

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение
высшего образования «Санкт-Петербургский политехнический
университет Петра Великого»

Институт компьютерных наук и кибербезопасности

Высшая школа технологий искусственного интеллекта

Направление: 02.03.01 «Математика и компьютерные науки»

Индивидуальное домашнее задание №4
по дисциплине
«Математическая статистика»
Вариант 27

Студент,

группы 5130201/20102

_____ Тищенко А. А.

Преподаватель

_____ Малов С. В.

« _____ » _____ 2025г.

Санкт-Петербург, 2025

1 Задание №1

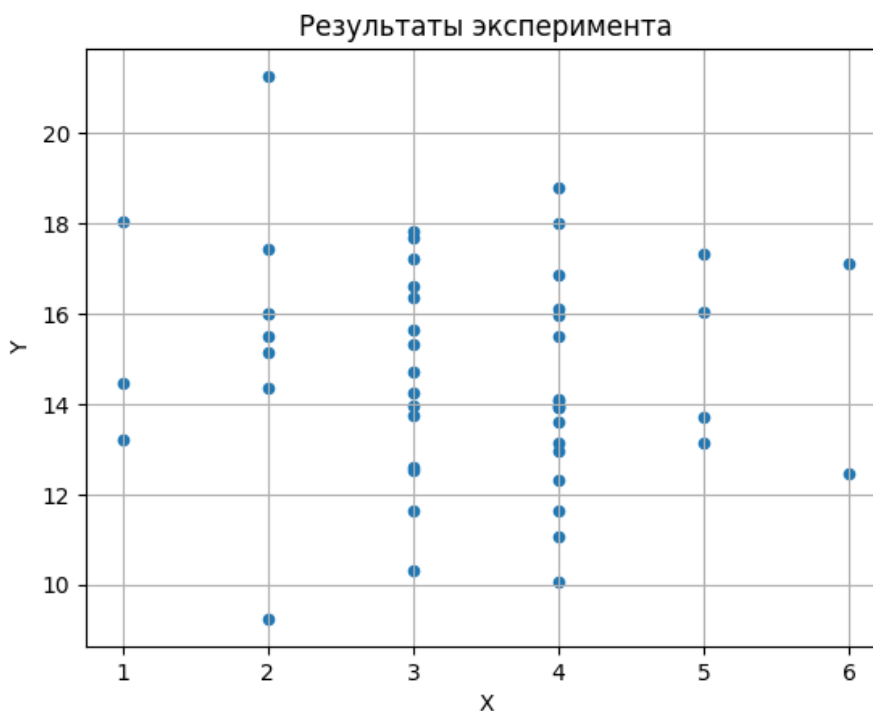
Вар. 27 (513020125)

1. Результаты статистического эксперимента приведены в таблице 1. Требуется оценить характер зависимости наблюдаемой переменной Y от ковариаты X .
 - a) Построить графически результаты эксперимента. Сформулировать линейную регрессионную модель переменной Y по переменной X . Построить МНК оценки параметров сдвига β_1 и масштаба β_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 χ^2 . Визуально оценить данный факт.
 - d) В предположении нормальности построить частные и совместные доверительные интервалы для параметров β_2 и β_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 $\alpha = 0.02; h = 1.40$.

No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Y	12.33	16.61	12.47	14.36	13.21	13.76	13.93	13.96	15.96	15.99	17.32	14.10	12.97	13.60	16.37	16.11	9.24
X	4	3	6	2	1	3	4	3	4	2	5	4	4	4	3	4	2
No	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Y	15.51	14.24	17.23	15.14	14.73	15.52	10.07	21.27	16.86	13.98	11.07	13.70	13.91	17.70	14.08	15.65	13.14
X	2	3	3	2	3	4	4	2	4	4	4	5	4	3	4	3	4
No	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	
Y	17.43	18.79	12.59	15.99	12.53	16.03	11.63	18.01	15.33	11.65	10.32	18.06	17.83	14.46	13.13	17.11	
X	2	4	3	2	3	5	3	4	3	4	3	1	3	1	5	6	

1.1 Пункт а



Формулировка линейной регрессионной модели Линейная регрессионная модель зависимости Y от X имеет вид:

