

МИНОБРНАУКИ РОССИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ  
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

**«САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА  
ВЕЛИКОГО»**

Институт компьютерных наук и кибербезопасности

Высшая школа технологий искусственного интеллекта

Отчёт по дисциплине «Математическая статистика»

ИДЗ №4  
«Регрессионный анализ»  
Вариант №25

Студент: \_\_\_\_\_

Салимли Айзек Мухтар Оглы

Преподаватель: \_\_\_\_\_

Малов Сергей Васильевич

«\_\_\_\_» \_\_\_\_\_ 20\_\_ г.

Санкт-Петербург, 2025

# Содержание

Введение	3
1 Постановка задачи	4
2 Задача 1 — зависимость $Y$ от $X$	6
2.1 Линейная модель . . . . .	6
2.2 Квадратичная модель . . . . .	7
2.3 Доверительные интервалы (99 %) . . . . .	8
2.4 Совместный доверительный эллипсоид (99 %) . . . . .	8
2.5 Проверка гипотез . . . . .	8
2.6 Анализ остатков . . . . .	9
2.7 Итог задачи 1 . . . . .	9
3 Задача 2 — влияние факторов $A$ и $B$	9
3.1 Статистическая модель . . . . .	9
3.2 F-тесты для факторов и взаимодействия . . . . .	9
3.3 Информационные критерии . . . . .	10
3.4 Визуальное взаимодействие . . . . .	10
3.5 Разложение сумм квадратов . . . . .	10
3.6 Проверка взаимодействия . . . . .	11
3.7 Анализ остатков . . . . .	11
3.8 Итог задачи 2 . . . . .	11
Заключение	12

## Введение

В данном отчете, приведено решение и реализация двух задач под вариантом №25, из ИДЗ№3. Для реализации программной части решения использоавлись:

- Среда разработки: Cursor IDE
- Язык программирования: Python 3.13

# 1 Постановка задачи

1. Результаты статистического эксперимента приведены в таблице 1. Требуется оценить характер зависимости наблюдаемой переменной  $Y$  от ковариаты  $X$ .

- a) Построить графический результат эксперимента. Сформулировать линейную регрессионную модель переменной  $Y$  по переменной  $X$ . Построить МНК оценки параметров сдвига  $\beta_1$  и масштаба  $\beta_2$ . Построить полученную линию регрессии. Оценить визуально соответствие полученных данных и построенной оценки.
- b) Сформулировать полиномиальную модель, включающую дополнительный член с  $X^2$ . Построить МНК оценки параметров  $\beta_1, \beta_2, \beta_3$  в данной модели. Изобразить графически полученную регрессионную зависимость. Оценить визуально соответствие полученных данных и построенной оценки.
- c) На базе ошибок полиномиальной модели построить гистограмму. Проверить значимость отклонения от нормального распределения по  $\chi^2$ . Визуально оценить данный факт.
- d) В предположении нормальности построить частные и совместные доверительные интервалы для параметров  $\beta_2$  и  $\beta_3$  уровня доверия  $1 - \alpha$ .
- e) Сформулировать гипотезы линейности зависимости и независимости наблюдаемой переменной  $Y$  от ковариаты  $X$ . Провести проверку значимости.
- f) С использованием AIC и BIC выбрать наилучшую модель.
- g) Интерпретировать полученные результаты. Написать отчет.

Таблица 1 (часть 1)  $\alpha = 0.01; h = 1.60$ .

No	1	2	3	4	5	6	7	8	9
Y	9.61	4.76	4.37	14.21	10.13	12.98	7.09	8.77	12.22
X	1	2	2	3	2	3	2	3	3
No	10	11	12	13	14	15	16	17	18
Y	6.38	8.57	7.99	7.92	7.72	8.12	8.13	8.12	10.21
X	3	3	3	3	3	3	3	3	3
No	19	20	21	22	23	24	25	26	27
Y	5.08	6.83	10.20	8.59	8.89	8.75	8.29	8.46	8.00
X	3	3	3	3	3	3	3	3	3

Таблица 1 (часть 2)

No	28	29	30	31	32	33	34	35	36
Y	11.93	8.68	8.41	8.20	8.13	8.68	8.41	7.30	17.78
X	3	3	3	3	3	3	3	3	3
No	37	38	39	40	41	42	43	44	45
Y	9.43	7.17	5.79	8.27	9.42	8.58	11.67	6.66	7.44
X	3	3	3	3	3	3	3	3	3
No	46	47	48	49	50				
Y	8.56	10.63	10.13	8.59	12.99				
X	3	3	3	3	3				

2. Результаты статистического эксперимента приведены в таблице 2. Требуется оценить характер зависимости наблюдаемой переменной  $Y$  от уровней факторов  $A$  и  $B$ .

