическими ожиданиями $\mathbb{E}x_{ij\cdot} = \mu + \alpha_i$, $\mathbb{E}(x_{ijk} - x_{ij\cdot}) = \beta_k + \gamma_{ik}$. В случае неполных данных эти модели становятся смещенными. В этом можно убедиться, доказав следующее утверждение.

**Лемма 0.1** *Если существует индивид $j$ из группы $i$ такой, что $n_{ij} \neq T$, то для индивидуального среднего имеет место смещение $\mathbb{E}x_{ij\cdot} = \mu + \alpha_i + W_{ij}$, где $W_{ij} = \frac{1}{n_{ij}} \sum_{t \in N_{ij}} (\beta_t + \gamma_{it})$.*

Из оценок параметров получим $\hat{\beta}_t + \hat{\gamma}_{it} = x_{i\cdot t} - x_{i\cdot\cdot}$ и оценим смещение

$$\hat{W}_{ij} = \frac{1}{n_{ij}} \sum_{t \in N_{ij}} (\hat{\beta}_t + \hat{\gamma}_{it}) = \frac{1}{n_{ij}} \sum_{t \in N_{ij}} \frac{1}{m_{it}} \sum_{l \in M_{it}} (x_{ilt} - x_{il\cdot}),$$

$$\mathbb{E}(x_{ij\cdot} - \hat{W}_{ij}) = \mu + \alpha_i + \frac{1}{n_{ij}} \sum_{t \in N_{ij}} \frac{1}{m_{it}} \sum_{l \in M_{it}} W_{ij}.$$

Можно повторить еще раз эту процедуру, рассмотрев математическое ожидание разности $\mathbb{E}\left( x_{ij\cdot} - \hat{W}_{ij} - \frac{1}{n_{ij}} \sum_{t \in N_{ij}} \frac{1}{m_{it}} \sum_{l \in M_{it}} \hat{W}_{ij} \right)$. Опять получим какое-то смещение меньше предыдущего, и так далее. В общем виде, обозначим через $A_{ij}(0) = \frac{1}{n_{ij}} \sum_{t \in N_{ij}} \frac{1}{m_{it}} \sum_{l \in M_{it}} (x_{ilt} - x_{il\cdot})$ и определим рекуррентным образом последовательность $A_{ij}(k+1) = \frac{1}{n_{ij}} \sum_{t \in N_{ij}} \frac{1}{m_{it}} \sum_{l \in M_{it}} A_{ij}(k)$. В качестве индивидуального смещения будем рассматривать $H_{ij} = \sum_{k=1}^{\infty} A_{ij}(k)$. В [3] доказано, что $H_{ij}$ конечно при условии полноты наблюдений хотя бы в один момент времени, вычислено его математическое ожидание. Это позволило использовать $H_{ij}$ в качестве индивидуальной поправки при небольшом ограничении, связанном с равномерностью пропущенных наблюдений по группам, которое в [5] было снято посредством введения дополнительной поправки.

## Поправки модели при неполноте данных

Оценку параметров общей модели (1) можно получить посредством введения двух несмещенных моделей, для которых

$$\mathbb{E}X_{ij} = \mu + \alpha_i \text{ и } \mathbb{E}(x_{ijk} - X_{ij}) = \beta_k + \gamma_{ik}\,. \tag{2}$$

В случае полных данных используется $X_{ij} = x_{ij\cdot}$, а в случае неполных данных [5] нужно использовать поправки. Пусть $J^i$ матрица инцидентности в $i$-й группе, $\Lambda_{iT}$ и $\Lambda_{\nu_i}$ диагональные матрицы с векторами на главной диагонали вида $(m_{it})_{t=1}^T$ и $(n_{ij})_{j=1}^{\nu_i}$, $R_i = \Lambda_{\nu_i} J^i$, $P_i = R_i \Lambda_{iT} (J^i)^T$ – стохастическая матрица с ненулевыми компонентами при условии наличия полных данных хотя бы в одной точке, $P_i^\infty = \lim\limits_{k\to\infty} P^k$, $Q_i = (I - P_i + P_i^\infty)^{-1}$, векторы $V_i = \{x_{ij\cdot}\}_{j=1}^{\nu_i}$, $U_i = \{x_{i\cdot t}\}_{t=1}^T$, $L = \{x_{\cdot\cdot t} - x_{\cdot\cdot\cdot}\}_{t=1}^T$, $K = \{x_{i\cdot\cdot} - x_{\cdot\cdot\cdot}\}_{i=1}^I$ и матрицы

