# Частично-итерационный метод множественной регрессии для неполных данных с применением в биологии

Гриненко Юрий Константинович, группа 22.М03-мм

Санкт-Петербургский государственный университет
Математическое моделирование, программирование, искусственный интеллект
Научный руководитель: к. ф.-м.н., доцент Н. П. Алексеева
Рецензент: Волканова Маргарита Дмитриевна

Санкт-Петербург
2024г.

# Содержание работы

Дипломная работа состоит из трех основных частей:

1. Рассмотрение частично-итерационного метода множественной регрессии как инструмента для получения предсказаний;
2. Предложение процедуры отбора частных предсказаний, способной учитывать кластеры в многомерных данных, для построения ансамбля моделей;
3. Сравнительный анализ полученных результатов, выводы об изменении интерпретации модели;

# [1] Задачи множественной линейной регрессии

Модель: Модель задает отображение $f : \mathbb{X}^p \to \mathbb{Y}$;

$$f_i(x) = \sum_{i=1}^{p} \beta_i x_i + \varepsilon_i, \tag{1}$$

Оценки $\hat{\beta}$:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{2}$$

# [1] Частично-итерационная множественная регрессия

Частное предсказание: $\hat{f}(X_{\tau 1} \dots X_{\tau k})$ — предсказание по подможеству независимых переменных $\tau \subseteq \Omega_p = (1, 2, \dots, p)$.

Максимальное число моделей (и полученных векторов с предсказаниями) $N = 2^p - 1$; предсказания стандартизуются.

| | | |
|---|---|---|
| $\hat{f}_1$ | $\dots$ | $\hat{f}_N$ |
| $\hat{y}_{11}$ | $\dots$ | $\hat{y}_{N1}$ |
| $\dots$ | $\dots$ | $\dots$ |
| $\hat{y}_{1n}$ | $\dots$ | $\hat{y}_{Nn}$ |

Корреляция между $y$ и $j$-м частным предсказанием:

$$l_{0j} = \frac{1}{n} \sum_{\nu=1}^{n} y_\nu \hat{f}_{j\nu}, \qquad j = 1, \dots, N. \tag{3}$$

# [1] Частные предсказания по комбинациям признаков

Матрица вторых моментов: $\mathbf{L}$, элементы — корреляции частных предсказаний $\hat{f}_1, \ldots, \hat{f}_N$

$$\mathbf{L} = \begin{pmatrix} 1 & l_{12} & \ldots & l_{1N} \\ l_{21} & 1 & \ldots & l_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ l_{N1} & l_{N2} & \ldots & 1 \end{pmatrix}.$$

Модель матрицы $\mathbf{L}$ [Алексеева, Н. П., Ал-Джубури, Ф. С. Ш. (2022)]:

$$\begin{pmatrix} 1 & r+\epsilon_{12} & \ldots & r+\epsilon_{1N} \\ r+\epsilon_{21} & 1 & \ldots & r+\epsilon_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ r+\epsilon_{N1} & r+\epsilon_{N2} & \ldots & 1 \end{pmatrix},$$

$\mathbb{E}\epsilon_{ij} = 0$, $\mathbb{D}\epsilon_{ij} = \sigma^2$, $\epsilon_{ij} = \epsilon_{ji}$.

# [1] Взвешенное предсказание

Взвешенное предсказание: $\hat{f}_\Theta(\hat{f}_1, \ldots, \hat{f}_N)$
[Алексеева, Н. П., Ал-Джубури, Ф. С. Ш. (2022)];
Существует матрица вида $\mathbf{U}_N$:

$$\mathbf{U}_N = \begin{pmatrix} 1 & r & \ldots & r & r \\ r & 1 & \ldots & r & r \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r & r & \ldots & 1 & r \\ r & r & \ldots & r & 1 \end{pmatrix}.$$

$$\hat{f}_\Theta = \Theta(\sigma)(\frac{1}{N}\sum_{j=1}^{N} l_{0j}\hat{f}_j), \Theta(\sigma) = \frac{N(1-r)^{N-1} + NG_{N-1}(\sigma)}{|U_N| + NrG_{N-1}(\sigma) + G_N(\sigma)}, \quad (4)$$

$$\mathbb{D}\epsilon_{ij} = \sigma^2; \qquad G_N(\sigma) = \sum_{k=1}^{[\frac{N}{2}]} C_N^{N-2k}(-1)^k \phi_k \sigma^{2k} a^{N-2k}; \qquad \phi_k = \frac{(2k)!}{2^k k!}. \quad (5)$$