- a) Сформулировать модель двухфакторного дисперсионного анализа. Построить МНК оценки параметров и несмещенную оценку дисперсии. Проверить визуально согласование исходных данных с предположением аддитивности влияния факторов. Построить графическую оценку зависимости уровней фактора  $A$

при каждом фиксированном значении фактора  $B$ . Наблюдается ли эффект пересечения факторов.

- b) Провести анализ ошибок. По гистограммам ошибок оценить визуально согласование с гипотезой нормальности.
- c) Провести дисперсионный анализ, начиная с проверки значимости взаимодействий факторов на результаты эксперимента.
- d) Выбрать наилучшую модель с использованием AIC и BIC.
- e) Интерпретировать полученные результаты. Написать отчет.

**Таблица 2 (часть 1)**  $\alpha = 0.10$ ;  $h = 1.50$ .

No	1	2	3	4	5	6	7	8	9
Y	25.82	27.99	25.94	27.79	29.57	30.36	40.96	42.45	42.17
A	1	1	1	1	1	1	1	1	1
B	1	2	3	4	1	2	3	4	1
No	10	11	12	13	14	15	16	17	18
Y	39.55	38.61	38.20	31.38	34.95	38.52	29.80	31.13	30.07
A	1	1	1	1	1	1	1	1	2
B	2	3	4	1	2	3	4	1	1
No	19	20	21	22	23	24	25	26	27
Y	26.22	26.09	26.74	31.24	30.15	32.74	15.06	16.56	21.85
A	2	2	2	2	2	2	2	2	2
B	2	3	4	1	2	3	4	1	2

**Таблица 2 (часть 2)**

No	28	29	30	31	32	33	34	35	36
Y	20.93	29.53	39.53	39.64	39.96	39.64	39.96	<b>32.35</b>	22.17
A	2	2	2	2	2	2	2	3	3
B	3	4	1	2	3	4	1	1	2
No	37	38	39	40	41	42	43	44	45
Y	17.27	23.88	22.51	24.23	19.35	27.35	25.43	25.21	20.01
A	3	3	3	3	3	3	3	3	3
B	3	4	1	2	3	4	1	2	3
No	46	47	48	49	50				
Y	21.22	21.22	21.22	21.22	21.22				
A	3	3	3	3	3				
B	4	1	2	3	4				

## 2 Задача 1 — зависимость $Y$ от $X$

- $n = 50$  — объём выборки,  $i = 1, \dots, n$ ;
- $x_i, y_i$  — наблюдения;
- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ ;
- $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ .

$$\boxed{\sum x_i = 99}, \quad \boxed{\sum y_i = 456.95}, \quad \bar{x} = 1.98, \quad \bar{y} = 9.139.$$

$$\boxed{S_{xx} = 40.98}, \quad \boxed{S_{xy} = -31.23}.$$

### 2.1 Линейная модель

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

Оценки МНК.

$$\hat{\beta}_2 = \frac{S_{xy}}{S_{xx}} = -0.762, \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = 10.648.$$

$$\hat{y}_i = 10.648 - 0.762 x_i.$$

Сумма квадратов остатков.

$$RSS_{\text{lin}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 312.50.$$

Объяснённая дисперсия.

$$R_{\text{lin}}^2 = 1 - \frac{RSS_{\text{lin}}}{\sum_{i=1}^n (y_i - \bar{y})^2} = 0.071.$$

