Министерство науки и высшего образования
Российской Федерации

Федеральное государственное бюджетное
образовательное учреждение высшего образования

«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра теоретической и прикладной информатики

РАСЧЕТНО-ГРАФИЧЕСКОЕ ЗАДАНИЕ

по дисциплине «Статистические методы анализа данных»

Факультет: ПМИ

Группа: ПМИ-02

Вариант: 17

Студент: Сидоров Даниил Игоревич

Преподаватель: Попов Александр Александрович.

Новосибирск

2026

1. Постановка задачи

Провести полный цикл исследований, связанных с построением регрессионной зависимости по имеющимся экспериментальным данным.

1. Выбор предварительного состава регрессоров с использованием корреляционных полей. В качестве регрессоров-кандидатов предположительно могут выступать: свободный член, сами факторы, их взаимодействия (двух-трех факторов), квадраты факторов;
2. Проверка данных на мультиколлинеарность;
3. Проверка данных на гетероскедастичность (предположительно, что чем дальше от центра эксперимента проведено наблюдение, то, возможно, дисперсия его больше);
4. Проверка данных на автокорреляцию (упорядоченность наблюдений по своим номерам считать упорядоченностью по времени);
5. Выбор модели оптимальной сложности с помощью программы ОДА;
6. Нахождение оценок оптимальной модели, учитывающих автокорреляцию;
7. Дополнительная проверка адекватности выбранной модели с использованием повторных наблюдений (последние 6 наблюдений выборки), по которым необходимо будет вычислить оценку дисперсии наблюдений;
8. Построение графиков остатков в различных координатах (по номеру наблюдений, по факторам, по отклику);
9. Определение с помощью построенной модели точки в факторном пространстве, имеющей максимальное значение математического ожидания отклика. Вычисление для этой точки доверительного интервала. Координаты такой точки не обязательно должны совпадать с какой-либо точкой из имеющихся в таблице наблюдений.

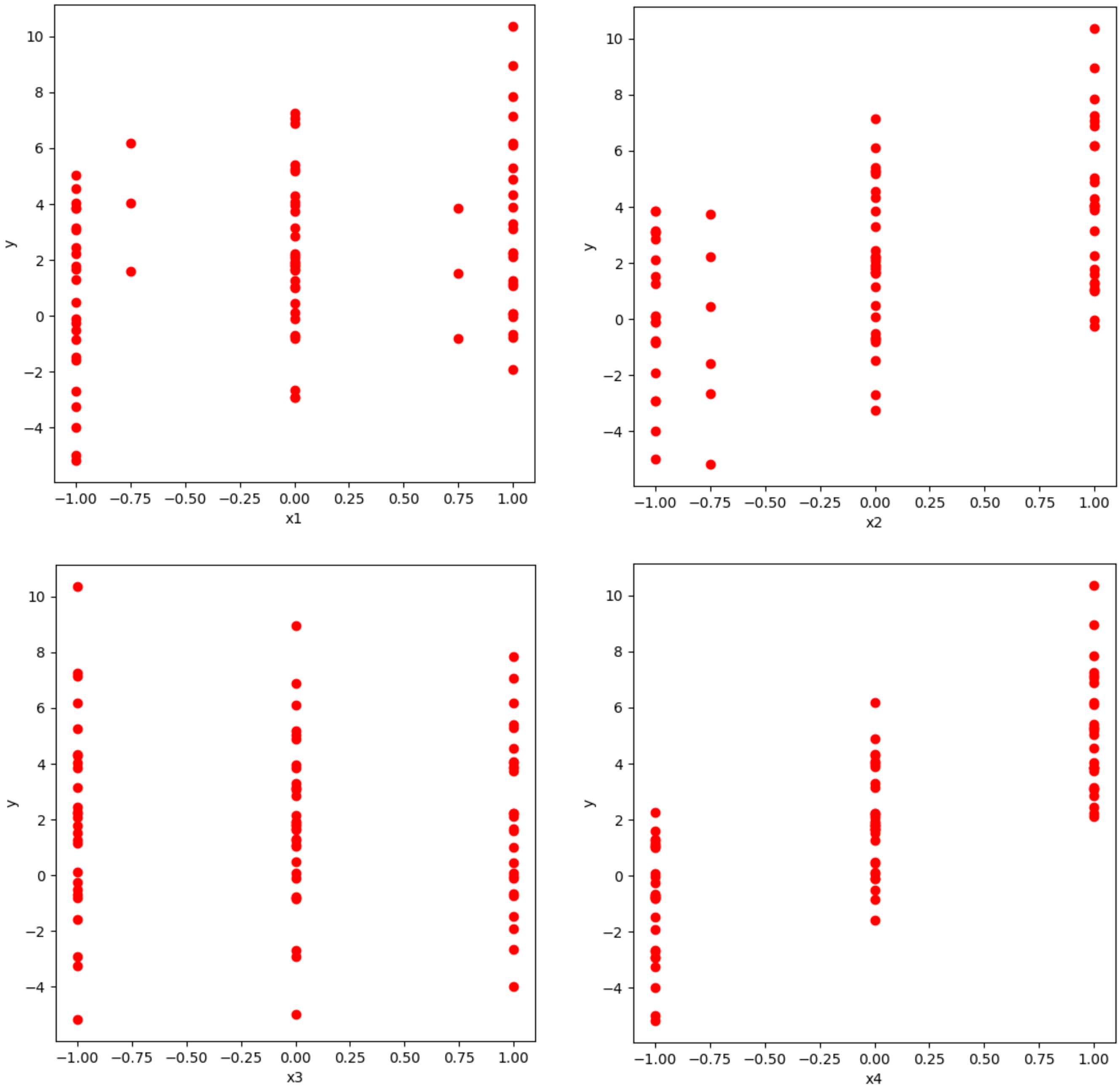
2. Исходные данные

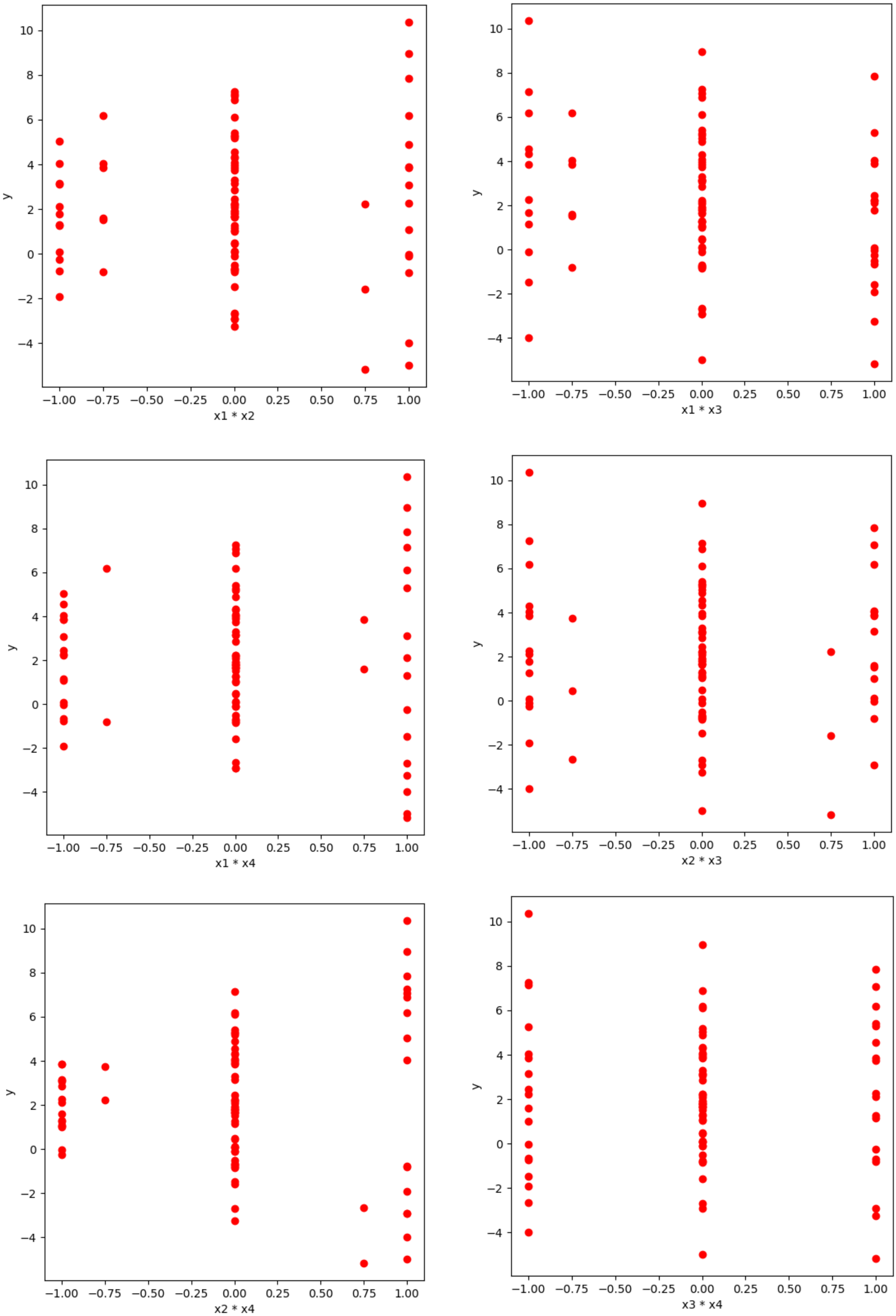
№	X1	X2	X3	X4	y
1	-1,00	-0,75	-1,00	-1,00	-5,17
2	-1,00	-0,75	-1,00	0,00	-1,57
3	-1,00	-0,75	-1,00	1,00	2,21
4	-1,00	-1,00	0,00	-1,00	-4,98
5	-1,00	-1,00	0,00	0,00	-0,86
6	-1,00	-1,00	0,00	1,00	3,06
7	-1,00	-1,00	1,00	-1,00	-3,99
8	-1,00	-1,00	1,00	0,00	-0,09
9	-1,00	-1,00	1,00	1,00	3,87
10	-1,00	0,00	-1,00	-1,00	-3,26
11	-1,00	0,00	-1,00	0,00	-0,53
12	-1,00	0,00	-1,00	1,00	2,44
13	-1,00	0,00	0,00	-1,00	-2,68
14	-1,00	0,00	0,00	0,00	0,50
15	-1,00	0,00	0,00	1,00	3,84
16	-1,00	0,00	1,00	-1,00	-1,47
17	-1,00	0,00	1,00	0,00	1,68
18	-1,00	0,00	1,00	1,00	4,54
19	-1,00	1,00	-1,00	-1,00	-0,25
20	-1,00	1,00	-1,00	0,00	1,78
21	-1,00	1,00	-1,00	1,00	4,05
22	-1,00	1,00	0,00	-1,00	1,29
23	-1,00	1,00	0,00	0,00	3,15
24	-1,00	1,00	0,00	1,00	5,02
25	-0,75	1,00	1,00	-1,00	1,59
26	-0,75	1,00	1,00	0,00	4,03
27	-0,75	1,00	1,00	1,00	6,18
28	0,00	-1,00	-1,00	-1,00	-2,91
29	0,00	-1,00	-1,00	0,00	0,10
30	0,00	-1,00	-1,00	1,00	3,16
31	0,00	-1,00	0,00	-1,00	-2,90
32	0,00	-1,00	0,00	0,00	-0,11
33	0,00	-1,00	0,00	1,00	2,85
34	0,00	-0,75	1,00	-1,00	-2,66
35	0,00	-0,75	1,00	0,00	0,44
36	0,00	-0,75	1,00	1,00	3,75
37	0,00	0,00	-1,00	-1,00	-0,68
38	0,00	0,00	-1,00	0,00	2,08
39	0,00	0,00	-1,00	1,00	5,26
40	0,00	0,00	0,00	-1,00	-0,79
41	0,00	0,00	0,00	0,00	2,16
42	0,00	0,00	0,00	1,00	5,19
43	0,00	0,00	1,00	-1,00	-0,73
44	0,00	0,00	1,00	0,00	2,21
45	0,00	0,00	1,00	1,00	5,40
46	0,00	1,00	-1,00	-1,00	1,26
47	0,00	1,00	-1,00	0,00	4,31
48	0,00	1,00	-1,00	1,00	7,25
49	0,00	1,00	0,00	-1,00	1,05
50	0,00	1,00	0,00	0,00	3,97
51	0,00	1,00	0,00	1,00	6,90
52	0,00	1,00	1,00	-1,00	1,02
53	0,00	1,00	1,00	0,00	4,06

