# Статистические модели анализа повторных наблюдений

Редкокош Кирилл Игоревич

гр. 21.М03-мм
Санкт-Петербургский государственный университет
Кафедра статистического моделирования
**Научный руководитель: к. ф.-м. н., доцент Алексеева Нина Петровна**
**Рецензент: к. т. н., доцент Белякова Людмила Анатольевна**

9 июня 2023 г.

# Введение. Цель и задачи

Целью данной работы является исследование методов анализа повторных наблюдений в условиях неполных данных, в частности адаптация метода **MANOVA Repeated Measures в случае неполных данных**.

# Введение. Цель и задачи

Целью данной работы является исследование методов анализа повторных наблюдений в условиях неполных данных, в частности адаптация метода **MANOVA Repeated Measures в случае неполных данных**.

Среди задач можно выделить:

- Рассмотрение теоретических аспектов темы;
- Систематизация материала;
- Моделирование для проверки точности критерия;
- Апробация на реальных данных.

# Общий вид данных

Многомерные повторные наблюдения с внутригрупповым временным фактором B и межгрупповым фактором группы A:

| Фактор A (группа) | Объекты | Фактор B(повторные наблюдения) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $B_1$ | | | $B_2$ | | | | $B_t$ | | |
| | | $Y_{11}$ | $\dots$ | $Y_{p1}$ | $Y_{12}$ | $\dots$ | $Y_{p2}$ | $\dots$ | $Y_{1t}$ | $\dots$ | $Y_{pt}$ |
| | $S_{11}$ | $y_{1111}$ | $\dots$ | $y_{11p1}$ | $y_{1112}$ | $\dots$ | $y_{11p2}$ | $\dots$ | $y_{111t}$ | $\dots$ | $y_{11pt}$ |
| | $S_{12}$ | $y_{1211}$ | $\dots$ | $y_{12p1}$ | $y_{1212}$ | $\dots$ | $y_{12p2}$ | $\dots$ | $y_{121t}$ | $\dots$ | $y_{12pt}$ |
| $A_1$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | | $\vdots$ |
| | $S_{1n_1}$ | $y_{1n_111}$ | $\dots$ | $y_{1n_1p1}$ | $y_{1n_112}$ | $\dots$ | $y_{1n_1p2}$ | $\dots$ | $y_{1n_1t}$ | $\dots$ | $y_{1n_1pt}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $S_{g1}$ | $y_{g111}$ | $\dots$ | $y_{g1p1}$ | $y_{g112}$ | $\dots$ | $y_{g1p2}$ | $\dots$ | $y_{g11t}$ | $\dots$ | $y_{g1pt}$ |
| $A_g$ | $S_{g2}$ | $y_{g211}$ | $\dots$ | $y_{g2p1}$ | $y_{g212}$ | $\dots$ | $y_{g2p2}$ | $\dots$ | $y_{g21t}$ | $\dots$ | $y_{g2pt}$ |
| | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | | $\vdots$ |
| | $S_{1n_g}$ | $y_{gn_g11}$ | $\dots$ | $y_{gn_gp1}$ | $y_{gn_g12}$ | $\dots$ | $y_{gn_gp2}$ | $\dots$ | $y_{gn_g1t}$ | $\dots$ | $y_{gn_gpt}$ |

# Описание модели

Рассмотрим один признак.
Модель для $j$-го индивида из $i$-ой группы в $k$-ый момент времени имеет вид:

$$y_{ijk} = \mu + \alpha_i + \epsilon^1_{ij} + \beta_k + \gamma_{ik} + \epsilon_{ijk},$$

$\mathbf{y}_{11}, \mathbf{y}_{12}, ..., \mathbf{y}_{1n}$ является выборкой из $N_p(\mu_1, \mathbf{\Sigma})$,
$\mathbf{y}_{21}, \mathbf{y}_{22}, ..., \mathbf{y}_{2n}$ — так же из $N_p(\mu_2, \mathbf{\Sigma})$ и это верно для всех $g$ групп.