# [1] Сравнение результатов частно-итерационного метода множественной регрессии

Таблица 1. Коэффициенты детерминации общей модели $f_\Theta$ и моделей с исключением некомплектных признаков $f_1$ или некомплектных наблюдений $f_2$

| $R^2(\hat{f}_\Theta, Y)$ | $R^2(\hat{f}_1, Y)$ | $R^2(\hat{f}_2, Y)$ |
|---|---|---|
| 0.76 | 0.63 | 0.64 |

# [2] Линии регрессии при кластерной структуре данных; моделирование

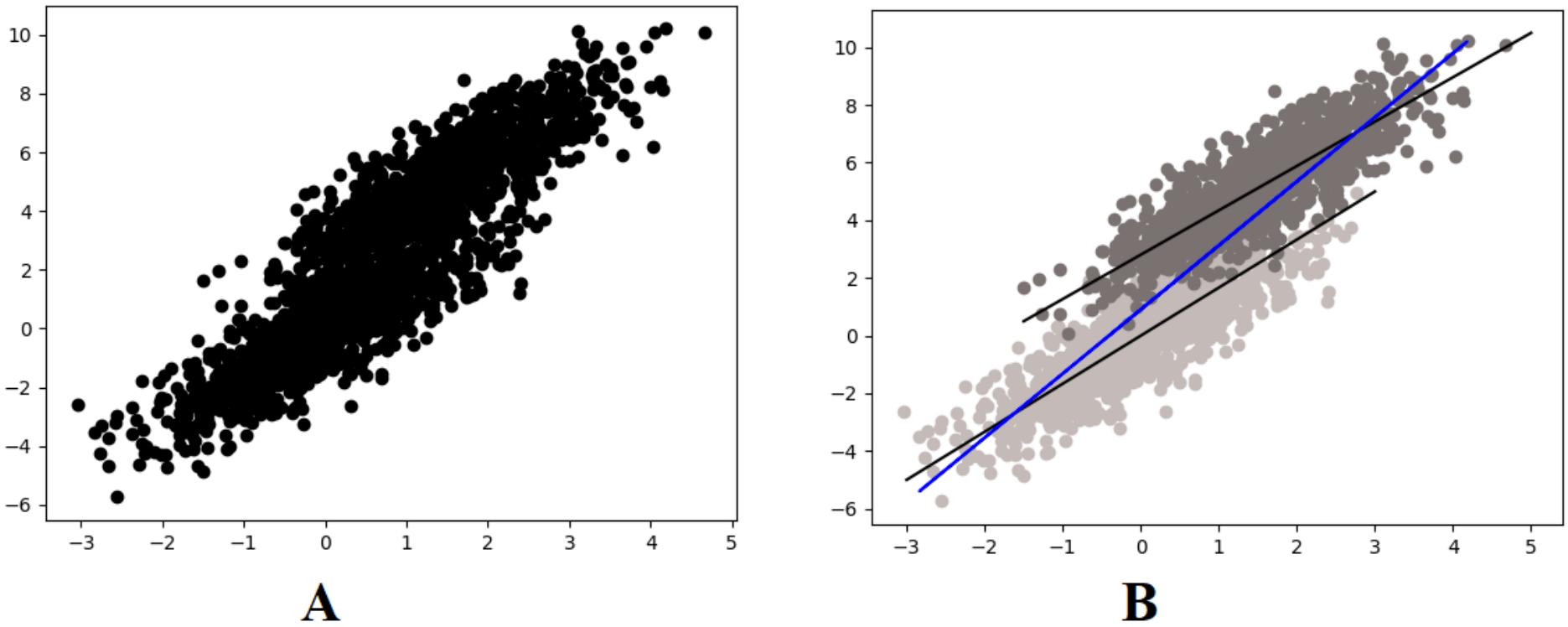



Рис. 1. Деление моделированных данных на группы

# [2] Выводы из нормального уравнения

Из нормального уравнения: $(\mathbf{X}^T\mathbf{X})\hat{\beta} = \mathbf{X}^T y,$

$$(\mathbf{X}^T\mathbf{X})\hat{\beta} = \mathbf{X}^T(\mathbf{X}\hat{\beta} + e),$$

$$\mathbf{X}^T e = 0.$$

Гиперплоскость регрессии проходит через средние наблюдения $\bar{X}, \bar{y}$:

$$\bar{e} = \bar{y} - \bar{X}\hat{\beta} = 0,$$

$$\bar{y} = \bar{X}\hat{\beta};$$

Свойства SSCP матрицы: Элементы вне главное диагонали зависят от ковариации независимых переменных.