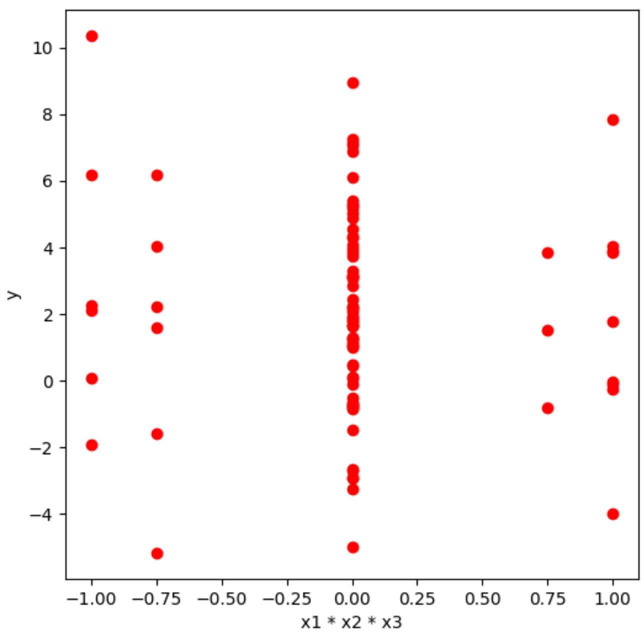
54	0,00	1,00	1,00	1,00	7,08
55	0,75	-1,00	-1,00	-1,00	-0,80
56	0,75	-1,00	-1,00	0,00	1,53
57	0,75	-1,00	-1,00	1,00	3,86
58	1,00	-1,00	0,00	-1,00	-0,78
59	1,00	-1,00	0,00	0,00	1,27
60	1,00	-1,00	0,00	1,00	3,10
61	1,00	-1,00	1,00	-1,00	-1,93
62	1,00	-1,00	1,00	0,00	0,09
63	1,00	-1,00	1,00	1,00	2,13
64	1,00	0,00	-1,00	-1,00	1,15
65	1,00	0,00	-1,00	0,00	4,35
66	1,00	0,00	-1,00	1,00	7,16
67	1,00	0,00	0,00	-1,00	0,10
68	1,00	0,00	0,00	0,00	3,29
69	1,00	0,00	0,00	1,00	6,11
70	1,00	0,00	1,00	-1,00	-0,67
71	1,00	0,00	1,00	0,00	2,23
72	1,00	0,00	1,00	1,00	5,30
73	1,00	1,00	-1,00	-1,00	2,26
74	1,00	1,00	-1,00	0,00	6,20
75	1,00	1,00	-1,00	1,00	10,36
76	1,00	1,00	0,00	-1,00	1,09
77	1,00	1,00	0,00	0,00	4,87
78	1,00	1,00	0,00	1,00	8,95
79	1,00	1,00	1,00	-1,00	-0,04
80	1,00	1,00	1,00	0,00	3,90
81	1,00	1,00	1,00	1,00	7,86
82	0	0	0	0	1,83
83	0	0	0	0	1,64
84	0	0	0	0	1,69
85	0	0	0	0	1,76
86	0	0	0	0	1,90
87	0	0	0	0	1,93

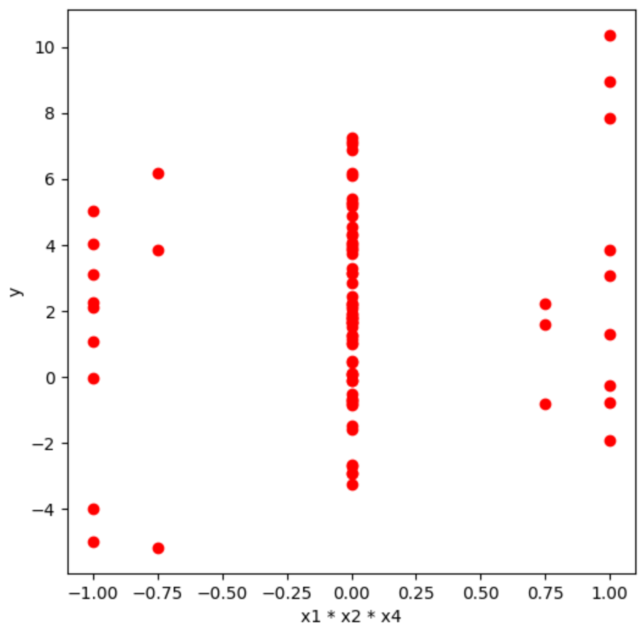
3. Ход работы

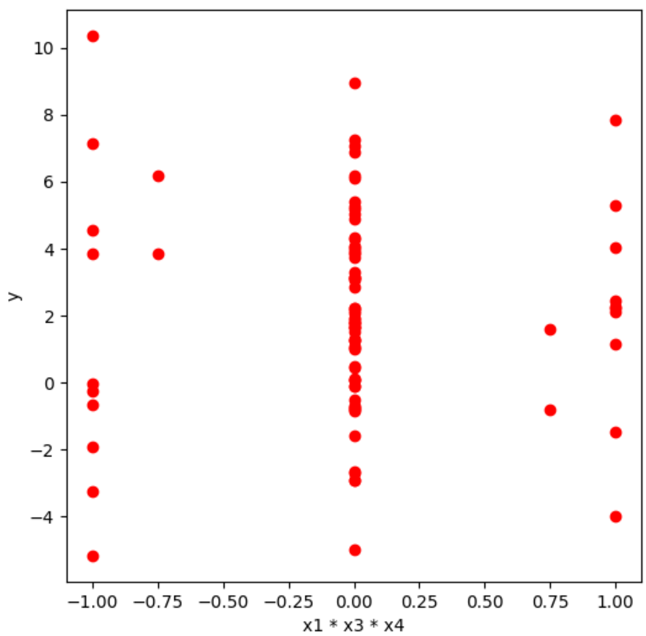
1. Корреляционные поля

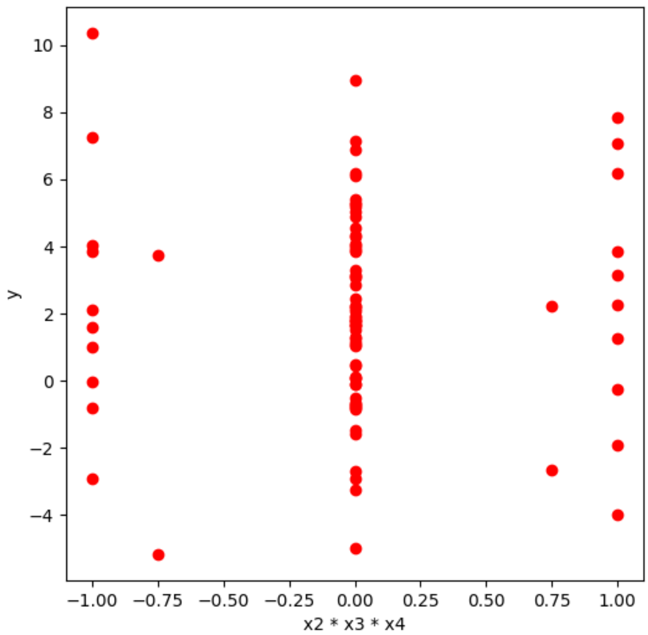


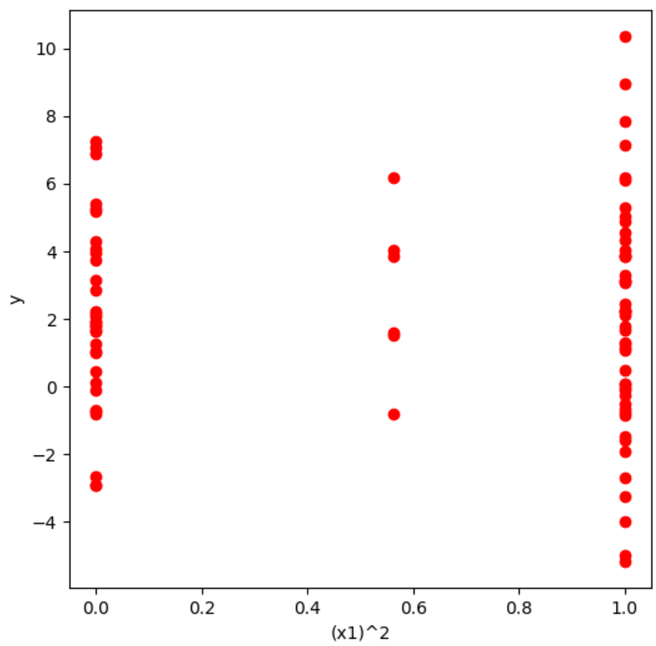

y
10
8
6
4
2
0
−2
−4
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
x1 * x2
x1 * x3
x1 * x4
x2 * x3
x2 * x4
x3 * x4