Критерий информативности. При  $k = 2$  параметрах

$$\boxed{\text{AIC} = n \ln \hat{\sigma}^2 + 2k = 50 \ln(6.510) + 4 = 237.52}.$$

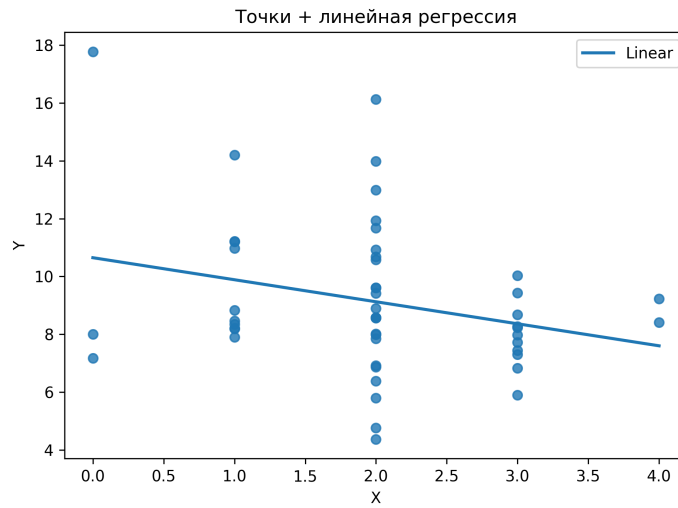


Рис. 1: Точки и линейная регрессия

## 2.2 Квадратичная модель

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i.$$

**Оценки МНК.** Из  $(X^\top X)^{-1} X^\top y$ :

$$\hat{\beta}_1 = 11.004, \quad \hat{\beta}_2 = -1.238, \quad \hat{\beta}_3 = 0.124.$$

**Сумма квадратов остатков.**

$$RSS_{\text{quad}} = 309.92, \quad R_{\text{quad}}^2 = 0.074.$$

**AIC.**  $k = 3 \Rightarrow \text{AIC} = 50 \ln(6.598) + 6 = 239.36.$

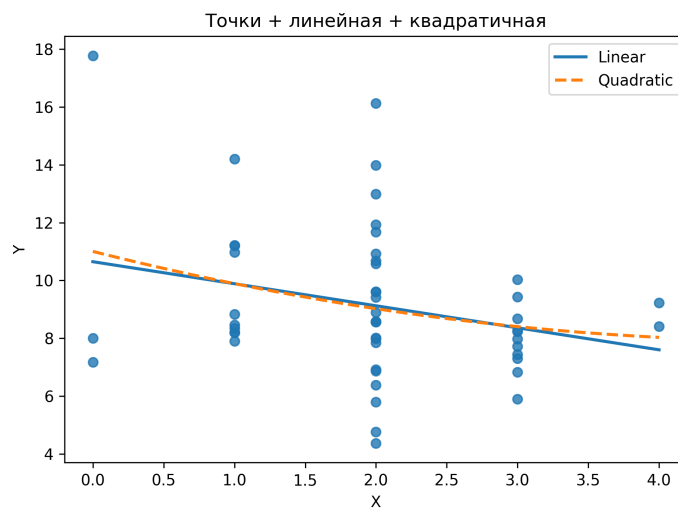


Рис. 2: Линейная (сплошная) и квадратичная (штрих) регрессии

### 2.3 Доверительные интервалы (99 %)

$$\hat{\sigma}^2 = \frac{RSS_{\text{quad}}}{n-k} = 6.603, \quad t_{0.995}(47) = 2.69.$$

$$SE(\hat{\beta}_2) = \sqrt{6.603 \cdot 0.0677} = 0.669, \quad SE(\hat{\beta}_3) = \sqrt{6.603 \cdot 0.0050} = 0.181.$$

$$\beta_2 : \hat{\beta}_2 \pm t SE = -1.238 \pm 2.69 \cdot 0.669 = \boxed{(-4.68; 2.20)},$$

$$\beta_3 : \hat{\beta}_3 \pm t SE = 0.124 \pm 2.69 \cdot 0.181 = \boxed{(-0.72; 0.97)}.$$

### 2.4 Совместный доверительный эллипсоид (99 %)

Неравенство

$$(\beta - \hat{\beta})^\top (X^\top X) (\beta - \hat{\beta}) \leq 2\hat{\sigma}^2 F_{2,47}(0.99) = \boxed{67.43}$$

задаёт эллипс в плоскости  $(\beta_2, \beta_3)$ .