# Проверяемые гипотезы

В модели отражено влияние двух факторов: группы и времени, и взаимодействия этих факторов. Как следствие мы имеем для проверки три гипотезы.
Первая из них– гипотеза об отсутствии **эффекта группы**:

$$H_{0_A} : \alpha_i = 0 \; \forall i = 1, \ldots, g.$$

Вторая– гипотеза об отсутствии **эффекта времени**:

$$H_{0_B} : \beta_k = 0 \; \forall k = 1, \ldots, t.$$

Третья– гипотеза об отсутствии **эффекта взаимодействия группы и времени**:

$$H_{0_{AB}} : \gamma_{ik} = 0 \; \forall i = 1, \ldots, g, \; \forall k = 1, \ldots, t.$$

# Групповая поправка

Как известно из литературы [**Alexeyeva N., 2017**] для борьбы с неполными данными вводятся **индивидуальная** и **групповая** поправки.

Введем $\mathbf{M}$ и $\mathbf{N}$:

$m_{ik}$– количество наблюдений в $i$-ой группе в момент времени $k$.

$$\mathbf{M} = \mathbf{M}_{(I,T)} \begin{bmatrix} \frac{m_{11}}{m_{1.}} & \dots & \frac{m_{1T}}{m_{1.}} \\ \vdots & \dots & \vdots \\ \frac{m_{I1}}{m_{I.}} & \dots & \frac{m_{IT}}{m_{I.}} \end{bmatrix} \mathbf{N} = \mathbf{N}_{(T,I)} \begin{bmatrix} \frac{m_{11}}{m_{.1}} & \dots & \frac{m_{I1}}{m_{.1}} \\ \vdots & \dots & \vdots \\ \frac{m_{1T}}{m_{.T}} & \dots & \frac{m_{IT}}{m_{.T}} \end{bmatrix}$$

$\mathbf{P}_0 = \mathbf{MN}, \mathbf{P}_0^{\infty} = \lim\limits_{n \to \infty} \mathbf{P}_0^n$– стационарная матрица.

Вектор **групповой поправки**:

$$G = \sum_{i=0}^{\infty} \mathbf{P}_0^i (\mathbf{M}L - \mathbf{P}_0 K), \text{ где}$$

$$L = (y_{..1} - y_{...}, \dots, y_{..T} - y_{...})^T, K = (y_{1..} - y_{...}, \dots, y_{I..} - _{...})^T$$

# Индивидуальная поправка

Обозначения: матрица инциденций– $\mathbf{J}_i$; диагональная матрица $\mathbf{\Lambda}_{\nu_i}$ с элементами $\frac{1}{n_{ij}}$; диагональная матрица $\mathbf{\Lambda}_{iT}$ с элементами $\frac{1}{m_{it}}$; матрица $\mathbf{R}_i = \mathbf{\Lambda}_{\nu_i}\mathbf{J}_i$ и матрицу $\mathbf{P}_i = \mathbf{R}_i\mathbf{\Lambda}_{iT}(\mathbf{J}_i)^T$;
$U_i = \{y_{i.t}\}_{t=1}^{T}, V_i = \{y_{ij.}\}_{j=1}^{\nu_i}$.

$A_i(k) = \mathbf{P}_i A_i(k-1)$, где $A_i(0) = \mathbf{R}_i U_i - \mathbf{P}_i V_i$.

Вектор **индивидуальной поправки** $i$-ой группы определим следующим образом:

$$H_i = \Sigma_{k=1}^{\infty} A_i(k)$$

# Полученные результаты. Применимость метода

Из литературы также известно, что метод применим, если хоть в один момент времени есть наблюдения для всех индивидов. В работе было доказано более общее утверждение:

Так как стохастическая матрица $\mathbf{P}_i$ является матрицей переходных вероятностей для цепи Маркова, то вопрос применимости данного метода эквивалентен вопросу о регулярности матрицы $\mathbf{P}_i$.

**Теорема**

*Если в каждой паре строк исходной матрицы данных найдётся хоть одна пара непропущенных значений и нет индивидов, у которых пропущены все значения, то $\mathbf{P}_i$– регулярная стохастическая матрица.*

# Получение линейной комбинации

Приведём многомерный случай к одномерному, для этого рассмотрим линейную комбинацию исходных признаков в $k$-ый $(k = 1, \ldots, t)$ момент времени:

$$Z_k = \mathbf{Y}_k A = a_1 Y_{1k} + \cdots + a_p Y_{pk}, \text{ где } A = (a_1, \ldots, a_p)^T,$$

где $a_1, \ldots, a_p$ – неизвестные коэффициенты линейной комбинации, которые нам и предстоит найти.
В результате введения «новых» признаков $Z_k$ мы приходим к модели одномерного дисперсионного анализа.