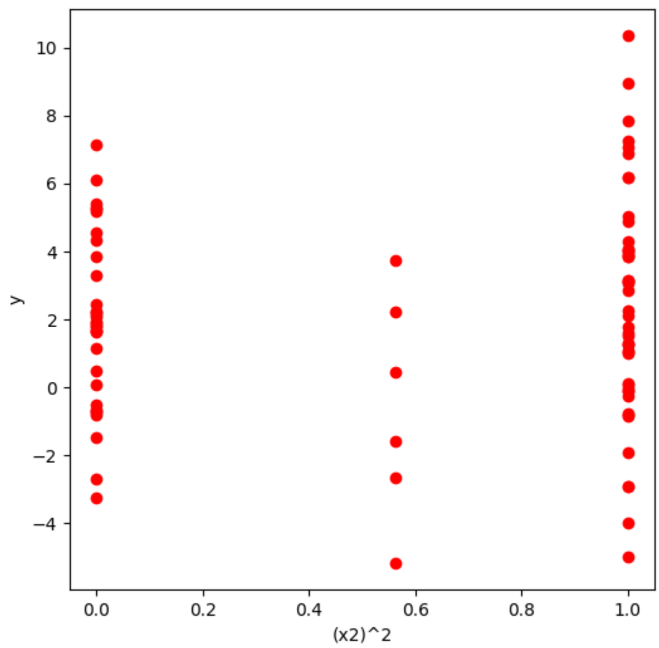

y
x1 * x2 * x3
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
−4 −2 0 2 4 6 8 10

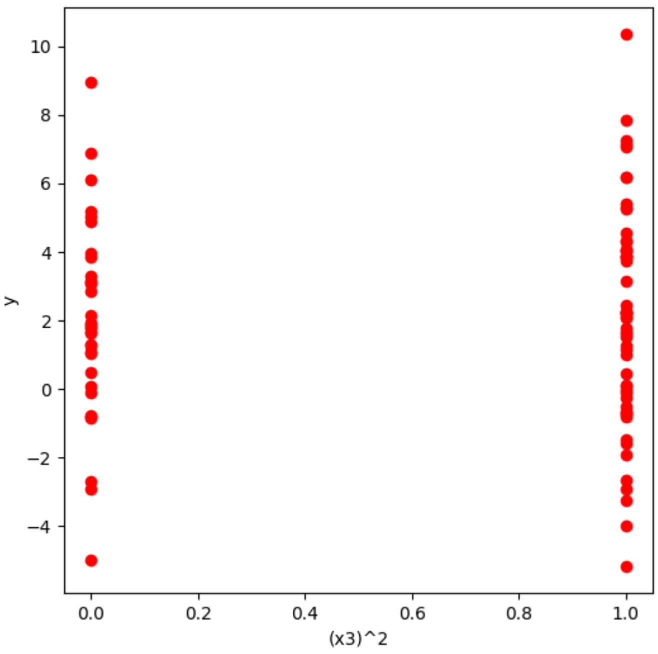

y
x1 * x2 * x4
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
−4 −2 0 2 4 6 8 10

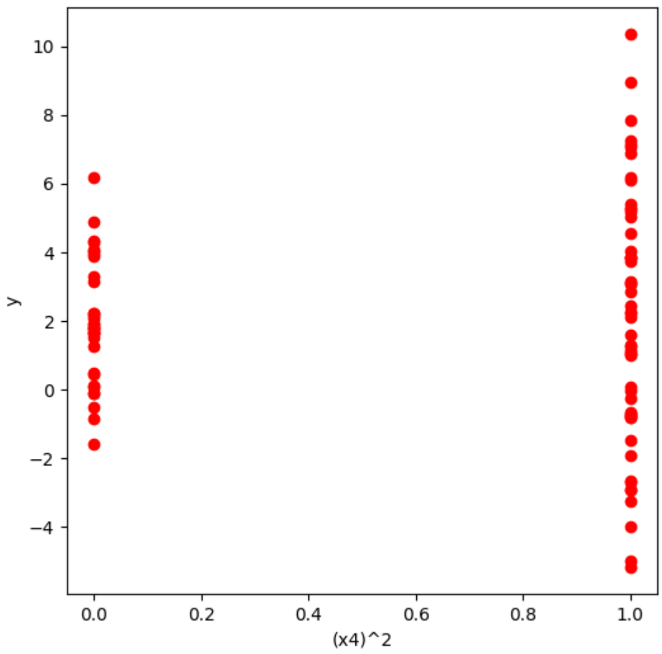

y
x1 * x3 * x4
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
−4 −2 0 2 4 6 8 10


y
x2 * x3 * x4
−1.00 −0.75 −0.50 −0.25 0.00 0.25 0.50 0.75 1.00
−4 −2 0 2 4 6 8 10


y
(x1)^2
0.0 0.2 0.4 0.6 0.8 1.0
−4 −2 0 2 4 6 8 10


y
(x2)^2
0.0 0.2 0.4 0.6 0.8 1.0
−4 −2 0 2 4 6 8 10





При построении учитывался тот факт, что в качестве регрессоров-кандидатов предположительно могут выступать: свободный член, сами факторы, их взаимодействия (двух-трех факторов), квадраты факторов.

Благодаря анализу корреляционных полей была выбрана предварительная модель, имеющая следующий вид:

$$f(x) = (1, x_1, x_2, x_4, x_1x_3, x_1x_2x_3, x_1x_2x_4).$$

Эти регрессоры имеют сильную связь с выходными данными, точки на соответствующих корреляционных полях формируют общий тренд.

2. Проверка данных на мультиколлинеарность

Показатели, характеризующим мультиколлинеарность.

1. Определитель нормированной информационной матрицы:

$$\left|\frac{X^TX}{trX^TX}\right| = 5.25471e - 07;$$

2. Минимальное собственное число нормированной информационной матрицы:

$$\lambda_{min}\left(\frac{X^TX}{trX^TX}\right) = 6.1697e - 02;$$

3. Мера обусловленности матрицы по Нейману-Голдстрейну.

$$\frac{\lambda_{max}}{\lambda_{min}} = 4.43239;$$

4. Максимальная парная сопряженность. Матрица сопряженности в общем виде:

$$R = \begin{bmatrix} 1 & \dots & r_{1m} \\ r_{21} & \dots & r_{2m} \\ r_{m1} & \dots & 1 \end{bmatrix},$$

где $r_{ij} = cos(x_i\,, x_j);$

Косинус угла между векторами можно интерпретировать как косинусное сходство между векторами. Построенная матрица сопряженности:

```
[ 1.          0.01459854  0.          0.         -0.10512484  0.        ]
[ 0.01459854  1.          0.          0.01811235  0.00584027  0.        ]
[ 0.          0.          1.          0.          0.          0.01471703]
[ 0.          0.01811235  0.          1.          0.02898399  0.        ]
[-0.10512484  0.00584027  0.          0.02898399  1.          0.        ]
[ 0.          0.          0.01471703  0.          0.          1.        ]]
```

Максимальная парная сопряженность:

$$\max_{i,j}|r_{ij}| = 0.10512, i \neq j;$$

5. Максимальная сопряженность. В качестве меры мультиколлинеарности можно взять $\max_i |R_i|$. Вычислить R_i можно следующим способом:

$$R^2{}_t = 1 - \frac{1}{R_{ii}^{-1}}, i = 1..m,$$

где R_{ii}^{-1} – диагональный элемент матрицы, обратный к сопряженной.

Максимальная сопряженность: 0.01194.

Рассмотрев множество показателей, особенно важными из которых являются: минимальное собственное число, мера обусловленности матрицы по Нейману-Голдстрейну и максимальная сопряженность, можно сделать вывод об <u>отсутствии мультиколлинеарности</u>.

3. Проверка данных на гетероскедастичность

1) Тест Бреуша-Пагана:

Вектор известных переменных: $z_t^T = \left(1, \sqrt{x_3^2 + \mathrm{x}_4^2}\right)$;
Вектор неизвестных параметров: $\alpha^T = (\alpha_0, \alpha_1)$;

Гипотеза о гомоскедастичности в нашем случае имеет вид: $\alpha_1 = 0$.

- оценим исходное уравнения по МНК:

$$e_t = y_t - f(x_t)^T \hat{\theta};$$

- оценим дисперсию:

$$\tilde{\sigma}^2 = \sum \frac{{e_t}^2}{n} = 0.0383.$$

- построим регрессию с откликом:

$$c_t = \frac{e_t^2}{\tilde{\sigma}^2};$$

- найдем предсказанные значения нормированных квадратов остатков:

$$\hat{c}_t = \hat{\alpha}^{\mathrm{T}} z_t;$$

Гипотеза о гомоскедастичности будет принята, если

$$\frac{ESS}{2} = \frac{\sum (c_t - \bar{c})^2}{2} \sim \chi^2_{0.05,1} = 3.95059;$$

Получаем:

$$\frac{\sum (c_t - \bar{c})^2}{2} = 0.94603 < 3.95059;$$

Гипотеза о гомоскедастичности принимается.