$$Y = \beta_1 + \beta_2 X + \epsilon,$$

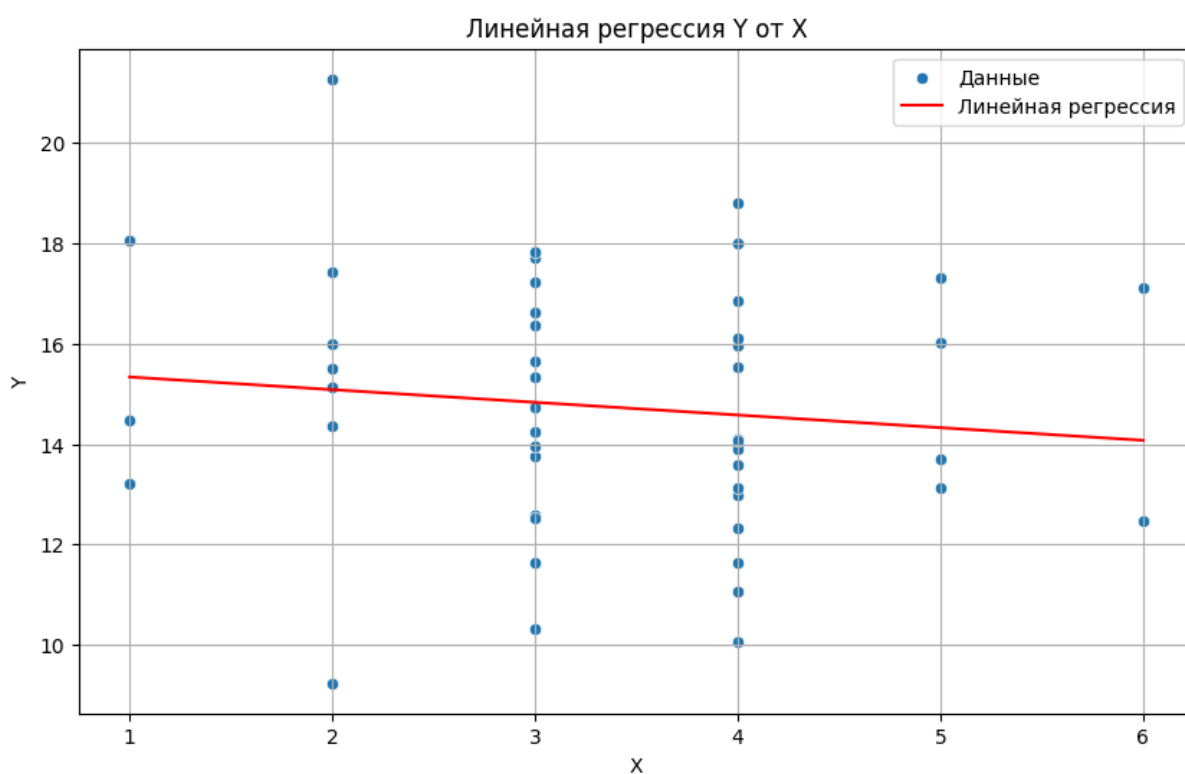
где: - β_1 — параметр сдвига, - β_2 — параметр масштаба, - ϵ — случайная ошибка.

Построение МНК-оценок параметров Метод наименьших квадратов (МНК) используется для нахождения оценок $\hat{\beta}_1$ и $\hat{\beta}_2$, которые минимизируют сумму квадратов остатков.

$$\beta_1 = 15.5869$$

$$\beta_2 = -0.2522$$

R^2 линейной модели: 0.0144



Распределение точек относительно линии Точки разбросаны, линия не отражает тренд, что говорит о плохом соответствии.

Наклон линии: Линия близка к горизонтальной, зависимость слабая.

Таким образом, Между X и Y нет линейной зависимости. Линейная модель не подходит для описания данных.

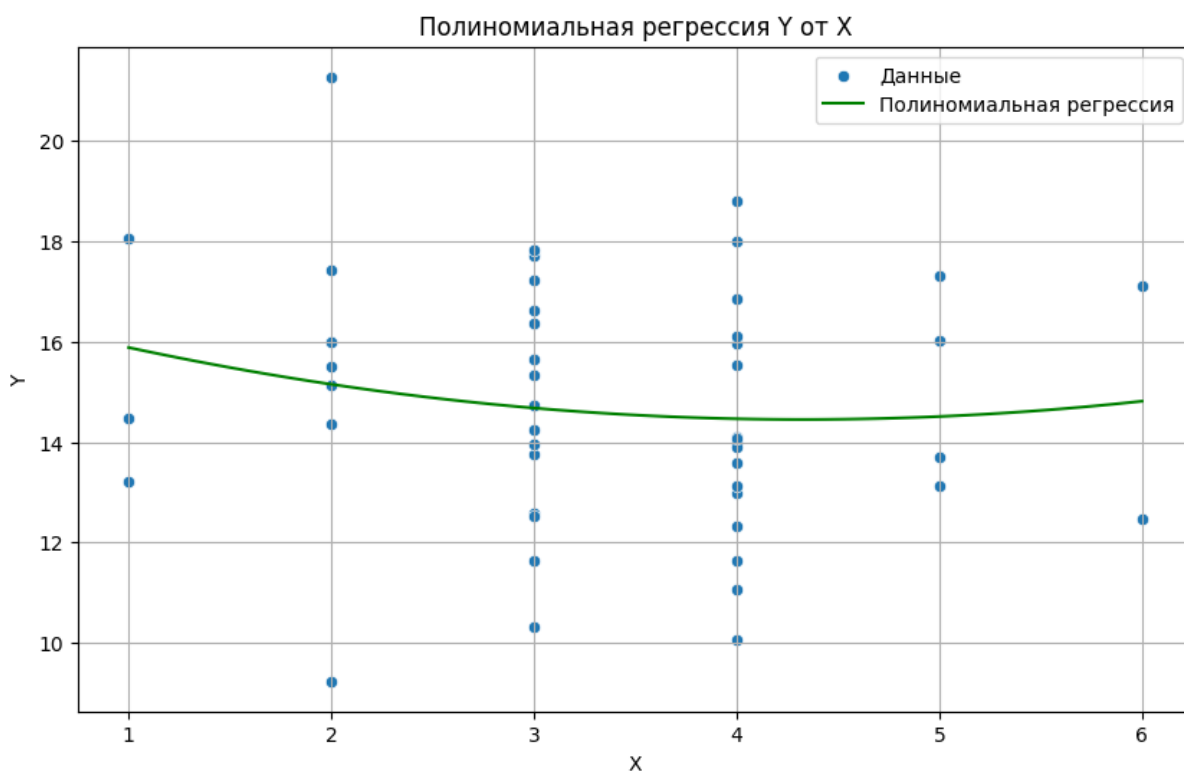
1.2 Пункт в

Формулировка полиномиальной регрессионной модели Полиномиальная регрессионная модель зависимости Y от X имеет вид:

$$Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \epsilon,$$

где:

- β_1 — параметр сдвига,
- β_2 — линейный коэффициент при X ,
- β_3 — квадратичный коэффициент при X^2 ,
- ϵ — случайная ошибка



Полиномиальная модель: $\beta_1 = 16.8727$ $\beta_2 = -1.1208$ $\beta_3 = 0.1296$

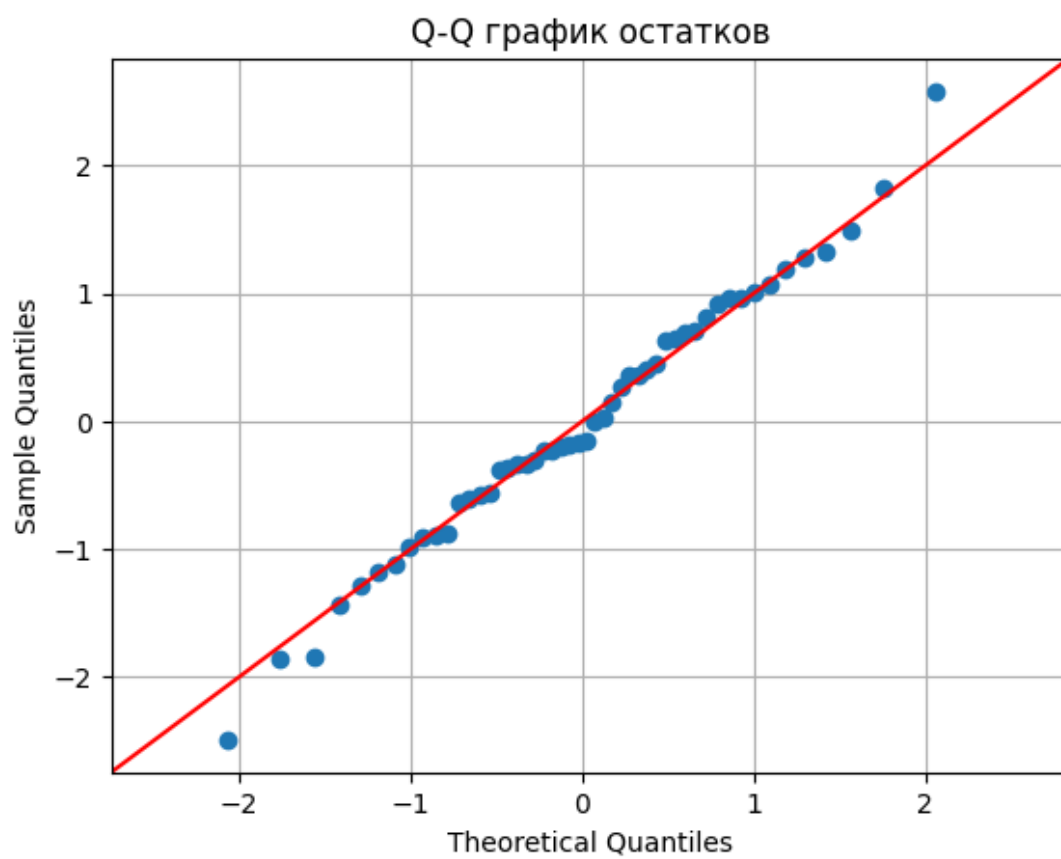
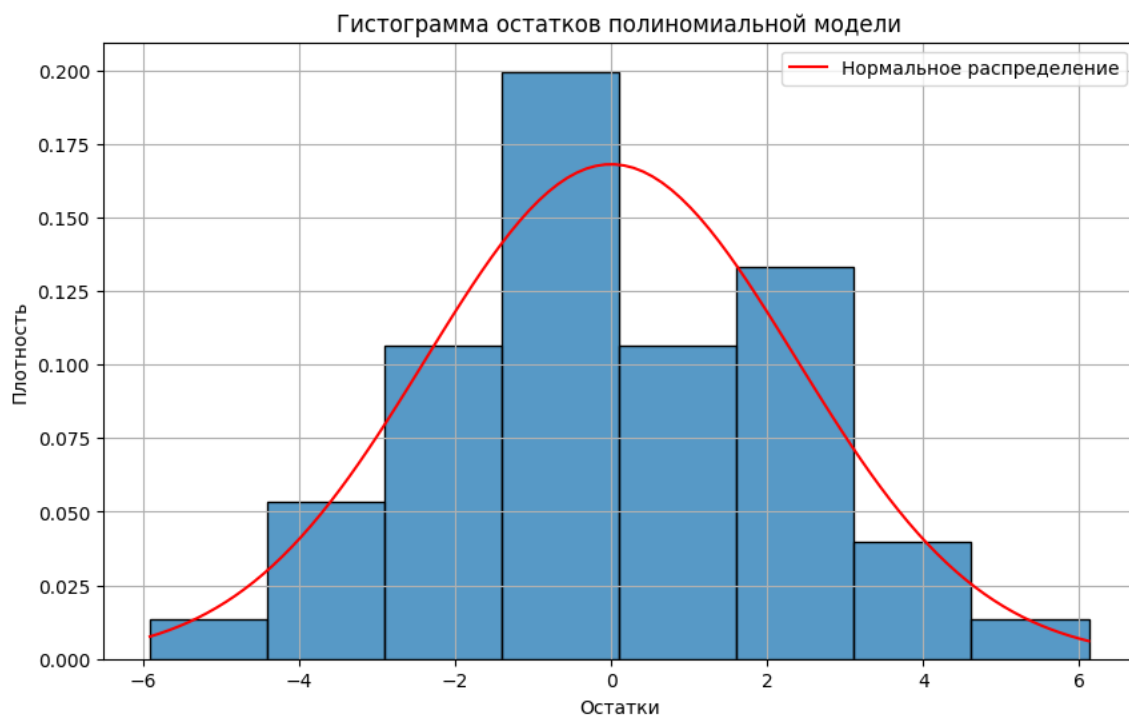
R^2 полиномиальной модели: 0.0240