$$\hat{\beta} = \underbrace{(\mathbf{X}^T\mathbf{X})}_{SSCP}{}^{-1}\mathbf{X}^T y.$$

# [2] Результаты, полученные при разделении пациентов на группы по типу травмы

Таблица 2. Коэффициенты детерминации общей модели $f_\Theta$ и моделей для групп пациентов.

| $R^2(\hat{f}_\Theta, Y)$ | $R^2(\hat{f}_{\Theta 1}, Y)$ | $R^2(\hat{f}_{\Theta 2}, Y)$ |
|---|---|---|
| 0.76 | 0.8242 | 0.8794 |

Таблица 3. Коэффициенты регрессии моделей

| | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ | $\hat{\beta}_5$ | $\hat{\beta}_6$ | $\hat{\beta}_7$ | $\hat{\beta}_8$ | $\hat{\beta}_9$ | $\hat{\beta}_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $f_\Theta$ | 0.05 | 0.03 | 0.0014 | 0.31* | 0.14 | -0.05 | 0.11 | 0.37 | -0.14 | 0.24 |
| $f_{\Theta 1}$ | -0.04 | -0.05 | 0.039 | 0.59* | 0.01 | 0.03 | -0.44* | 0.02 | 0.14* | 0.20* |
| $f_{\Theta 2}$ | -0.11 | 0.02 | -0.16 | 0.60* | -0.03 | -0.54* | 0.41* | 0.74* | -0.05 | 0.25 |