2) Тест Голдфельда-Квандтона:

Пусть источник нарушения гомоскедастичности взят в форме:

$$E\left(\varepsilon_i^2\right) = \sqrt{\mathrm{x}_3^2 + \mathrm{x}_4^2}\,;$$

- Упорядочим последовательность наблюдений в соответствии с этой величиной:

- опустим n_c наблюдений, оказавшихся в середине упорядоченной выборки:

$$n_c = \frac{n}{3} = 29;$$

- Гипотеза о гомоскедастичности будет принята, если:

$$\frac{RSS_2}{RSS_1} \sim F_{\alpha, \frac{n-n_c-2k}{2}, \frac{n-n_c-2k}{2}} = \ 2.04777;$$

где k = m – число параметров в регрессии. Получаем:

$$\frac{RSS_2}{RSS_1} = 0.79866 < 2.04777;$$

Гипотеза о гомоскедастичности принимается.

На основе теста Бреуша-Пагана и Голдфельда-Квандтона можно сделать вывод об <u>отсутствии гетероскедастичности.</u>

4. Проверка данных на автокорреляцию

Тест Дарбина-Уотсона:

$$DW = \sum_{t=2}^{n} (\hat{e}_t - \hat{e}_{t-1})^2 \,/ \sum_{t=1}^{n} \hat{e}_t^{\ 2} = 0.54192;$$

При отсутствии автокорреляции статистика DW близка к двум, при положительной автокорреляции она близка к нулю, а при отрицательной к 4. Дарбин и Уотсон показали, что есть определенные границы, зависящие только от числа наблюдений n, количества регрессоров m и уровня значимости.

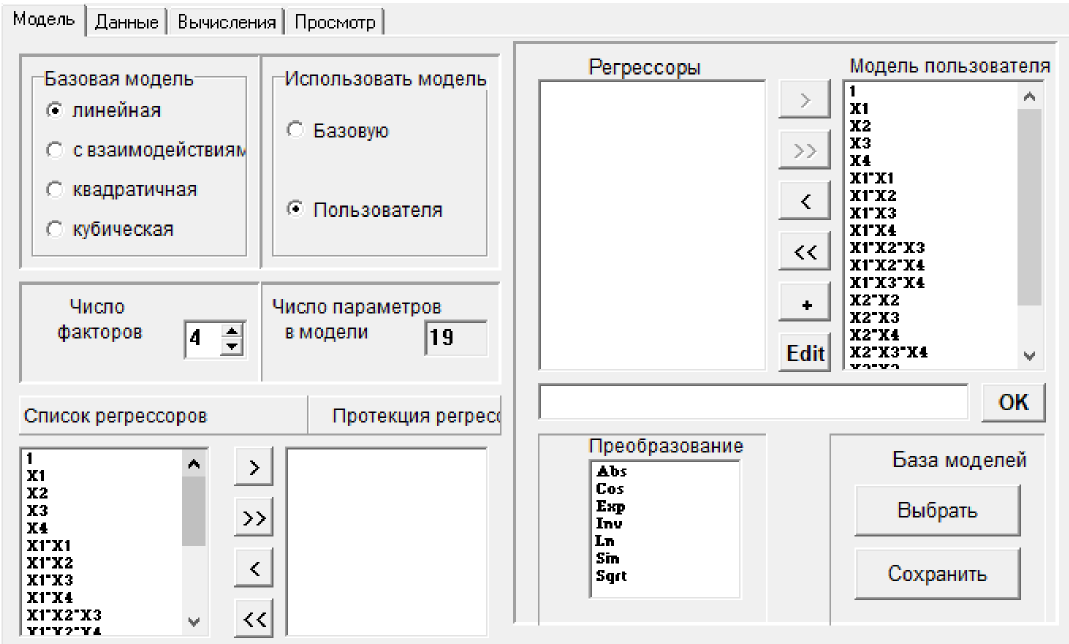
Найдем эти границы по таблице: $d_L \approx 1.57, d_u \approx 1.75$.

Получаем случай $0 < DW < d_L$, исходя из этого имеем положительную автокорреляцию ошибок.

5. Выбор оптимальной модели по ОДА

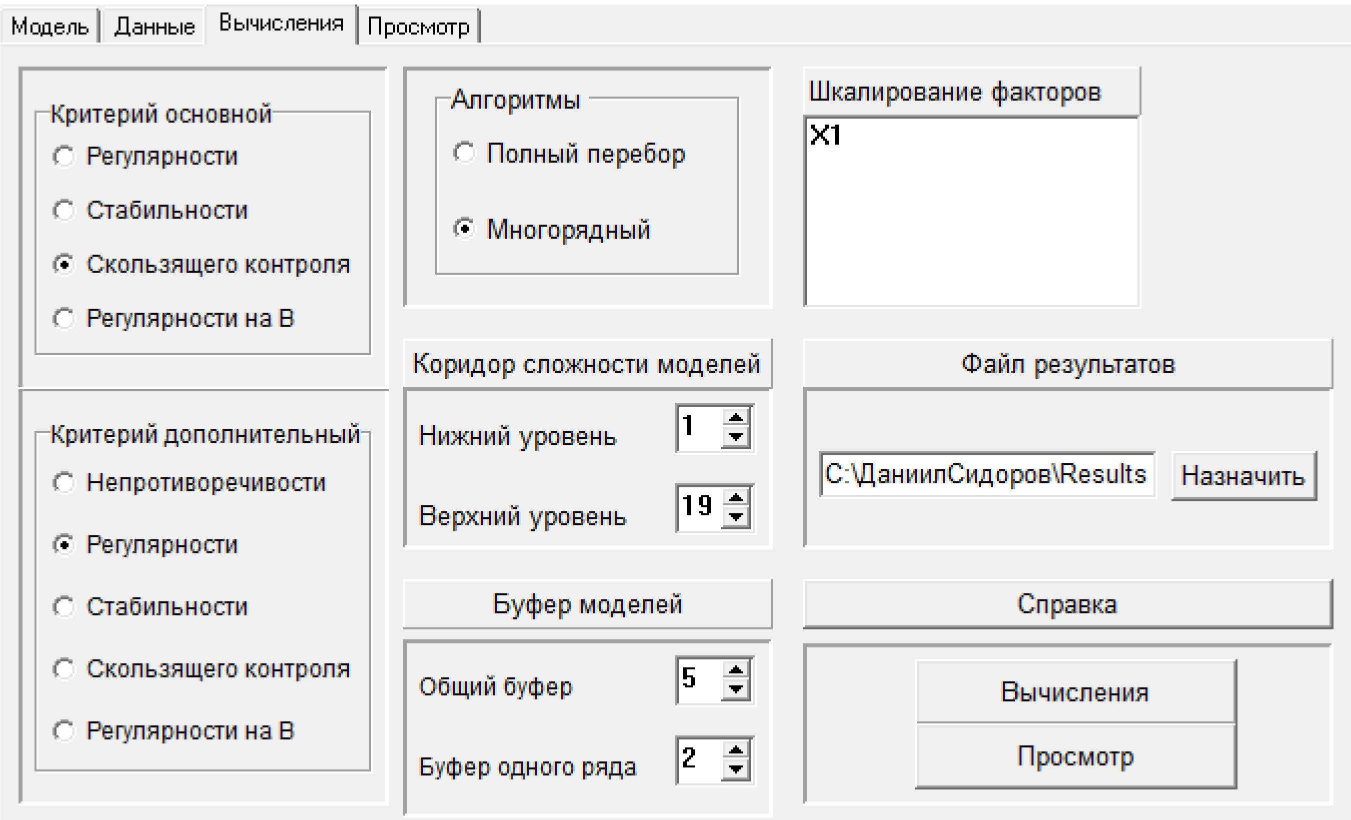
Используем программу ОДА для поиска модели оптимальной сложности.

1) Введем пользовательскую модель в программную систему ОДА. Она состоит из всех регрессоров, которые использовались в процессе анализа корреляционных полей + свободный член:



2) Вставим исходные данные:

3) Выполним вычисления. Основным критерием был выбран критерий скользящего контроля, а дополнительным - критерий регулярности. Число лучших моделей, отбираемых по основному критерию, равно 5. Среди них будет отбираться наиболее подходящая:



4) Файл результата:

Вектор оценок параметров полной модели
 1.96E+0000 1.14E+0000 1.99E+0000-2.61E-0002 3.01E+0000
-9.19E-0002-6.57E-0003-1.02E+0000-5.38E-0003-6.47E-0002
 9.75E-0001-1.83E-0002 7.42E-0002-4.50E-0002-3.78E-0003
-1.90E-0002 4.90E-0002 1.04E-0002 5.49E-0002
 Сумма квадратов отклонений для полной модели= 2.87E+0000

 Список регрессоров:
 1): 1
 2): X1
 3): X2
 4): X3
 5): X4
 6): X1*X1
 7): X1*X2
 8): X1*X3
 9): X1*X4
10): X1*X2*X3
11): X1*X2*X4
12): X1*X3*X4
13): X2*X2
14): X2*X3
15): X2*X4
16): X2*X3*X4
17): X3*X3
18): X3*X4

19): X4*X4

Список регрессоров лучших моделей

m	1 -я мод.	2 -я мод.	3 -я мод.	4 -я мод.	5 -я мод.
1):	1	1	1	1	1
2):	2	2	2	2	2
3):	3	3	3	3	3
4):	4	4	0	0	0
5):	5	5	5	5	5
6):	6	6	6	6	6
7):	0	0	0	0	0
8):	8	8	8	8	8
9):	0	0	0	0	0
10):	10	10	10	10	0
11):	11	11	11	11	11
12):	0	0	0	0	0
13):	13	13	13	13	13
14):	14	14	14	0	14
15):	0	0	0	0	0
16):	0	0	0	0	0
17):	0	0	0	0	0
18):	0	0	0	0	0
19):	0	19	0	0	0

Значения 1внешнего критерия:
0.0431 0.0433 0.0429 0.0434 0.0436
Лучшая по 1 критерию 3 модель.