Распределение точек относительно линии: Точки разбросаны, линия не отражает тренд, что говорит о плохом соответствии.

Низкий R^2 означает, что квадратичная модель плохо описывает связь между X и Y .

Результаты указывают на то, что квадратичная модель не подходит для описания данных.

1.3 Пункт с



Проверка нормальности с помощью критерия χ^2

Этапы:

1. Гипотезы:

- H_0 : Остатки имеют нормальное распределение.
- H_1 : Остатки не имеют нормального распределения.

2. Разделить данные на интервалы (бины): Используем те же интервалы, что и в гистограмме.

3. Рассчитать наблюдаемые (O_i) и ожидаемые (E_i) частоты:

- $E_i = N \cdot P$ (для i -го интервала), где P — вероятность из нормального распределения $N(\mu, \sigma^2)$.

4. Вычислить статистику χ^2 :

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}.$$

5. Сравнить с критическим значением χ^2 : Если $\chi^2 > \chi^2_{\text{крит}}$, отвергаем H_0 .

Хи-квадрат статистика: 2.7737

Критическое значение: 13.3882

p-value: 0.7348

Не отвергаем H_0 : распределение нормальное

Визуально: Остатки близки к нормальному распределению.

Статистически: Критерий χ^2 не выявил значимых отклонений от нормальности на уровне $\alpha = 0.02$.

Предположение о нормальности ошибок выполняется.

1.4 Пункт d

Частные интервалы строятся для каждого параметра отдельно, используя t-распределение.

Формула:

$$\hat{\beta}_j \pm t_{1-\alpha/2, n-p} \cdot SE(\hat{\beta}_j),$$

где:

- $\hat{\beta}_j$ - оценка параметра,
- $SE(\hat{\beta}_j)$ - стандартная ошибка параметра,
- $t_{1-\alpha/2}$ - критическое значение t-распределения,
- n - число наблюдений,

- p - число параметров модели (для квадратичной модели $p = 3$).

Доверительные интервалы (уровень 0.98):

- Доверительный интервал для β_2 (98.0%): [-4.2930, 2.0514]
- Доверительный интервал для β_3 (98.0%): [-0.3310, 0.5902]

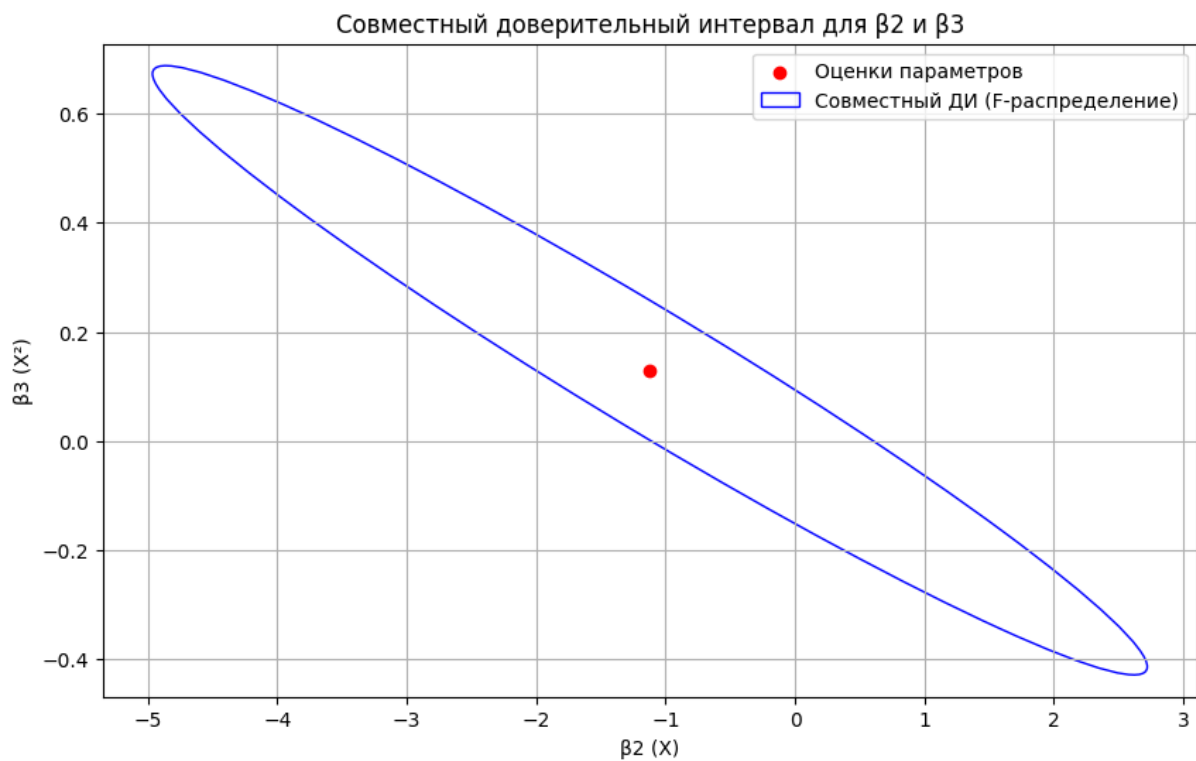
Совместные доверительные интервалы Совместные интервалы учитывают корреляцию между оценками параметров. Используем метод Бонферрони или F-распределение.