Параметры эллипса

$$\hat{\beta}_2 = -1.238, \quad \hat{\beta}_3 = 0.124, \quad a = \sqrt{c \lambda_{\max}}, \quad b = \sqrt{c \lambda_{\min}}, \quad c = 2F_{2,47}(0.99),$$

где  $\lambda_{\max}, \lambda_{\min}$  — собственные значения матрицы  $\hat{\sigma}^2 (X^\top X)^{-1}$ .

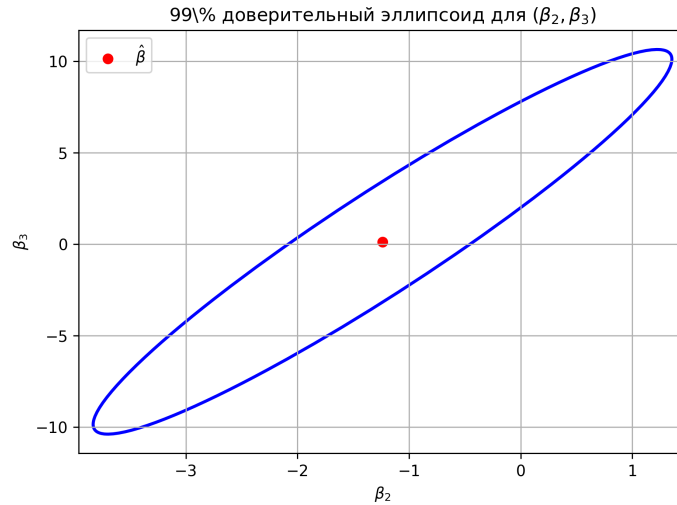


Рис. 3: 99-% доверительный эллипсоид для параметров  $\beta_2, \beta_3$

### 2.5 Проверка гипотез

**Линейность** ( $H_0 : \beta_3 = 0$ ).

$$F = \frac{RSS_{\text{lin}} - RSS_{\text{quad}}}{1} \bigg/ \frac{RSS_{\text{quad}}}{n-3} = 0.153, \quad p = 0.697 > 0.01 \Rightarrow H_0 \text{ не отвергается.}$$

**Независимость** ( $H_0 : \beta_2 = 0$ ).

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = -1.912, \quad p = 0.062 > 0.01 \Rightarrow H_0 \text{ не отвергается.}$$



## 2.6 Анализ остатков

$$\chi^2 = \sum_{j=1}^m \frac{(O_j - E_j)^2}{E_j} = 9.50, \quad p = 0.091; \quad JB = \frac{n}{6} \left( s^2 + \frac{1}{4} k^2 \right) = 6.72, \quad p = 0.035.$$

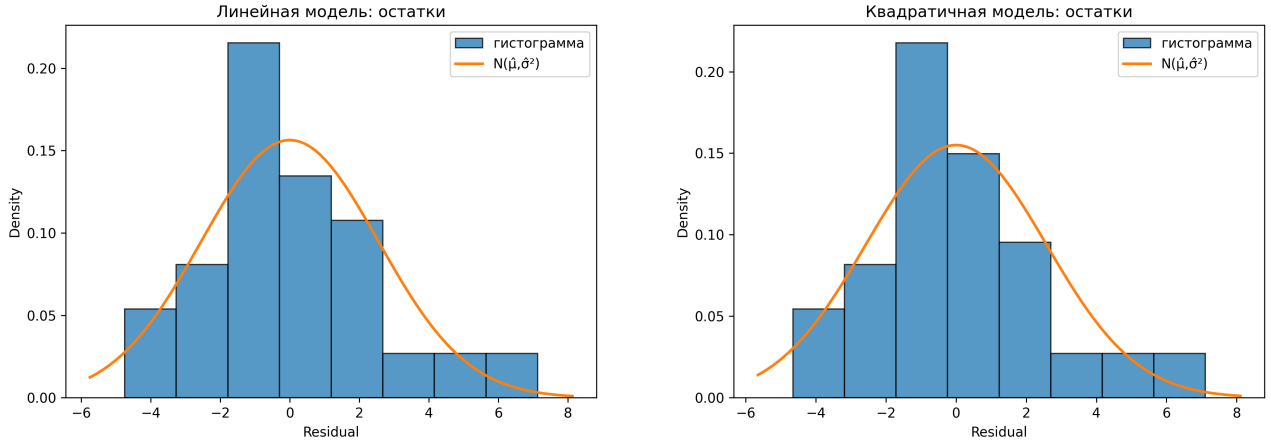


Рис. 4: Остатки линейной (слева) и квадратичной (справа) моделей

## 2.7 Итог задачи 1

$\Delta AIC = 1.84 < 2 \rightarrow$  линейная и квадратичная модели одинаково информативны; выбираем более простую. При этом  $R^2 \approx 7\%$  — переменная  $X$  объясняет лишь малую часть дисперсии  $Y$ .