Значения 2 внешнего критерия:
12.8188 12.7728 7.2820 8.0860 14.1873
Лучшая по 2 критерию 3 модель.

Остаточные суммы квадратов для лучших моделей:
3.0098 2.9375 3.0468 3.1139 3.1333

Критерий Маллоуса для лучших моделей:
6.3764 6.6623 5.2546 4.8445 5.3040

Оценки параметров лучших моделей:

1	2.0152	1.9859	2.0153	2.0108	2.0180
2	1.1408	1.1410	1.1400	1.1391	1.1445
3	1.9939	1.9941	1.9939	1.9940	1.9939
4	-0.0262	-0.0262	0.0000	0.0000	0.0000
5	3.0050	3.0050	3.0050	3.0050	3.0050
6	-0.0832	-0.0900	-0.0834	-0.0753	-0.0843
7	0.0000	0.0000	0.0000	0.0000	0.0000
8	-1.0159	-1.0156	-1.0158	-1.0174	-1.0173
9	0.0000	0.0000	0.0000	0.0000	0.0000
10	-0.0652	-0.0647	-0.0663	-0.0694	0.0000
11	0.9771	0.9771	0.9771	0.9771	0.9771
12	0.0000	0.0000	0.0000	0.0000	0.0000
13	0.0841	0.0774	0.0843	0.0838	0.0796
14	-0.0442	-0.0449	-0.0453	0.0000	-0.0481
15	0.0000	0.0000	0.0000	0.0000	0.0000
16	0.0000	0.0000	0.0000	0.0000	0.0000
17	0.0000	0.0000	0.0000	0.0000	0.0000
18	0.0000	0.0000	0.0000	0.0000	0.0000
19	0.0000	0.0602	0.0000	0.0000	0.0000

Сравнение прогноза и отклика для модели 3

N	Y	Y^	Y-Y^	(Y-Y^)/Y
1	-3.99	-4.10	-0.11	2.88
2	3.87	3.86	-0.01	-0.28
3	-0.25	-0.19	0.06	-22.11
4	4.05	3.86	-0.19	-4.66
5	-1.93	-1.77	0.16	-8.29
6	2.13	2.29	0.16	7.32
7	2.26	2.30	0.04	1.57
8	10.36	10.26	-0.10	-0.97
9	-0.04	0.04	0.08	-201.15
10	7.86	8.00	0.14	1.84
11	-5.17	-5.39	-0.22	4.33
12	2.21	2.08	-0.13	-5.81
13	1.59	1.69	0.10	6.01
14	6.18	6.23	0.05	0.81
15	-0.80	-0.69	0.11	-13.54
16	3.86	3.85	-0.01	-0.19
17	-4.98	-5.10	-0.12	2.41
18	3.06	2.86	-0.20	-6.39
19	-0.09	-0.12	-0.03	36.54
20	-1.57	-1.66	-0.09	5.49
21	-0.86	-1.12	-0.26	29.97
22	-3.26	-3.23	0.03	-0.95
23	-0.53	-0.22	0.31	-57.74
24	2.44	2.78	0.34	13.98
25	-2.68	-2.21	0.47	-17.42
26	0.50	0.79	0.29	58.37
27	3.84	3.80	-0.04	-1.12
28	-1.47	-1.20	0.27	-18.55
29	1.68	1.81	0.13	7.60
30	4.54	4.81	0.27	6.01
31	1.78	1.83	0.05	2.99
32	1.29	0.84	-0.45	-34.72
33	3.15	2.87	-0.28	-8.89
34	5.02	4.90	-0.12	-2.43

35	4.03	3.96	-0.07	-1.79
36	-2.91	-2.94	-0.03	1.19
37	0.10	0.06	-0.04	-39.68
38	3.16	3.07	-0.09	-3.00
39	-2.90	-2.90	0.00	-0.02
40	-0.11	0.11	0.22	-196.04
41	2.85	3.11	0.26	9.15
42	-2.66	-2.40	0.26	-9.63
43	0.44	0.60	0.16	36.64
44	3.75	3.61	-0.14	-3.83
45	-0.68	-0.99	-0.31	45.55
46	2.08	2.02	-0.06	-3.11
47	5.26	5.02	-0.24	-4.56
48	-0.79	-0.99	-0.20	25.29
49	2.16	2.02	-0.14	-6.70
50	5.19	5.02	-0.17	-3.27
51	-0.73	-0.99	-0.26	35.58
52	2.21	2.02	-0.19	-8.81
53	5.40	5.02	-0.38	-7.03
54	1.26	1.13	-0.13	-10.02
55	4.31	4.14	-0.17	-3.97
56	7.25	7.14	-0.11	-1.46
57	1.05	1.09	0.04	3.66
58	3.97	4.09	0.12	3.11
59	6.90	7.10	0.20	2.88
60	1.02	1.04	0.02	2.26
61	4.06	4.05	-0.01	-0.29
62	7.08	7.05	-0.03	-0.38
63	1.53	1.58	0.05	3.30
64	-0.78	-0.87	-0.09	10.99
65	1.27	1.16	-0.11	-8.49
66	3.10	3.19	0.09	2.91
67	0.09	0.26	0.17	186.67
68	1.15	1.08	-0.07	-5.86
69	4.35	4.09	-0.26	-6.03
70	7.16	7.09	-0.07	-0.94
71	0.10	0.07	-0.03	-33.19
72	3.29	3.07	-0.22	-6.63
73	6.11	6.08	-0.03	-0.54
74	-0.67	-0.95	-0.28	41.65
75	2.23	2.06	-0.17	-7.80
76	5.30	5.06	-0.24	-4.51
77	6.20	6.28	0.08	1.25
78	1.09	1.17	0.08	7.15
79	4.87	5.15	0.28	5.75
80	8.95	9.13	0.18	2.03
81	3.90	4.02	0.12	3.14
82	1.83	2.02	0.19	10.12
83	1.64	2.02	0.38	22.88
84	1.69	2.02	0.33	19.25
85	1.76	2.02	0.26	14.50
86	1.90	2.02	0.12	6.07
87	1.93	2.02	0.09	4.42

5) Аналогичные действия для предварительной модели:

Значения 1 внешнего критерия:

0.0447

Значения 2 внешнего критерия:

4.2688

Остаточные суммы квадратов для лучших моделей:

3.3366

Критерий Маллоуса для лучших моделей:

7.0000

6) Сравнение моделей

$$f_{\text{предв}}(x) = (1, x_1, x_2, x_4, x_1x_3, x_1x_2x_3, x_1x_2x_4).$$

$$f_{\text{опт}}(x) = (1, x_1, x_2, x_4, x_1x_3, x_2x_3, x_1x_2x_3, x_1x_2x_4, x_1^2, x_2^2);$$

Оптимальная модель оказывается лучше предварительной.

6. Нахождение оценок для оптимальной модели

Ранее мы обнаружили положительную автокорреляцию ошибок. Это необходимо учитывать:

$$\hat{\theta} = (X^{*T}X^*)^{-1}X^{*T}y^*,$$

где $y^* = Py, X^* = PX$;

Матрица P определяется как:

$$P = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1 & 0 & \dots & 0 & 0 \\ 0 & -\rho & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}$$

Оценка коэффициента автокорреляции определялась из соотношения:

$$DW = \sum_{t=2}^{n} (\hat{e}_t - \hat{e}_{t-1})^2 \,/ \sum_{t=1}^{n} \hat{e}_t{}^2 \cong 2(1-\hat{p});$$

В результате получаем следующие оценки:

$$\hat{\theta} = (2.046, 1.111, 2.006, 3.010, -0.979, -0.022, -0.016, 1.005, -0.07, 0.021)^{\mathrm{T}};$$

7. Проверка адекватности выбранной модели

Пусть имеются повторные наблюдения. Тогда всего $\mathrm{N}_c = 82$ серий и в каждую v-ю серию входят r_v повторных опытов. Общее число опытов $\mathrm{N} = 87$.

Гипотеза об адекватности модели имеет вид:

$$\text{Нулевая гипотеза:}$$

$$H_0: E\{\hat{\sigma}_{LF}^2\} = E\{\hat{\sigma}_E^2\}$$

$$\text{Альтернативная:}$$

$$H_1: E\{\hat{\sigma}_{LF}^2\} > E\{\hat{\sigma}_E^2\}$$

Вычислим оценки дисперсий: на основе меры степени неадекватности представления экспериментальных данных с помощью выбранной модели и на основе ошибки эксперимента в случае, когда имеются повторные опыты:

$$\hat{\sigma}_{LF}^2 = SS_{LF}/f_{LF} = \left(Q\bar{y} - X\hat{\theta}\right)^T\left(Q\bar{y} - X\hat{\theta}\right)/(N_c - m) = 11.35128;$$

$$\hat{\sigma}_E^2 = SS_E/f_E = (y - Q\bar{y})^T(y - Q\bar{y})/(N - N_c) = 164.49767;$$

Матрица Q имеет размер N x N_c и в нашем случае имеет вид:

$$Q = \begin{pmatrix} 1 & . & . & . & 0 \\ . & 1 & & & . \\ . & & . & & . \\ . & & & . & . \\ & & & & 1 \\ & & & & 1 \\ & & & & 1 \\ 0 & . & . & . & 1 \\ & & & & 1 \\ & & & & 1 \end{pmatrix}$$

Гипотеза H_0 не отвергается, если $\hat{\sigma}_{LF}^2/\hat{\sigma}_E^2 \leq F_{\alpha,f_{LF},\mathrm{f_E}}$.