$$M = \left\{ \frac{m_{i1}}{m_{i\cdot}}, \frac{m_{i2}}{m_{i\cdot}}, \ldots, \frac{m_{iT}}{m_{i\cdot}} \right\}_{i=1}^{I}, \quad N = \left\{ \frac{m_{1t}}{m_{\cdot t}}, \ldots, \frac{m_{it}}{m_{\cdot t}} \right\}_{t=1}^{T}. \tag{3}$$

Тогда для $i$-й группы вектор индивидуальной поправки $H_i$ с компонентами $\{H_{ij}\}_{i=1,j=1}^{I,\ \nu_i}$ вводится как

$$H_i = (I - C_i)V_i + D_i U_i, \quad \text{где} \quad C_i = P_i^\infty + Q_i, D_i = Q_i R_i. \tag{4}$$

Групповые поправки $G_i$ определяются как компоненты вектора $G = \sum_{i=0}^\infty (MN)^i (ML - MNK)$.

**Теорема 0.1** *[5] Пусть $M$, $N$ матрицы из (3) и $P_0 = MN$ стохастическая матрица, $Q_0 = (I - P_0 + P_0^\infty)^{-1}$, $A = P_0 Q_0$, $B = Q_0 M$, $a_{ik}$ и $b_{it}$ соответственно элементы этих матриц и $d^i_{j\tau}$ элементы матриц $C_i$, $D_i$ из (4). Будем рассматривать два вида наблюдений*

$$X_{ij} = x_{ij\cdot} - (H_{ij} + G_i) \quad u \quad y_{ijt} = x_{ijt} - X_{ij}. \tag{5}$$

*Пусть $\Delta_{ij} = e_{ij\cdot} - \mathcal{E}_{ij} - \varepsilon_i$, $\mathcal{E}_{ij} = H_{ij} - \mathbb{E}H_{ij}$, $\varepsilon_i = G_i - \mathbb{E}G_i$. Тогда*

$$\mathcal{E}_{ij} = e_{ij\cdot} - \sum_{l=1}^{\nu_i} c^i_{jl} e_{il\cdot} + \sum_{\tau=1}^{T} d^i_{j\tau} e_{i\cdot\tau}\,,\ \varepsilon_i = \sum_{k=1}^{I}\sum_{t=1}^{T} m_{kt} \left( \frac{b_{it}}{m_{t\cdot}} - \frac{a_{ik}}{m_{k\cdot}} \right) e_{i\cdot t}\,,$$

*и имеют место несмещенные модели вида*

$$X_{ij} = \mu + \alpha_i + e^1_{ij} + \Delta_{ij} \quad u \quad y_{ijt} = \beta_t + \gamma_{it} + e_{ijt} - \Delta_{ij}\,. \tag{6}$$

Ковариационная матрица $\sigma^2\Lambda$ ошибок $\varepsilon_{ijt} = e_{ijt} - \Delta_{ij}$ получена в [5].

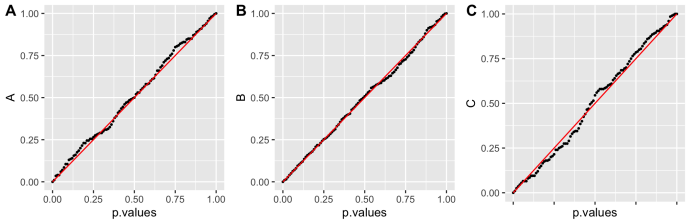


Рис. 1: Распределение $p$-значений для проверки значимости дифференциальных эффектов факторов: $A$ группы, $B$ времени, $C$ взаимодействия. Объем выборки $m_{\cdot 1} = 72$, три временные точки, две группы, 6% процентов пропущенных данных.