# [2] Отбор частных предсказаний по $R^2$

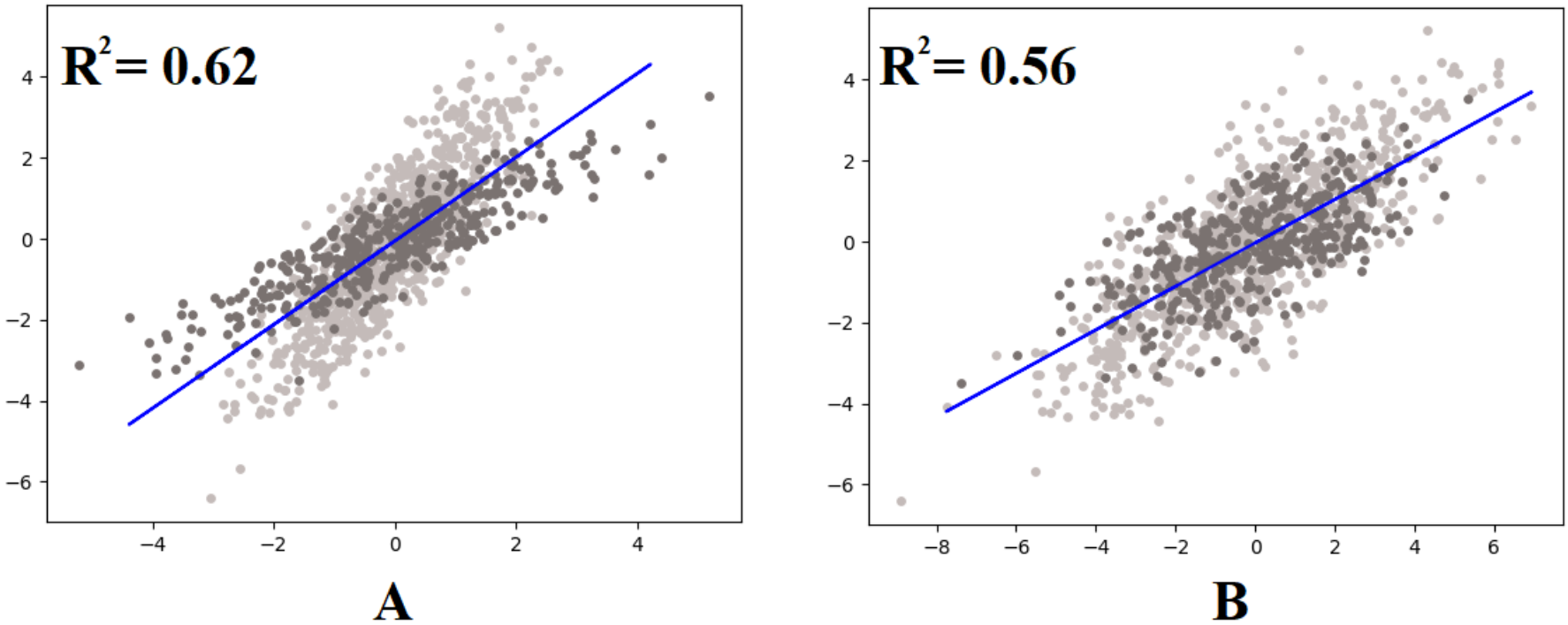



Рис. 2. Линии регрессии в данных с кластерами

# [2] Моделирование данных; расположение групп

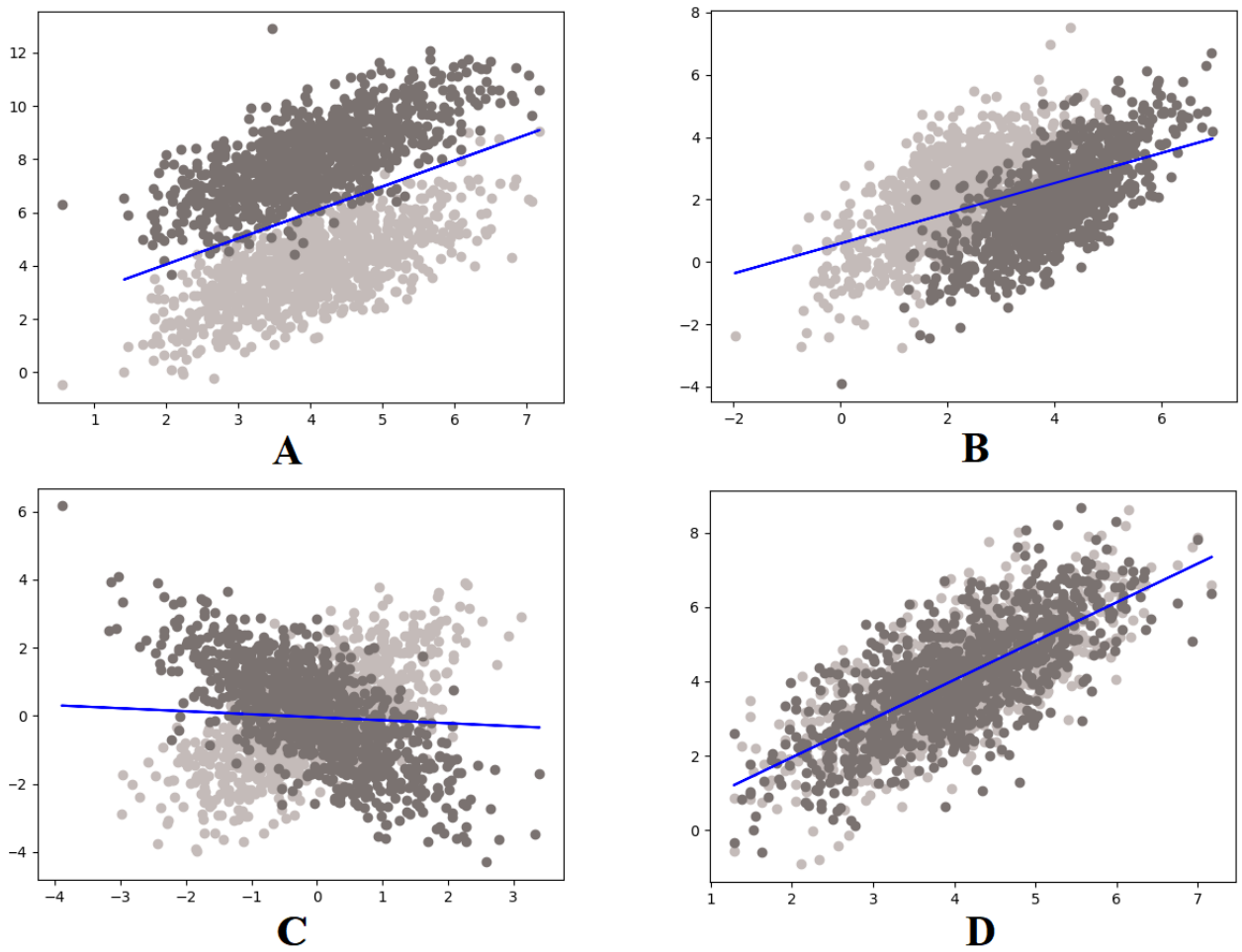



Рис. 3. (**A**): $\Sigma_1 = \Sigma_2$, $\mu_1 \neq \mu_2$; (**B**): $\Sigma_1 \neq \Sigma_2$, $\mu_1 \neq \mu_2$; (**C**): $\Sigma_1 \neq \Sigma_2$, $\mu_1 = \mu_2$; (**D**): $\Sigma_1 = \Sigma_2$, $\mu_1 = \mu_2$