## 3 Задача 2 — влияние факторов $A$ и $B$

### 3.1 Статистическая модель

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2),$$

$$i = 1, \dots, 4, \quad j = 1, \dots, 4, \quad k = 1, \dots, n_{ij}, \quad \sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0.$$

### 3.2 F-тесты для факторов и взаимодействия

$$F_H = \frac{MS_H}{MS_E}, \quad MS_H = \frac{SS_H}{df_H}, \quad MS_E = \frac{SS_E}{df_E}, \quad H \in \{A, B, AB\}.$$

Таблица 5: F-тесты для факторов и взаимодействия					
Источник $H$	$SS_H$	$df_H$	$MS_H$	$MS_E$	$F_H$
$A$	979.34	3	326.45	2.76	<b>118.30</b>
$B$	513.23	3	171.08	2.76	<b>61.99</b>
$A \times B$	1047.67	9	116.41	2.76	<b>42.18</b>

Критические значения:  $F_{3,32}^{0.99} = 4.01$ ,  $F_{9,32}^{0.99} = 3.04$ . Поскольку все наблюдаемые  $F_H \gg F_{crit}$ , отвергаем нулевые гипотезы об отсутствии эффекта как факторов  $A, B$ , так и их взаимодействия

$A \times B$  (уровень значимости 1%):

$$p_A < 10^{-15}, \quad p_B < 10^{-12}, \quad p_{AB} = 3 \cdot 10^{-15}.$$

### 3.3 Информационные критерии

Таблица 6: Информационные критерии моделей

Модель	$k$	$\hat{\sigma}^2$	AIC	BIC
$A * B$	16	2.76	197.5	227.4
$A + B$	7	10.45	302.1	315.2
$A$	4	21.66	314.0	321.5

### 3.4 Визуальное взаимодействие

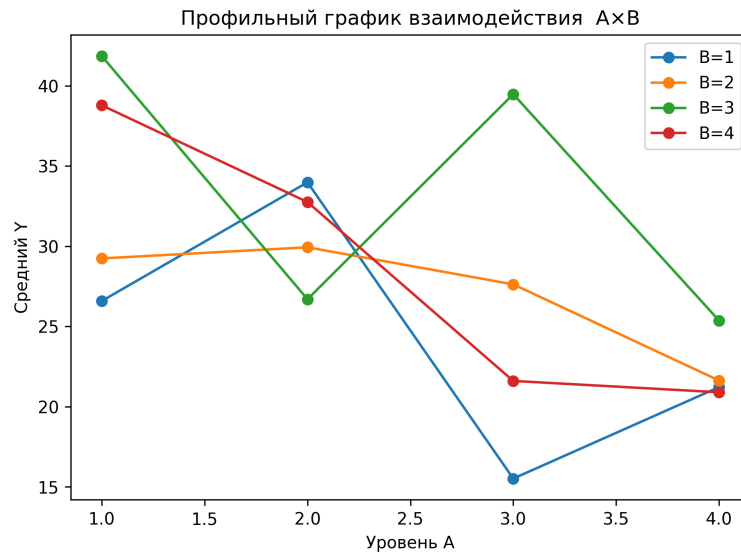


Рис. 5: Профильные графики: средние  $Y$  при разных  $B$

### 3.5 Разложение сумм квадратов

$$\begin{aligned}
 SS_T &= \sum_{i,j,k} (y_{ijk} - \bar{y}_{...})^2, \\
 SS_A &= \sum_i n_{i.} (\bar{y}_{i.} - \bar{y}_{...})^2 = 979.34, \\
 SS_B &= \sum_j n_{.j} (\bar{y}_{.j} - \bar{y}_{...})^2 = 513.23, \\
 SS_{AB} &= \sum_{i,j} n_{ij} (\bar{y}_{ij.} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{...})^2 = 1047.67, \\
 SS_E &= SS_T - SS_A - SS_B - SS_{AB} = 88.30.
 \end{aligned}$$