Метод Бонферрони

Формула:

$$\hat{\beta}_j \pm t_{1-\alpha/(2k), n-p} \cdot SE(\hat{\beta}_j),$$

где $k = 2$ (число параметров β_2 и β_3).



Ковариационная матрица для β_2 и β_3 :

	X	X2
X	1.734960	-0.245172
X2	-0.245172	0.036575

Совместные интервалы (Бонферрони):

- β_2 : [-4.657, 2.415]

- β_3 : [-0.384, 0.643]

Метод F-распределения

Формула:

$$(\hat{\beta} - \beta)^T \cdot Cov(\hat{\beta})^{-1} \cdot (\hat{\beta} - \beta) \leq F_{1-\alpha, 2, n-p},$$

где $F_{1-\alpha, 2, n-p}$ - критическое значение F-распределения.

Полная ковариационная матрица:

	const	X	X2
const	4.7543	-2.7403	0.3629
X	-2.7403	1.7350	-0.2452
X2	0.3629	-0.2452	0.0366

Вектор оценок параметров $[\beta_2, \beta_3]$: [-1.120772, 0.129577]

1.5 Пункт е

Гипотеза линейности

- H_0 : Зависимость Y от X линейна ($\beta_3 = 0$).
- H_1 : Зависимость нелинейна ($\beta_3 \neq 0$).

Гипотеза независимости

- H_0 : Y не зависит от X линейна ($\beta_2 = \beta_3 = 0$).
- H_1 : Y зависит от X линейна (хотя бы один из $\beta_2, \beta_3 \neq 0$).

Проверка гипотезы линейности ($H_0 : \beta_3 = 0$):

- t-статистика: 0.6775
- p-значение: 0.5014
- Нет оснований отвергать гипотезу о линейности ($p > 0.02$).

Проверка гипотезы независимости ($H_0 : \beta_2 = 0$):

- t-статистика: -0.8509
- p-значение: 0.3991
- Нет оснований отвергать гипотезу о независимости ($p > 0.02$).

1.6 Пункт f

Сравнение моделей по AIC и BIC:

Модель	AIC	BIC
Линейная	232.83	236.66
Квадратичная	234.35	240.08

AIC/BIC линейной модели меньше, она лучше описывает данные.

1.7 Пункт g

Характер зависимости Y от X

- **Линейная модель:**

$$Y = 15.59 - 0.25X, R^2 = 0.014.$$

- Крайне низкий R^2 (1.4%) указывает на отсутствие линейной зависимости.
- Коэффициент $\beta_2 = -0.25$ статистически незначим (доверительный интервал $[-4.29, 2.05]$ включает ноль).

- **Квадратичная модель:**

$$Y = 16.87 - 1.12X + 0.13X^2, R^2 = 0.024.$$

- $R^2 = 2.4\%$ показывает, что модель объясняет лишь незначительную часть вариации.
- Коэффициенты:
 - * $\beta_2 = -1.12$ (линейный член): интервал $[-4.29, 2.05]$ включает ноль.
 - * $\beta_3 = 0.13$ (квадратичный член): интервал $[-0.33, 0.59]$ включает ноль.

Проверка гипотез

Остатки близки к нормальному распределению. Критерий χ^2 не выявил значимых отклонений от нормальности на уровне $\alpha = 0.02$.

Предположение о нормальности ошибок выполняется.

AIC/BIC

Модель	AIC	BIC
Линейная	232.83	236.66
Квадратичная	234.35	240.08

- **Линейная модель** имеет более низкие AIC/BIC, чем квадратичная.

Аномалии в результатах

- **Парадокс низкого R^2 :**
 - Обе модели объясняют менее 3% вариации, что ставит под сомнение их практическую применимость.