# [2] Критерий Хотеллинга ($\alpha = 0.05$)

Случайные вектора $\xi^{(1)}, \xi^{(2)} \in \mathbb{R}^{p+1}$; проверяем гипотезу

$$H_0 : \mathbb{E}\xi^{(1)} = \mathbb{E}\xi^{(2)}$$
$$(H_0 : \mu^{(1)} = \mu^{(2)}).$$

Статистика:

$$T^2 = (\bar{x}^{(1)} - \bar{x}^{(2)})^T \left( \frac{\hat{\mathbf{S}}_1}{n_1} + \frac{\hat{\mathbf{S}}_2}{n_2} \right)^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}) \sim T^2(p, n_1 + n_2 - 2), \qquad (6)$$

Через распределение Фишера:

$$F = \frac{k - p + 1}{p} T^2 \sim F(p, k - p + 1). \qquad (7)$$

# [2] M-тест Бокса ($\alpha = 0.1$)

$$H_0 : \Sigma_1 = \ldots = \Sigma_g.$$

Статистика:

$$Box's = -2(1 - C)ln(\mathbf{M}) \sim \chi^2_{df}, df = (g - 1)p(p + 1)/2 \tag{8}$$

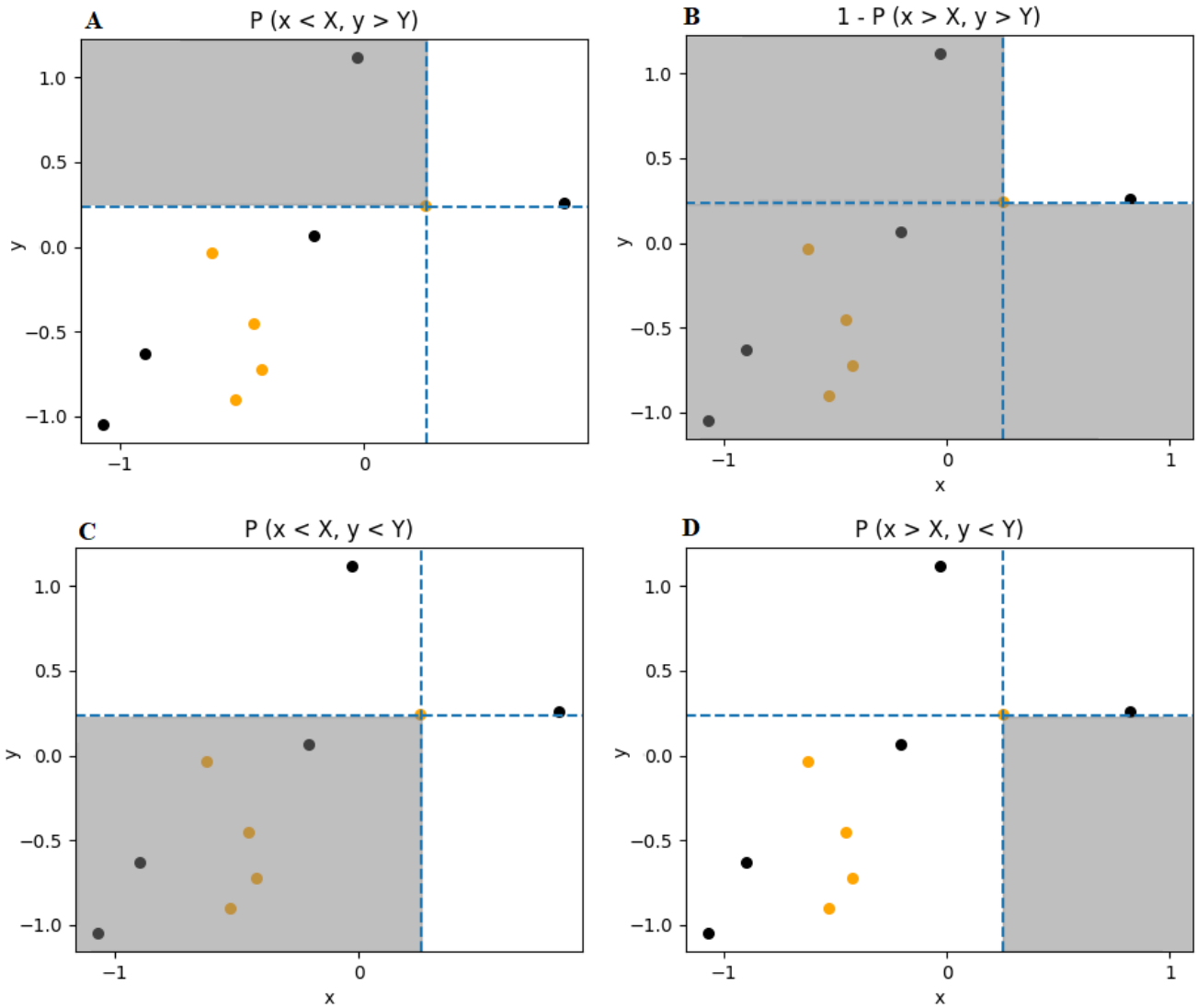
$$C = \left( \sum_{i=1}^{g} \frac{1}{n_i - 1} - \frac{1}{(\sum_{i=1}^{g} n_i) - g} \right) \frac{2p^2 + 3p - 1}{6(p + 1)(g - 1)}. \tag{9}$$