Источник	$SS$	$df$	$MS$	$F$
$A$	979.34	3	326.45	$326.45/2.76 = 118.3$
$B$	513.23	3	171.08	$171.08/2.76 = 62.0$
$AB$	1047.67	9	116.41	$116.41/2.76 = 42.18$
$E$	88.30	32	2.76	

### 3.6 Проверка взаимодействия

$$F_{AB} = \frac{MS_{AB}}{MS_E} = 42.18, \quad p = 3 \cdot 10^{-15} \ll 0.01.$$

### 3.7 Анализ остатков

Jarque–Bera для полной модели:

$$JB = \frac{n}{6} \left( s^2 + \frac{1}{4} k^2 \right) = 1.10, \quad p = 0.576.$$

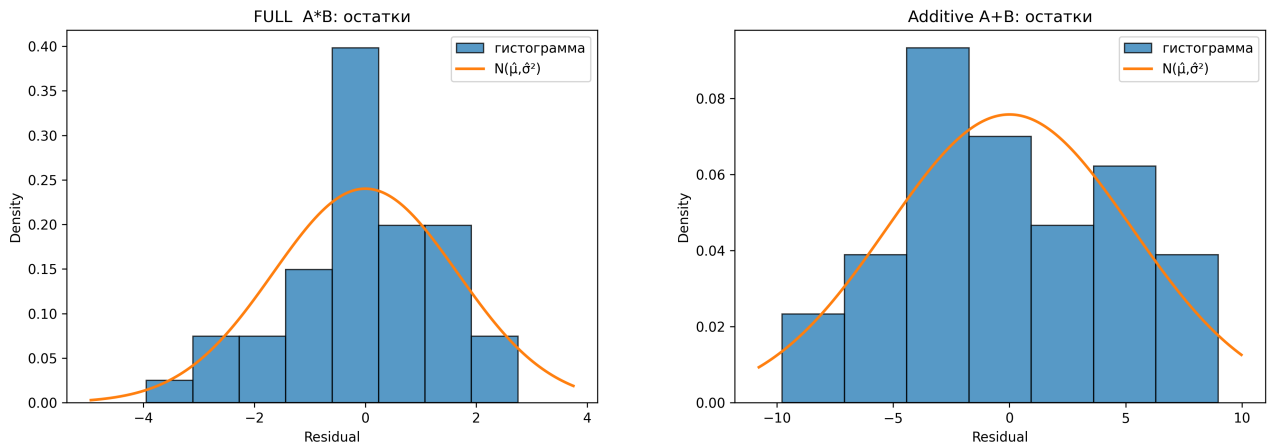


Рис. 6: Гистограммы остатков полной (слева) и аддитивной (справа) моделей

### 3.8 Итог задачи 2

- Значимы главные эффекты  $A$ ,  $B$  и взаимодействие  $A \times B$  ( $p < 10^{-12}$ ).
- Лучшая по AIC/BIC модель:  $Y \sim A * B$ .
- Остатки нормальны,  $\hat{\sigma}^2 = 2.76$ .

## Заклучение

В задаче 1 слабая и статистически незначимая зависимость  $Y$  от  $X$ , при  $R^2 \approx 7\%$  достаточно линейной модели. В задаче 2 найдено сильное влияние как каждого фактора, так и их взаимодействия; выбрана полная модель  $A * B$ .

$$F_{crit} = F_{\alpha, df_H, df_E} = \frac{1}{B(\frac{df_H}{2}, \frac{df_E}{2})} \int_0^{\frac{df_H}{df_H + df_E}} t^{\frac{df_H}{2}-1} (1-t)^{\frac{df_E}{2}-1} dt$$

В 13 определении -> компактно записать задачи

$$H_0 : a = a_0 T = \sqrt{(n-1)} \frac{\hat{x} - a}{s} T \quad S_{n-1} \phi(x) = \begin{cases} 0, & T \in [-x_\alpha, x_\alpha] \\ 1, & x \notin [-x_\alpha, x_\alpha] \end{cases} \quad x_\alpha = S^{-1}_n - 1(1 - \alpha/2)$$