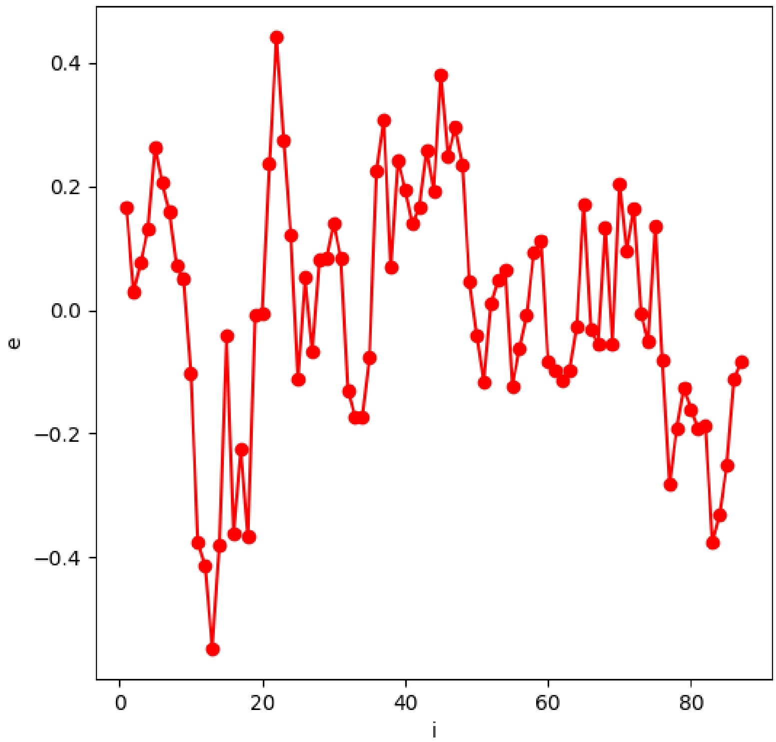
$$\frac{\hat{\sigma}_{LF}^2}{\hat{\sigma}_E^2} = 0.06901;$$

$$F_{\alpha,f_{LF},\mathrm{f_E}} = 4.42046.$$

<u>Модель адекватна.</u>

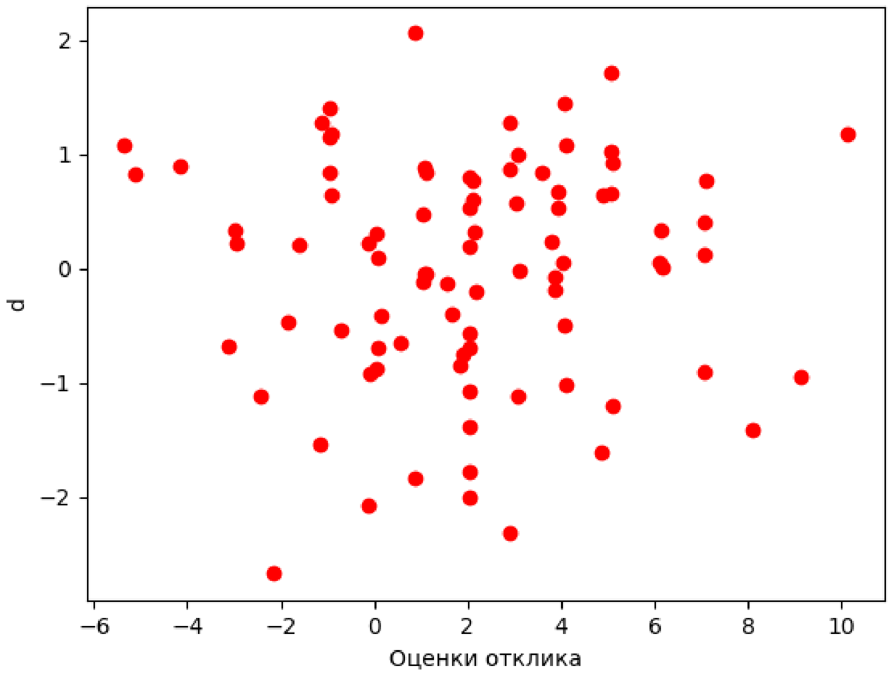
8. Построение графиков остатков

1) График остатков от времени:



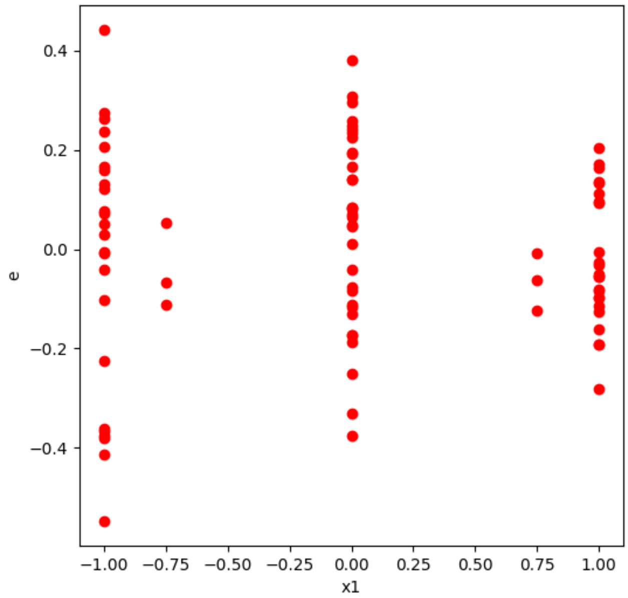
Отчетливо видна положительная автокорреляция.

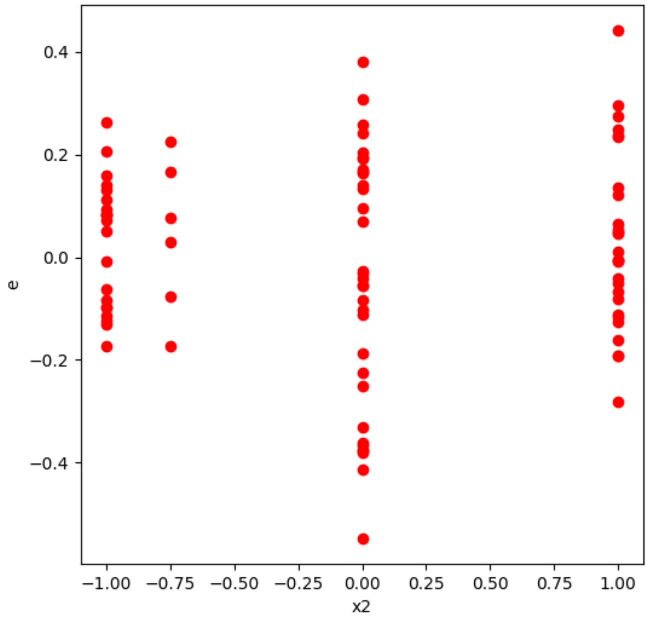
2) График остатков от оценок отклика:

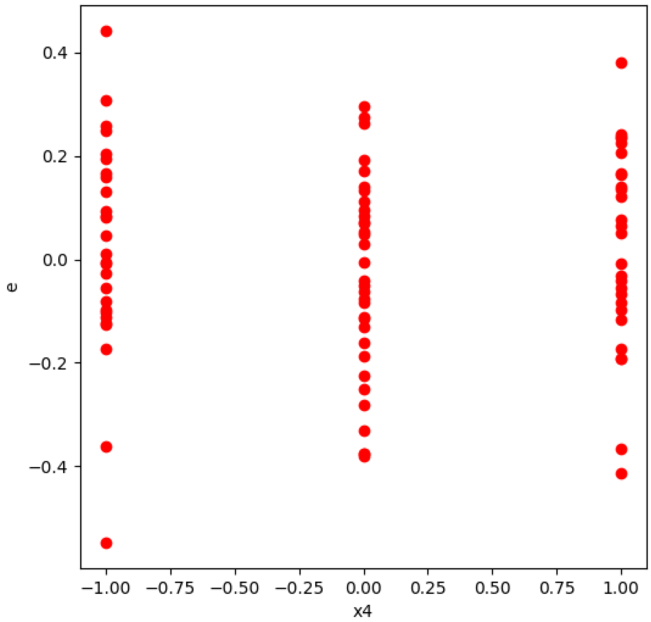


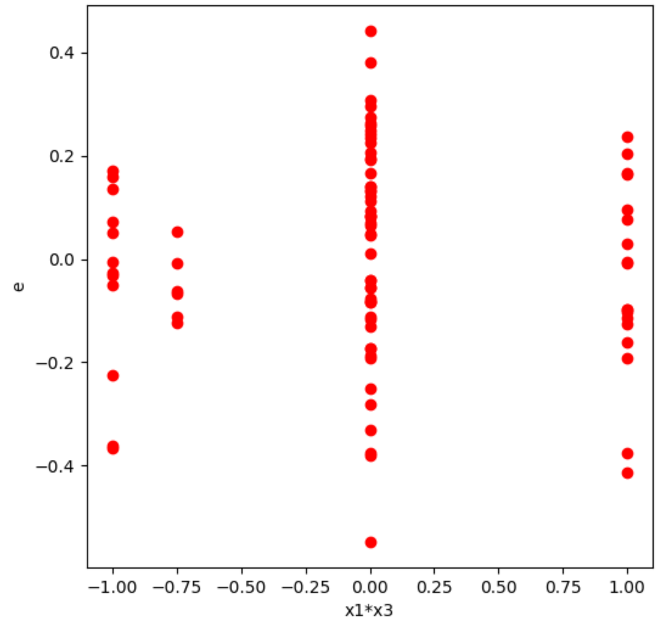
Явное непостоянство отсутствует, что указывает на гомоскедастичность модели.