# Получение наиболее значимой комбинации. Численное решение

Для получения линейной комбинации исходных признаков с коэффициентами $a_1, \ldots, a_p$, для которой достигается наибольшее различие индивидов необходимо решить следующую задачу:

- Для **эффекта группы**:
  $F_A(a_1, \ldots, a_p) \to \max\limits_{a_1,\ldots,a_p}$ .
- Для **эффекта времени**:
  $F_B(a_1, \ldots, a_p) \to \max\limits_{a_1,\ldots,a_p}$ .
- Для **эффекта взаимодействия группы и времени**:
  $F_{AB}(a_1, \ldots, a_p) \to \max\limits_{a_1,\ldots,a_p}$ .

Где $F_A, F_B, F_{AB}$– статистики для проверки гипотез об отсутствии соответствующего эффекта.

# Получение наиболее значимой комбинации. Матричный метод

$$F = F(A) = \frac{A^T \mathbf{H} A / \nu_{\mathbf{H}}}{A^T \mathbf{E} A / \nu_{\mathbf{E}}} \sim F_{\nu_{\mathbf{H}}, \nu_{\mathbf{E}}}, \text{ где}$$

**H**— матрица межгрупповых отклонений, а **E**— внутригрупповых отклонений, $\nu_{\mathbf{H}}$ и $\nu_{\mathbf{E}}$ соответсвующие степени свободы.

Для нахождение признака максимально разделяющего наши данные получим задачу:

$$\frac{A^T \mathbf{H} A}{A^T \mathbf{E} A} \to \max_A$$

Решение: $A_1 = \arg\max \frac{A^T \mathbf{H} A}{A^T \mathbf{E} A}$, где $A_1$– собственный вектор $\mathbf{E}^{-1}\mathbf{H}$, который соответствует максимальному собственному числу– $\lambda_1$. Полученная линейная комбинация $\mathbf{Y}_k A_1$– первая каноническая переменная.

# Полученные результаты. Распределение p-value для Wilks's lambda

Рассмотрим $p - value$ как случайную величину:
$\alpha_{\mathbf{I}} = P_{H_0}(H_0$ отвергается$) =$
$\alpha \Leftrightarrow P_{H_0}(\alpha > p) = \alpha \Leftrightarrow P_{H_0}(p < \alpha) = \alpha$
Если верна $H_0$, то $p - value$ **равномерно распределено** на [0,1], критерий является **применимым**.