$$\mathbf{M} = \left( (\sum_{i=1}^{g} n_i) - g \right) ln|\mathbf{S}_p| - \sum_{i=1}^{g} (n_i - 1) ln|\hat{\mathbf{S}}_i|, \tag{10}$$

где

$$\mathbf{S}_p = \left( (\sum_{i=1}^{g} n_i) - g \right)^{-1} \sum_{i=1}^{g} (n_i - 1)\hat{\mathbf{S}}_i. \tag{11}$$

# [2] Альтернативный критерий отбора частных предсказаний



# [2] Альтернативный критерий отбора частных предсказаний

Критерий Фазано-Франческини ($\alpha = 0.05$):

$$H_0 : F_1 = F_2;$$

Статистика:

$$D = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left( \frac{D_{FF1} + D_{FF2}}{2} \right), \tag{12}$$

Тогда p-value находится как доля статистик $D_i \geq D$,

$$\hat{p} = \frac{1 + \sum_{i=1}^{M} \mathbb{1}(D_i \geq D)}{1 + M}, \tag{13}$$

где

$$\mathbb{1}(x \geq y) = \begin{cases} 1, x \geq y \\ 0, x < y. \end{cases} \tag{14}$$

# [3] Сравнение моделей

Таблица 4. Сравнение коэффициентов регрессии при различном отборе частных предсказаний

| | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ | $\hat{\beta}_5$ | $\hat{\beta}_6$ | $\hat{\beta}_7$ | $\hat{\beta}_8$ | $\hat{\beta}_9$ | $\hat{\beta}_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $f_{\Theta R^2}$ | 0.052 | 0.026 | 0.0014 | 0.31* | 0.14 | -0.045 | 0.11 | 0.37 | -0.14 | 0.24 |
| $f_{\Theta Hotelling/Box}$ | 0.063 | 0.15 | — | 0.44* | 0.26* | -0.029 | -0.0003 | — | — | 0.13 |
| $f_{\Theta F-F}$ | 0.011 | 0.097 | 0.13 | 0.53* | 0.16 | -0.15 | — | 0.12 | -0.0024 | 0.12 |

Таблица 5. Коэффициент детерминации моделей при различном отборе частных предсказаний

| $R^2(\hat{f}_{\Theta R^2}, Y)$ | $R^2(\hat{f}_{\Theta Hotelling/Box}, Y)$ | $R^2(\hat{f}_{\Theta F-F}, Y)$ |
|---|---|---|
| 0.76 | 0.74 | 0.76 |

# [3] Результаты

- Применен итерационно-частичный метод множественной регрессии для неполных данных;
- Рассмотрено влияние кластеров в обсуждаемых данных на решение задачи регрессии;
- Предложены и проверены процедуры отбора частных предсказаний;
- Используя новый подход, мы способны учитывать расслоение в данных;
- Полученные результаты дают возможность предлагать разнообразные интерпретации модели исследователям без ухудшения точности предсказаний.