## Модель с дифференциальными эффектами

Рассмотрим элементы $y_{ijt} = x_{ijt} - X_{ij}$ из второго уравнения (5) $y_{ijt} = \beta_t + \gamma_{it} + \varepsilon_{ijt}$, для каждой пары $i, t$ соберем $m_{it}$ элементов $y_{ijt}$ в векторе $Y_{it}$ и обозначим через $Y$ вектор $(Y_{11}, Y_{12}, \ldots, Y_{1T},\ Y_{21}, Y_{22},\ \ldots, Y_{2T} \ldots\ Y_{I1}, Y_{I2}, \ldots, Y_{IT})$. Согласно (5), вектор $Y = \{y_{ijk}\}$ состоит из $m_{\cdot\cdot}$ компонент, из которых линейно независимые $n = m_{\cdot\cdot} - m_{\cdot 1}$, поэтому ранг матрицы $\Lambda$ равен $n$.

Матрицу частичного плана обозначим через $X = [X_1|X_2]$, где $X_1, X_2$ соответственно матрицы плана усеченных моделей, вектор ошибок $\varepsilon_{ijk}$ через $\varepsilon$, вектор параметров через $\Theta = (\beta_1, \ldots, \beta_{T-1}, \gamma_{1,1}, \ldots, \gamma_{1,T-1}, \ldots, \gamma_{I-1,1}, \ldots, \gamma_{I-1,T-1})$. В матричном виде модель $y_{ijt} = \beta_t + \gamma_{it} + \varepsilon_{ijt}$ имеет вид $Y = X\Theta + \varepsilon$, где $\mathbb{E}\varepsilon\varepsilon^T = \sigma^2\Lambda$. При помощи ортогонального преобразования преобразуем эту модель к виду $V = H\Theta + \delta$, где $\delta$ независимые ошибки. Минимизируя выражение $(V - H\Theta)^T(V - H\Theta)$, получаем остаточные суммы квадратов

$$
\begin{aligned}
R_0 = (V - H\hat{\Theta})^T(V - H\hat{\Theta}) = \delta^T(\mathbb{I} - H(H^TH)^-H^T)\delta = \delta^T A\delta\,,\\
R_1 = (V - H_1\hat{\beta})^T(V - H_1\hat{\beta}) = (H_2\gamma + \delta)^T A_1(H_2\gamma + \delta),\\
R_2 = (V - H_2\hat{\gamma})^T(V - H_2\hat{\gamma}) = (H_1\beta + \delta)^T A_2(H_1\beta + \delta),
\end{aligned}
\qquad (7)
$$

рассматривая $A_i = \mathbb{I}_n - H_1(H_1^TH_1)^-H_1^T$, $i = 0, 1, 2$, $A_0 = A$. Кроме того $R_2 - R_0 = (H_1\beta + H_2\gamma + \delta)^T(\mathbb{I} - A_1)(H_1\beta + H_2\gamma + \delta)$.

Используя обратное ортогональное преобразование, можно выразить компоненты $R_0$, $R_1$, $R_2$ через матрицы плана $X = X_0$, $X_1$, $X_2$ и матрицу $\Lambda$. В результате имеем $R_i = Y^TB_iY$, где $i = 0, 1, 2$, $B_i = \Lambda^{-1} - \Lambda^{-1}X_i(X_i^T\Lambda^{-1}X_i)^-X_i^T\Lambda^{-1}$. Далее применяются стандартное отношение Фишера. На рис.1 представлены согласованные с равно-

мерным распределением функции распределения $p$-значений, полученных при проверке значимости отклонения от нуля дифференциальных эффектов на модельных данных с числом итераций 400.