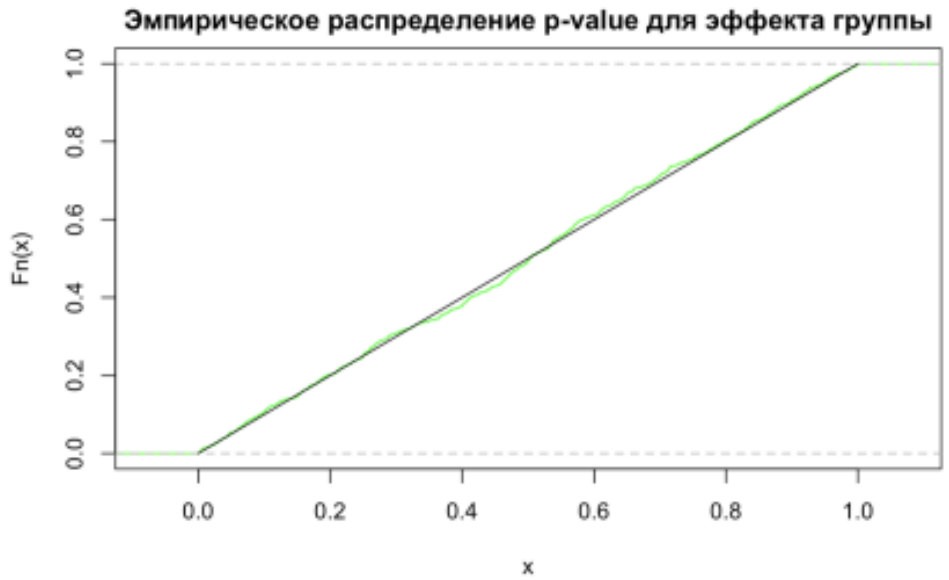
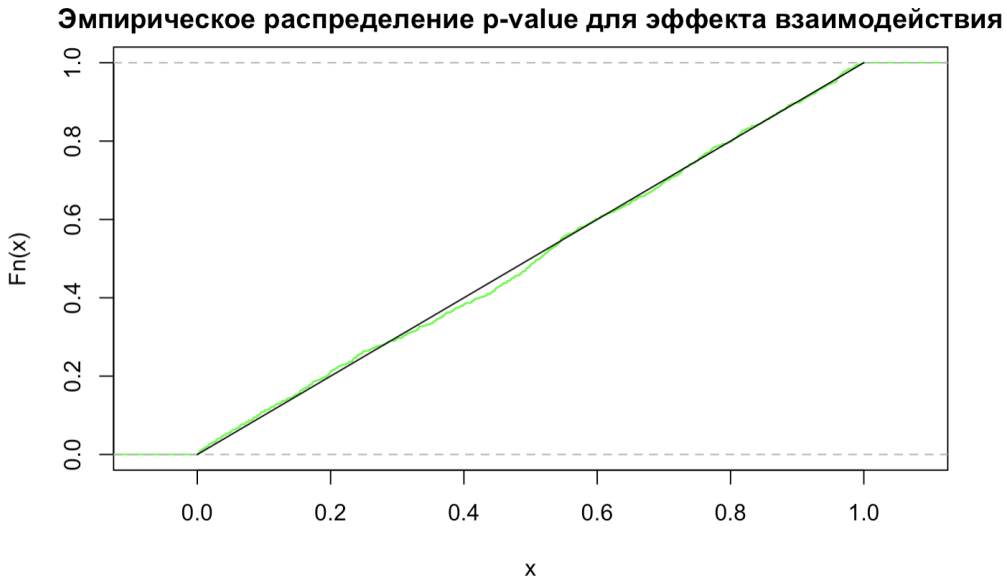



Рис. 3.4. График эмпирической функции распределения $p - value$ (эффект группы), 50% пропусков

# Полученные результаты. Одномерный случай для Roy's Largest root

Прикладное значение канонических переменных приводит к использованию их как отдельных признаков, что требует рассмотрения также статистики основанной на одном собственном числе– Roy's Largest root.




Критерий является **применимым** для одномерного дисперсионного анализа.

# Полученные результаты. Многомерный случай для Roy's Largest root

Перейдём к многомерному случаю, рассмотрим два признака ($n = 2$):

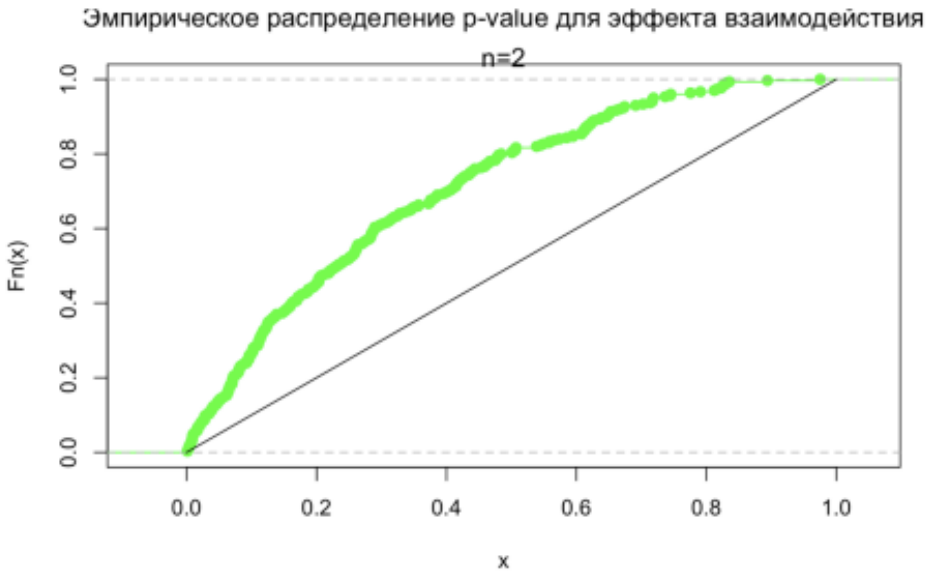



Рис. 3.13. График эмпирической функции распределения $p - value$ (эффект взаимодействия), $n = 2$

Как видно критерий является **радикальным**, что делает невозможным его применение.