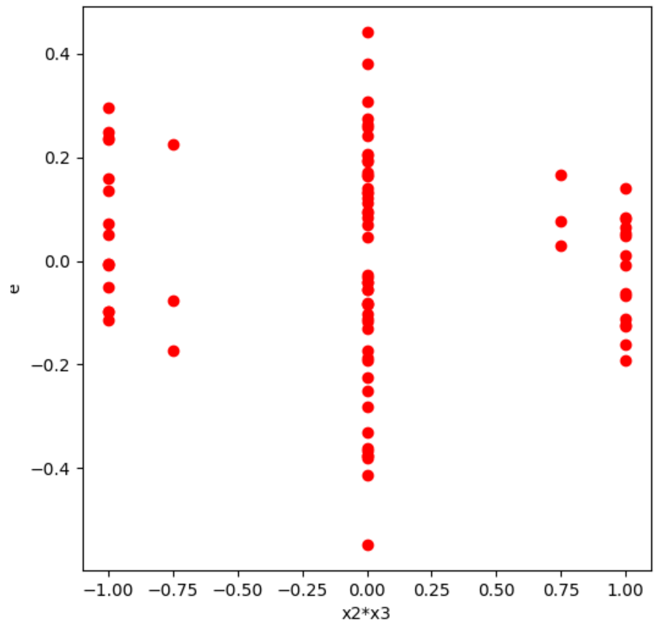
3) Графики остатков от регрессоров:

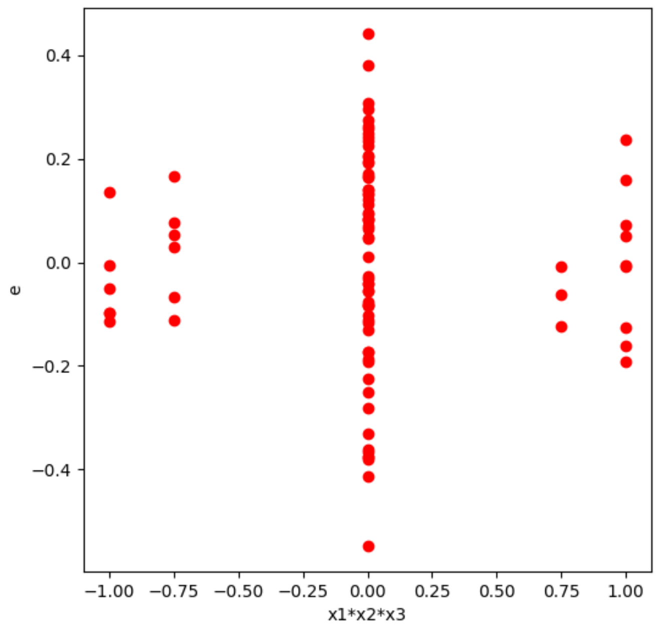


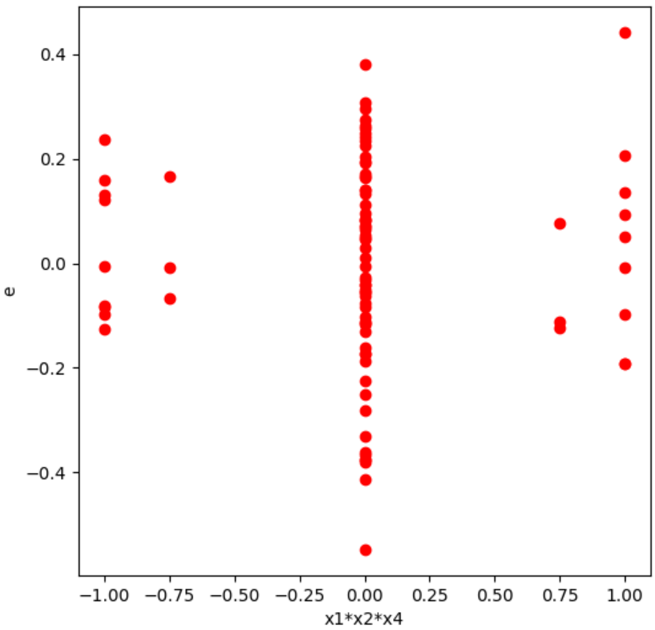


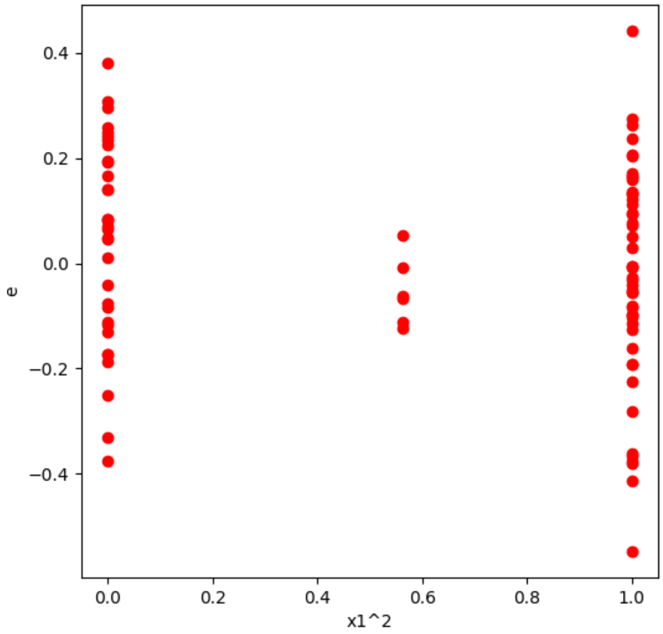


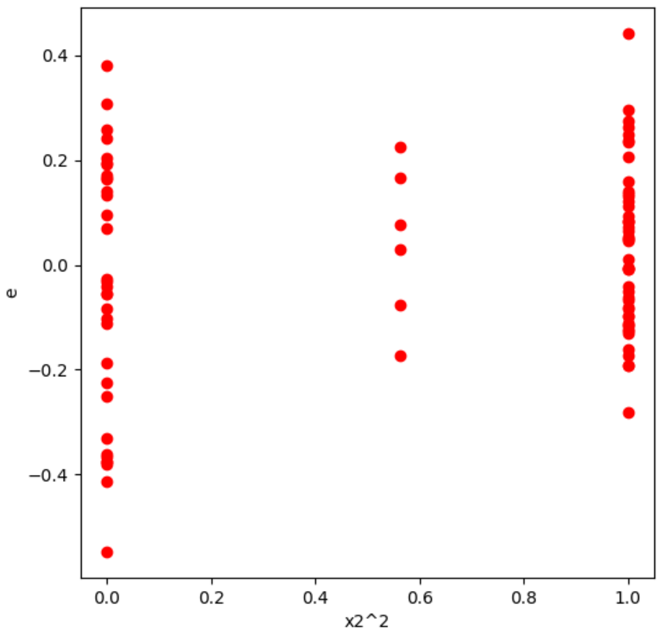












Построение доверительного интервала для максимального математического ожидания отклика

Для поиска максимального математического ожидания отклика воспользуемся полным перебором.

Результат полного перебора:

$$x_{max} = (1, 1, -1, 1)^T;$$

$$\eta_{max}(x, \hat{\theta}) = 10.14687;$$

Доверительный интервал:

$$\eta_{max}(x,\hat{\theta}) - t_{\alpha/2,n-m}\sigma\left(\eta(x,\hat{\theta})\right) \leq \ \eta_{max}(x,\theta) \leq \ \eta_{max}(x,\hat{\theta}) + t_{\alpha/2,n-m}\sigma\left(\eta(x,\hat{\theta})\right),$$

$$\text{где } \sigma\left(\eta(x,\hat{\theta})\right) = \hat{\sigma}\sqrt{f^T(x) * (X^T X)^{-1} f(x)}$$

$$9.68876 \leq \eta_{max}(x,\hat{\theta}) \leq 10.60498$$

4. Код программы

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from scipy.stats import f

from scipy.stats import t


np.random.seed(20)


# Определение функции, которую нужно максимизировать

def objective(x):

    return (theta_hat_star[0] + theta_hat_star[1] * x[0] + theta_hat_star[2] * x[1] + theta_hat_star[3] 
* x[3]

              + theta_hat_star[4] * x[0] * x[2] + theta_hat_star[5] * x[1] * x[2] + theta_hat_star[6] * 
x[0] * x[1] * x[2]

              + theta_hat_star[7] * x[0] * x[1] * x[3] + theta_hat_star[8] * x[0]**2 + theta_hat_star[9] 
* x[1]**2)


# Число наблюдений и число серий при повторных наблюдениях

N=87

N_c = 82

# Чтение данных из файла Excel

df = pd.read_excel('17_var.xlsx')

x1 = df['X1'].to_numpy()

x2 = df['X2'].to_numpy()

x3 = df['X3'].to_numpy()

x4 = df['X4'].to_numpy()

y = df['y'].to_numpy()


# ЗАДАНИЕ №1. Корреляционный анализ

# Создание списка входных данных

inputs = [x1, x2, x3, x4]


# Построение корреляционных полей факторов
```

```
for i in range(len(inputs)):
    plt.figure(figsize=(6, 6))
    plt.scatter(inputs[i], y, color='red')
    plt.xlabel(f'x{i+1}')
    plt.ylabel('y')
    plt.show()

# Построение корреляционных полей для взаимодействия двух факторов
for i in range(len(inputs)):
    for j in range(i+1, len(inputs)):
        interaction = inputs[i] * inputs[j]
        plt.figure(figsize=(6, 6))
        plt.scatter(interaction, y, color='red')
        plt.xlabel(f'x{i+1} * x{j+1}')
        plt.ylabel('y')
        plt.show()

# Построение корреляционных полей для взаимодействия трех факторов
for i in range(len(inputs)):
    for j in range(i+1, len(inputs)):
        for k in range(j+1, len(inputs)):
            interaction = inputs[i] * inputs[j] * inputs[k]
            plt.figure(figsize=(6, 6))
            plt.scatter(interaction, y, color='red')
            plt.xlabel(f'x{i+1} * x{j+1} * x{k+1}')
            plt.ylabel('y')
            plt.show()

# Построение корреляционных полей для квадратов факторов
for i in range(len(inputs)):
    square = inputs[i] ** 2
    plt.figure(figsize=(6, 6))
    plt.scatter(square, y, color='red')
    plt.xlabel(f'(x{i+1})^2')
    plt.ylabel('y')
    plt.show()