Итоговый вывод

- **Отсутствие значимой связи:** Ни линейная, ни квадратичная модели не демонстрируют статистически значимой зависимости Y от X на уровне $\alpha = 0.02$.
- **Рекомендации:**
 - Проверить данные на наличие выбросов или ошибок.
 - Рассмотреть другие предикторы или преобразования.
 - Увеличить объем данных для повышения надежности тестов.

2 Задание 2

2. Результаты статистического эксперимента приведены в таблице 1. Требуется оценить характер зависимости наблюдаемой переменной Y от уровней факторов A и B .
- Сформулировать модель двухфакторного дисперсионного анализа. Построить МНК оценки параметров и несмещенную оценку дисперсии.
 - Проверить визуально согласование исходных данных с предположением аддитивности влияния факторов. Построить графически оценку зависимости уровней фактора A при каждом фиксированном значении фактора B . Наблюдается ли эффект пересечения факторов.
 - Провести анализ ошибок. По гистограмме ошибок оценить визуально согласование с гипотезой нормальности.
 - Провести дисперсионный анализ, начиная с проверки значимости взаимодействий факторов на результаты эксперимента.
 - Выбрать наилучшую модель с использованием AIC и BIC.
 - Интерпретировать полученные результаты. Написать отчет.

Таблица 2 $\alpha_1 = 0.02$; $h = 0.82$.

No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Y	13.17	11.78	11.70	12.54	11.59	11.21	9.57	9.07	10.10	10.60	9.22	7.91	17.17	14.74	16.37	15.34	16.72
A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
B	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3

No	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Y	16.53	11.08	12.01	12.62	11.07	11.36	11.78	14.85	14.60	15.40	13.23	15.32	13.23	21.08	20.70	23.04	21.22
A	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
B	3	4	4	4	4	4	4	1	1	1	1	1	1	2	2	2	2

No	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Y	23.35	22.51	20.08	18.89	21.47	19.55	20.88	20.01	17.06	18.76	18.05	17.83	17.33	18.30
A	2	2	2	2	2	2	2	2	2	2	2	2	2	2
B	2	2	3	3	3	3	3	3	4	4	4	4	4	4

2.1 Пункт а

1. Формулировка модели двухфакторного дисперсионного анализа

Модель с взаимодействием факторов:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

где:

- Y_{ijk} — наблюдаемое значение переменной Y для i -го уровня фактора A , j -го уровня фактора B , k -го повторения,
- μ — общее среднее,
- α_i — эффект i -го уровня фактора A ,
- β_j — эффект j -го уровня фактора B ,
- $(\alpha\beta)_{ij}$ — эффект взаимодействия факторов A и B ,
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ — случайная ошибка.

2. Построение МНК-оценок параметров

Оценки параметров полной модели:

Intercept	11.998333
C(A) [T.2]	2.440000
C(B) [T.2]	-2.586667
C(B) [T.3]	4.146667
C(B) [T.4]	-0.345000
C(A) [T.2] : C(B) [T.2]	10.131667
C(A) [T.2] : C(B) [T.3]	1.561667
C(A) [T.2] : C(B) [T.4]	3.795000

3. Несмещенная оценка дисперсии

Несмещенная оценка дисперсии ошибок:

$$\hat{\sigma}^2 = \frac{SS_{\text{res}}}{df_{\text{res}}} = 0.757,$$

где:

- SS_{res} — сумма квадратов остатков,
- $df_{\text{res}} = n - p$ — степени свободы (n — число наблюдений, p — число параметров).