# Полученные результаты. Нормирующий множитель для Roy's Largest root

Введем **нормирующий множитель**, имеющей вид:

$$\frac{1}{n \cdot \sqrt{log_m(n+1)}},$$

где $n$- это количество признаков, а $m$- количество моментов времени.

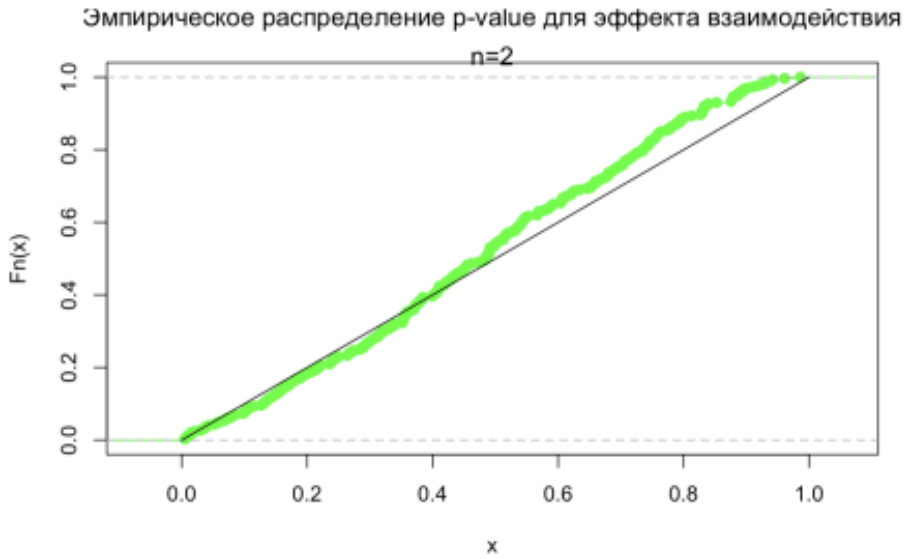



Рис. 3.27. График эмпирической функции распределения $p-value$ (эффект взаимодействия), $m = 3$, с поправкой

# Полученные результаты. Описание данных

Данные о $n = \mathbf{160}$ женщинах, которым проводилось лечение препаратами железа, данные разделены на $g = \mathbf{3}$ группы:

- 0– контрольная группа;
- 1– препараты вводились внутривенно;
- 2– препараты в форме таблеток.

Далее нас будут интересовать следующие признаки, измеренные в $t = \mathbf{5}$ моментов времени:
HB – гемоглобин, MCV – средний объём эритроцита, MCH – среднее содержание гемоглобина в отдельном эритроците, MCHC – средняя концентрация гемоглобина в эритроцитарной массе, HT – гематокрит, Lk – лейкоциты, Tr – тромбоциты, Er – эритроциты.

# Полученные результаты. Одномерный анализ

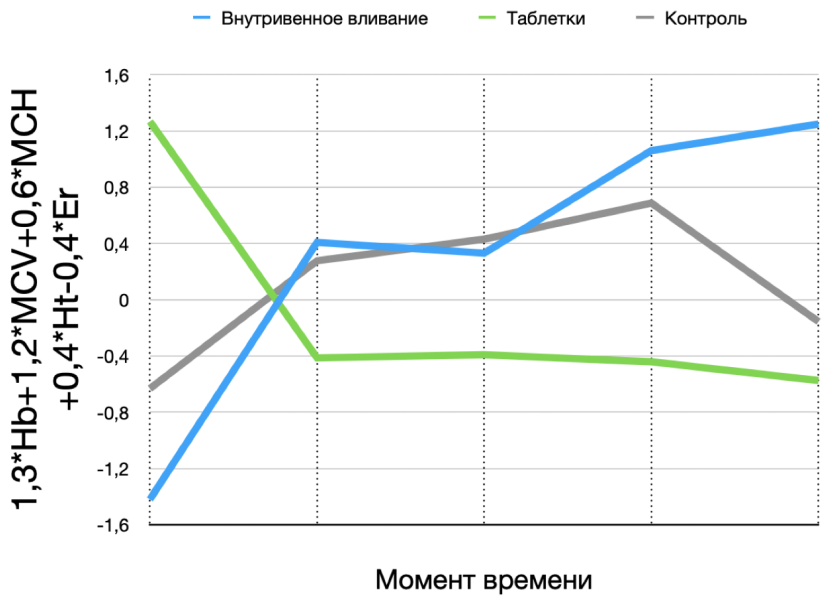
Проведём одномерный анализ каждой переменной. В столбцах таблицы приведены $p - value$ для соответствующих эффектов.