## Проверка значимости случайных эффектов

В отличии от модели (1) эффекты факторов времени и взаимодействия будем считать случайными, то есть

$$x_{ijk} = \mu + \alpha_i + \delta_{ij} + b_k + g_{ik} + e_{ijk}, \tag{8}$$

где $b_k$, $g_{ik}$ независимые нормально распределенные центрированные случайные величины с дисперсиями $\sigma_b^2$ и $\sigma_g^2$ соответственно. Вычислим математические ожидания статистик (7), полученных в анализе модели с дифференциальными эффектами,

$$\begin{aligned} \mathbb{E}R_0 &= \sigma^2 \mathbf{Tr}(A) = \sigma^2(n - I(T-1)), \\ \mathbb{E}R_1 = \mathbb{E}(H_2 g + \delta)^T A_1 (H_2 g + \delta) &= \mathbb{E}g^T H_2^T A_1 H_2 g + \mathbb{E}\delta^T A_1 \delta = \\ = \mathbf{Tr}(H_2^T A_1 H_2)\sigma_g^2 + \mathbf{Tr}(A_1)\sigma^2\,,&\ \text{где } \mathbf{Tr}A_1 = n - (T-1)\,. \\ \mathbb{E}(R_1 - R_0) &= \mathbf{Tr}(H_2^T A_1 H_2)\sigma_g^2 + (I-1)(T-1)\sigma^2\,. \end{aligned}$$

Отсюда видно, для проверки гипотезы $H_0 : \sigma_g^2 = 0$ можно применить стандартный критерий, а в случае $H_0 : \sigma_b = 0$ для построения статистики Фишера нужно выполнение равенства $(I-1)\mathbf{Tr}(H_2^T(\mathbb{I} - A_1)H_2) = \mathbf{Tr}(H_2^T A_1 H_2)$, которое справедливо в случае одинакового числа наблюдений $m_{it} = m$ в каждой точке и в каждой группе. Если это равенство не выполняется, то нужно исследовать возможность выравнивания коэффициентов через оценки дисперсий $\sigma_g^2$ и $\sigma^2$, но этот вопрос является предметом уже дальнейшего исследования.

## Заключение

В большинстве статистических моделей изучения влияния факторов на зависимую переменную метрического типа является закономерным требование одинакового числа наблюдений в ячейке. В анализе повторяющихся наблюдений это требование удалось сузить до более приемлемого требования наличия полных данных хотя бы в одной точке за счет корректировки индивидуального среднего по времени в зависимости от числа имеющихся наблюдений. Вычисление корреляционной

матрицы ошибок позволило построить статистики для проверки значимости отклонения от нуля дифференциальных эффектов. Моделирование свидетельствует об адекватности предлагаемого метода проверки гипотез.

Предпринята попытка применить данный подход в модели со случайными эффектами. Как и ожидалось, статистика для проверки значимости случайного эффекта взаимодействия совпала с соответствующей статистикой в модели с дифференциальными эффектами. Для проверки значимости случайного эффекта фактора времени приведено равенство, при выполнении которого допустимо построение отношения Фишера. В дальнейшем предполагается исследовать область допустимых отклонений от этого равенства для получения приближенных результатов и удостовериться в адекватности данного подхода при увеличении числа дополнительных факторов.

## Литература


[1] Longitudinal Data Analysis // Ed. by G. Fitzmaurice, M. Davidian, G. Verbeke, G. Molenberghs. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Boca Raton: Chapman and Hall/CRC, 2008.

[2] Alexeyeva N.P., Tatarinova A.A., Bondarenko B.B. et al. (2011). Analysis of repeated cardiological incomplete data based on ergodic centralization of model. Bulleten of Almazov Center, 3(8):59–63.

[3] Alexeyeva N. (2013). Analysis of biomedical systems. Reciprocity. Ergodicity. Synonymy. Publishing of the Saint-Petersburg State University, Saint-Petersburg.

[4] Ufliand A., Alexeyeva N. (2014). The dependence of the ergodicity on the time effect in the repeated measures anova with missing data based on the unbiasedness recovery. In Topics in Statistical Simulation, Springer Proceedings in Mathematics and Statistics 114, pages 517–527. Springer Science + Business Media New York 2014.

[5] Alexeyeva N. (2017) Dual balance correction in repeated measures ANOVA with missing data January Electronic Journal of Applied Statistical Analysis 10(1):146-159 DOI:10.1285/i20705948v10n1p146

[6] Scheffe H. (1999). The Analysis of Variance. John Wiley and Sons, Canada.