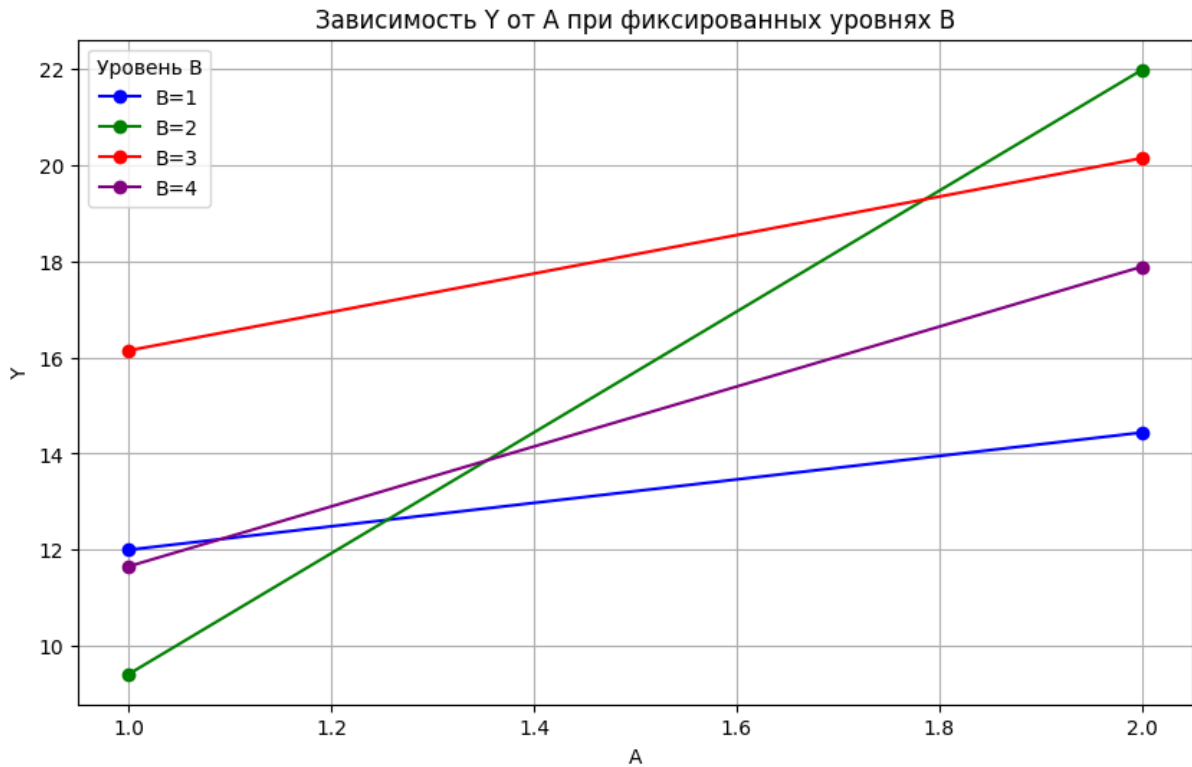
2.2 Пункт b

Сводная таблица средних значений Y :

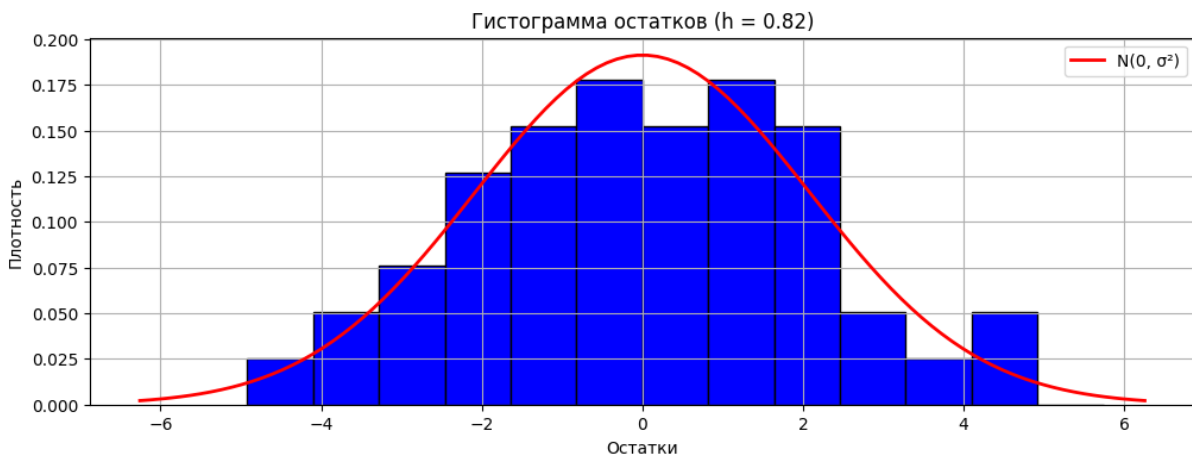
B	1	2	3	4
A				
1	11.998333	9.411667	16.145000	11.653333
2	14.438333	21.983333	20.146667	17.888333

Визуальная проверка аддитивности:

- Пересечение линий: График зависимости Y от A при фиксированных B показывает, что линии для разных уровней B пересекаются, особенно при $B = 4$. Это указывает на наличие взаимодействия между факторами.
- Следствия: Взаимодействие факторов может означать, что влияние одного фактора на зависимую переменную Y зависит от другого фактора.