Таблица: Результат применения одномерного дисперсионного анализа

| Переменная | Эф.группы | Эф.времени | Эф.взаимодействия |
|---|---|---|---|
| Hb | 10e-16 | 0.684 | **10e-16** |
| MCV | 6e-5 | 0.215 | **8e-10** |
| MCH | 1e-5 | 0.294 | **1e-6** |
| MCHC | 0.003 | 0.441 | 0.89 |
| Ht | 10e-16 | 0.608 | **4e-10** |
| Er | 0.018 | 0.946 | **0.02** |
| Tr | 0.096 | 0.646 | 0.16 |
| Lk | 0.285 | 0.923 | 0.594 |

# Полученные результаты. Эффект взаимодействия: иллюстрация

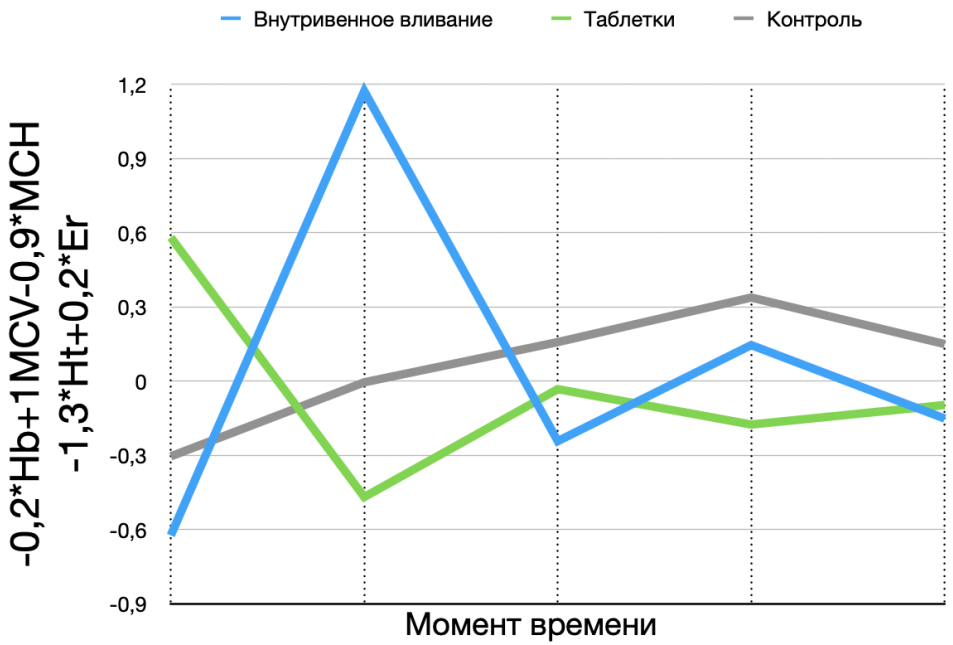
Рассмотрим первую каноническую переменную для эффекта взаимодействия $(p - value = 1e\text{-}4)$– 1,3*Hb+1,2*MCV+0,6*MCH +0,4*Ht-0,4*Er:




Интерпретация: при внутривенном вливание резко увеличивается гемоглобин, а также наблюдается увеличение качества эритроцитов в смысле наполняемости кислородом.

# Полученные результаты. Вторая компонента: иллюстрация

Рассмотрим вторую каноническую переменную для эффекта взаимодействия ($p - value$ = 1e-4)– -0,2*Hb+1*MCV-0,9*MCH-1,3*Ht+0,2*Er:




Интерпретация: компонента связана с избыточной жидкостью в организме, которая повышается при внутривенном вливания во второй точке.

# Полученные результаты. Заключение

В результате исследования были достигнуты следующие результаты:

- Доказана теорема об условиях применимости дисперсионного анализа в случае неполных данных;
- Реализован комплекс программ на языке программирования R;
- Предложен нормирующий множитель, позволяющий применять критерии;
- Метод апробирован на реальных данных;
- Получена линейная комбинация, соответствующая второму собственному числу.