# Результат корреляционного анализа
print('Предварительный состав регрессоров по результатам корреляционного анализа: 1, x1, x2, x4, x1*x3, x1*x2*x3, x1*x2*x4')

# МНК-оценки
X = np.column_stack((np.ones(len(x1)), x1, x2, x4, x1*x3, x1*x2*x3, x1*x2*x4))
```

```
theta_hat = np.linalg.inv(X.T @ X) @ X.T @ y


# Задание №2. Мультиколлинеарность

# Вычисление информационной матрицы и её определителя, нормированного на след

info_matrix = np.dot(X.T, X)

det_info_norm_matrix = np.linalg.det(info_matrix / np.trace(info_matrix))

print("\nНормированный определитель информационной матрицы: ", det_info_norm_matrix)


# Вычисление собственных чисел информационной матрицы и сохранение max и min собственного числа

eigvals = np.linalg.eigvals(info_matrix / np.trace(info_matrix))

min_eigval = min(eigvals)

print("Минимальное собственное число информационной матрицы: {:.5e}".format(min_eigval))


# Вычисление меры обусловленности матрицы по Нейману-Голдстейну

neuman_goldstein_measure = max(eigvals)/ min_eigval

print("Мера обусловленности матрицы по Нейману-Голдстейну: {:.5e}".format(neuman_goldstein_measure))


# Вычисление максимальной парной сопряженности

R = np.zeros((X.shape[1] - 1, X.shape[1] - 1))

for i in range(1, X.shape[1]):

    for j in range(i, X.shape[1]):

        R[i - 1, j - 1] = (np.dot(X[:, i], X[:, j]) / (np.linalg.norm(X[:, i]) * np.linalg.norm(X[:,
j]))).round(8)

        R[j - 1, i - 1] = R[i - 1, j - 1]

print("Матрица сопряженности:")

print(R)

np.fill_diagonal(R, 0)

max_pairwise_conjugacy = np.max(np.abs(R))

print("Максимальная парная сопряженность: ", max_pairwise_conjugacy)


# Вычисление максимальной сопряженности

np.fill_diagonal(R, 1)

R_inv = np.linalg.inv(R)

R_2 = [(1 - 1 / R_inv[i, i]).round(8) for i in range(0, R.shape[0])]

max_conjugacy = max(R_2)

print("Вектор R_2: ", R_2)

print("Максимальная сопряженность: ", max_conjugacy)


# Задание №3. Гетероскедастичность

# Тест Бреуша-Пагана

e_t = y - X @ theta_hat

sigma_hat_squared = np.sum(pow(e_t, 2) / N)

c_t = e_t**2 / sigma_hat_squared
```

```
z_t = np.column_stack((np.ones(len(x1)), np.sqrt(x3**2 + x4**2)))
alpha_hat = np.linalg.inv(z_t.T @ z_t) @ z_t.T @ c_t
c_hat = z_t @ alpha_hat
ESS = np.sum((c_hat - np.mean(c_t))**2)

# Вычисление статистики Бреуша-Пагана и критического значения
BP = ESS / 2
F_crit = f.ppf(1 - 0.05, 1, N)

print("\n{0} < {1} => Гипотеза о гомоскедастичности принимается".format(BP.round(5), F_crit.round(5)))

# Тест Голдфельда-Квандтона
order = np.argsort(np.sqrt(x3**2 + x4**2))
X_ordered = X[order]
y_ordered = y[order]

n_c = N // 3
X1 = X_ordered[:(N-n_c)//2]
y1 = y_ordered[:(N-n_c)//2]
X2 = X_ordered[(N+n_c)//2:]
y2 = y_ordered[(N+n_c)//2:]

# Оценка параметров модели для каждой части данных
theta_hat1 = np.linalg.inv(X1.T @ X1) @ X1.T @ y1
theta_hat2 = np.linalg.inv(X2.T @ X2) @ X2.T @ y2

RSS1 = np.sum((y1 - X1 @ theta_hat1)**2)
RSS2 = np.sum((y2 - X2 @ theta_hat2)**2)

# Вычисление статистики теста Голдфельда-Квандтона и критического значения
GQ = RSS2 / RSS1
F_crit = f.ppf(1-0.05, (N - n_c - 2 * len(theta_hat))/2, (N - n_c-2 * len(theta_hat))/2)

print("{0} < {1} => Гипотеза о гомоскедастичности принимается".format(GQ.round(5), F_crit.round(5)))

# Задание №4. Автокорреляция
e_t = y - X @ theta_hat
# Вычисление разности между соседними остатками
diff_e_t = np.diff(e_t)
# Статистика DW
DW = np.sum(diff_e_t**2) / np.sum(e_t**2)
print('\nСтатистика DW:', DW)
```

```python
# для n = 85 крит. знач. d_L = 1.55, d_U = 1.75 при уровне значимости 0.05
print('DW < d_L => Положительная корреляция')


# Задание №5. Оптимальная модель по ОДА
# Состав регрессоров по ОДА
print('\nОптимальный состав регрессоров по ОДА: 1, x1, x2, x4, x1*x3, x2*x3, x1*x2*x3, x1*x2*x4, x1^2, x2^2');
X = np.column_stack((np.ones(len(x1)), x1, x2, x4, x1*x3, x2*x3, x1*x2*x3, x1*x2*x4, x1**2, x2**2))


# Задание №6. Оценки с учетом автокорреляции
rho_hat = 1 - DW/2
P = np.eye(len(y))
for i in range(1, len(y)):
    P[i, i] = 1
    P[i, i-1] = -rho_hat
P[0, 0] = np.sqrt(1 - rho_hat**2)
# Преобразование данных
y_star = P @ y
X_star = P @ X
theta_hat_star = np.linalg.inv(X_star.T @ X_star) @ X_star.T @ y_star
print('Оптимальные оценки параметров(с учетом автокорреляции):', theta_hat_star)


# Задание №7. Повторные наблюдения и адекватность
f_E = N - N_c
f_LF = N_c - len(theta_hat_star)


# Создаем матрицу Q: единичная матрица до строки N_c, после чего идет столбец из 6 единиц (начиная с последнего диагонального элемента)
Q = np.zeros((N, N_c))
Q[:N_c, :N_c] = np.eye(N_c)
Q[N_c-1:, N_c-1] = 1
# SS_LF
y_bar = np.mean(y)
Q_y_bar = Q @ np.full((N_c, ), y_bar)
SS_LF = (Q_y_bar - X @ theta_hat_star).T @ (Q_y_bar - X @ theta_hat_star)
# SS_E
SS_E = (y - Q_y_bar).T @ (y - Q_y_bar)


sigma_E_hat_2 = SS_E / f_E
sigma_LF_hat_2 = SS_LF / f_LF


F = sigma_LF_hat_2 / sigma_E_hat_2
F_crit = f.ppf(1 - 0.05, f_LF, f_E)
```

```
print("\n{0} < {1} => Гипотеза об адекватности модели не отвергается, модель
адекватна".format(F.round(5), F_crit.round(5)))


# Задание №8. Графики остатков

# Построение графика зависимости остатков от номера итерации

plt.figure(figsize=(6, 6))

plt.plot(range(1, len(e_t)+1), e_t, marker='o', linestyle='-', color='red')  # изменение цвета точек

plt.xlabel('i')

plt.ylabel('e')

plt.show()


# Построение графиков зависимости остатков от регрессоров

regressors = ['1', 'x1', 'x2', 'x4', 'x1*x3', 'x2*x3', 'x1*x2*x3', 'x1*x2*x4', 'x1^2', 'x2^2']

for i in range(1, X.shape[1]):

    plt.figure(figsize=(6, 6))

    plt.scatter(X[:, i], e_t, color='red')  # изменение цвета точек

    plt.xlabel(regressors[i])

    plt.ylabel('e')

    plt.show()


e_t = y - X @ theta_hat_star

# "Стьюдентизированные" остатки

d_t = e_t / np.sqrt((e_t.T @ e_t / (N - len(theta_hat_star))) * (1 - np.diag(X @ np.linalg.inv(X.T @ X)
@ X.T)))


# Построение графика зависимости остатков от оценок отклика

plt.scatter(X @ theta_hat_star, d_t, color='red')  # изменение цвета точек

plt.xlabel('Оценки отклика')

plt.ylabel('d')

plt.show()


# Задание №9. Построение доверительного интервала для макс. мат. ожидания отклика

# Создание сетки значений

x1 = np.linspace(-1, 1, 30)

x2 = np.linspace(-1, 1, 30)

x3 = np.linspace(-1, 1, 30)

x4 = np.linspace(-1, 1, 30)


# Поиск максимума (полный перебор)

max_value = -np.inf

max_point = None
```

```
for i in x1:

    for j in x2:

        for k in x3:

            for l in x4:

                value = objective([i, j, k, l])

                if value > max_value:

                    max_value = value

                    max_point = [i, j, k, l]


print(f'\nТочка с максимальным значением математического ожидания отклика: {max_point}')

print('Максимальное математическое ожидание отклика: ', max_value)


# Построение доверительного интервала

t_quantile = t.ppf(1 - 0.05 / 2, df=len(y) - len(theta_hat_star))

sigma_hat_squared_unbiased = e_t.T @ e_t / (len(y) - len(theta_hat_star))

f_x = np.array([1, max_point[0], max_point[1], max_point[3], max_point[0]*max_point[2], max_point[1]*max_point[2],

                max_point[0]*max_point[1]*max_point[2], max_point[0]*max_point[1]*max_point[3], max_point[0]**2, max_point[1]**2])

left_interval_pred = max_value - t_quantile * np.sqrt(sigma_hat_squared_unbiased * (1 + f_x @ np.linalg.inv(X.T @ X) @ f_x.T))

right_interval_pred = max_value + t_quantile * np.sqrt(sigma_hat_squared_unbiased * (1 + f_x @ np.linalg.inv(X.T @ X) @ f_x.T))

print(f'Доверительный интервал для предсказанного значения отклика: ({left_interval_pred}, {right_interval_pred})')
```