2.3 Пункт с

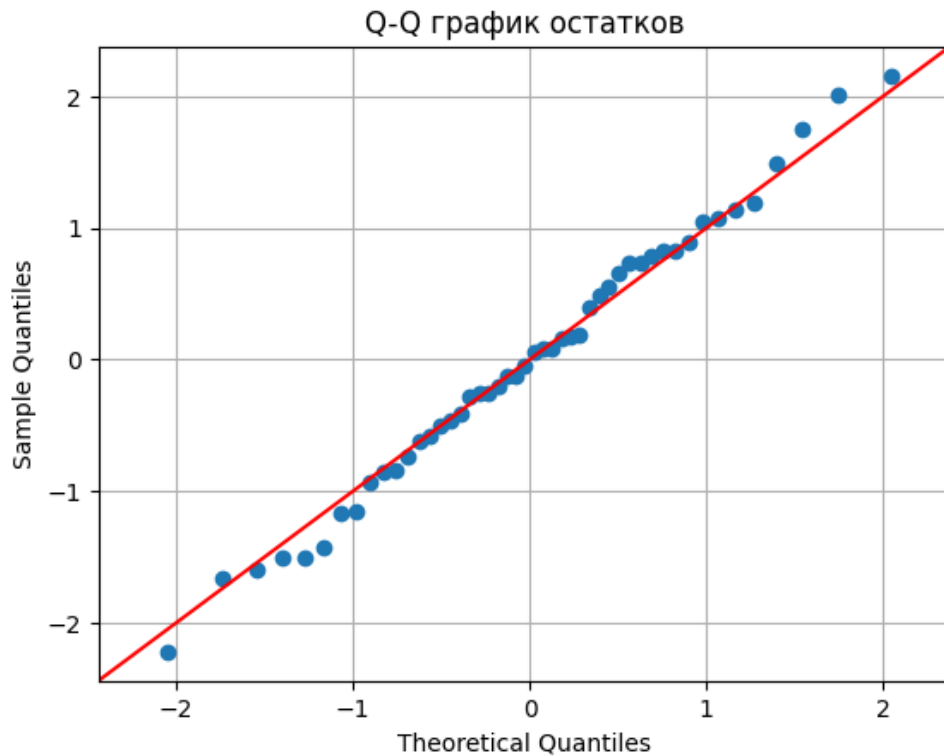


Тест Шапиро-Уилка: $p\text{-value} = 0.949$

Не отвергаем H_0 : остатки нормальны.

Результаты:

- Гистограмма: Распределение остатков близко к нормальному, совпадает с наложенной кривой $N(0, \sigma^2)$.
- Q-Q график: Точки лежат вдоль линии $y = x$, что подтверждает нормальность.
- Тест Шапиро-Уилка: гипотеза о нормальности не отвергается.



2.4 Пункт d

Таблица ANOVA:

df	sum_sq	mean_sq	F	PR(>F)	
C(A)	1.0	478.108752	478.108752	631.694471	4.061068e-26
C(B)	3.0	153.241356	51.080452	67.489330	1.051893e-15
C(A):C(B)	3.0	178.558140	59.519380	78.639144	8.022881e-17
Residual	40.0	30.274683	0.756867	NaN	NaN

Результаты ANOVA

- Фактор A:

$F = 631.69$, $p\text{-value} < 0.001 \rightarrow$ значимо влияет на Y .

- Фактор B:

$F = 67.49$, $p\text{-value} < 0.001 \rightarrow$ значимо влияет на Y .

- Взаимодействие $A \times B$:

$F = 78.64$, $p\text{-value} < 0.001 \rightarrow$ значимо влияет на Y .

- Вывод: На уровне значимости $\alpha = 0.02$ все факторы (A, B) и их взаимодействие **значимо** ($p < 0.02$). Это означает, что влияние фактора A на Y зависит от уровня фактора B, и наоборот.

2.5 Пункт e

Для выбора оптимальной модели используются критерии:

- AIC оценивает баланс между качеством подгонки модели и её сложностью, накладывая штраф за избыточное количество параметров.
- BIC работает аналогично AIC, но применяет более строгий штраф за сложность, особенно при больших объемах данных.

Сравниваем две модели:

1. Полная модель (с взаимодействием):

$$Y \sim A + b + A : B.$$

2. Аддитивная модель (без взаимодействия):

$$Y \sim A + B.$$

Модель	AIC	BIC
Полная	130.10	145.07
Аддитивная	216.79	226.15

Вывод о сравнении моделей

- **Результаты AIC и BIC:**

- Полная модель имеет $AIC = 130.10$, в то время как аддитивная модель имеет $AIC = 216.79$. Это указывает на значительное преимущество полной модели.
- Полная модель также имеет $BIC = 145.07$, а аддитивная модель — $BIC = 226.15$. Разница подтверждает выбор полной модели.

- **Заключение:**

- Полная модель **предпочтительнее**, так как она лучше соответствует данным, что подтверждается меньшими значениями AIC и BIC.
- Аддитивная модель не учитывает взаимодействие факторов.

2.6 Пункт f

1. Основные эффекты факторов A и B

- **Фактор A:** Оказал сильное статистически значимое влияние на Y ($F = 631.69, p < 0.001$).

- **Фактор В:** Также значимо влияет на Y ($F = 67.49, p < 0.001$).

2. Взаимодействие факторов $A \times B$

- **Статистическая значимость:** Взаимодействие значимо ($F = 78.64, p < 0.001$).
- **Визуальное подтверждение:** График зависимости Y от A при фиксированных B показывает пересечение линий (особенно для $B = 4$), что указывает на неаддитивность эффектов.

3. Выбор оптимальной модели

AIC/BIC:

Модель	AIC	BIC
Полная (с взаимодействием)	130.10	145.07
Аддитивная	216.79	226.15

Разница $\Delta AIC = 86.69$ и $\Delta BIC = 81.08$ явно указывает на преимущество полной модели.

Аддитивная модель не учитывает взаимодействие, что приводит к потере информации.

4. Нормальность остатков

- Тест Шапиро-Уилка:

$p\text{-value} = 0.949 \implies$ гипотеза о нормальности остатков не отвергается.

- Графическая проверка: Гистограмма остатков близка к нормальной форме.
- Q-Q график показывает совпадение точек с линией $y = x$.

Рекомендации: Для прогнозирования Y необходимо учитывать взаимодействие $A \times B$, так как его игнорирование приведет к систематической ошибке.

Итоговый вывод

1. Полная модель с взаимодействием предпочтительна по критериям AIC/BIC и объясняет данные лучше аддитивной.
2. Нормальность остатков подтверждена тестами и графиками.

Рекомендации:

- Проверить данные на наличие выбросов для уровня $B = 4$.
- Использовать полную модель для прогнозирования и анализа эффектов.