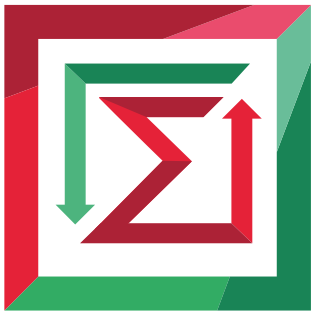
Министерство науки и высшего образования
Российской Федерации

Федеральное государственное бюджетное образовательное
учреждение высшего образования
«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

НГТУ
НЭТИ

Кафедра теоретической и прикладной информатики

Расчётно-графическая работа
по дисциплине «Статистические методы анализа данных»



ФАКУЛЬТЕТ:	ПМИ
ГРУППА:	ПМИ-62
СТУДЕНТ:	Ершов П. К.
ВАРИАНТ:	5
ПРЕПОДАВАТЕЛЬ:	Попов А.А.

Новосибирск
2019

1. **Задача**

   Провести полный цикл исследований по построению регрессионной зависимости по имеющимся экспериментальным данным.

В перечень исследований как обязательные части должны входить:

1. Проверка данных на мультиколлинеарность;
2. Проверка данных на гетероскедастичность (предположительно, что чем дальше от центра эксперимента проведено наблюдение, то возможно дисперсия его больше);
3. Проверка данных на автокорреляцию (упорядоченность наблюдений по своим номерам считать упорядоченностью по времени);
4. 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;
5. Выбор модели оптимальной сложности с использованием критериев Мэллоуса, скорректированного коэффициента детерминации, внешних критериев;
6. Проверка адекватности выбранной модели с использованием повторных наблюдений (последние 6 наблюдений выборки), по которым необходимо будет вычислить оценку дисперсии наблюдений;
7. Построение графиков остатков в различных координатах (по номеру наблюдений, по факторам, по отклику);
8. Определение, опираясь на построенную модель, 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.

Дополнительные комментарии.

Экспериментальные данные представляются в виде таблицы наблюдений типа "вход-выход" в формате xls. Номер варианта задания соответствует порядковому номеру студента в списке группы.

2. Ход работы

Исходные данные:

№	X1	X2	X3	X4	y
1	-1	-1	-1	-1	2,03
2	-1	-0,9	-1	0	0,63
3	-1	-0,8	-1	1	-0,04
4	-1	-0,7	0	-1	0,53
5	-1	-0,6	0	0	-0,22
6	-1	-1	0	1	-1,05
7	-1	-1	1	-1	-0,08
8	-1	-1	1	0	-1,11
9	-1	-1	1	1	-2,02
10	-1	0	-1	-1	0,03
11	-1	0	-1	0	-0,07
12	-1	0	-1	1	0,01
13	-1	0	0	-1	-1,15
14	-1	0	0	0	-1,14
15	-1	0	0	1	-0,93
16	-1	0	1	-1	-1,98
17	-1	0	1	0	-1,99
18	-1	0	1	1	-2,07
19	-1	1	-1	-1	-1,83
20	-1	1	-1	0	-0,91
21	-1	1	-1	1	0,12
22	-1	1	0	-1	-3,04
23	-0,9	1	0	0	-1,83
24	-0,8	1	0	1	-0,31
25	-0,7	1	1	-1	-3,21
26	-0,6	1	1	0	-2,14
27	-1	1	1	1	-1,69
28	0	-1	-1	-1	4,03
29	0	-0,9	-1	0	2,86
30	0	-0,8	-1	1	2,33
31	0	-0,7	0	-1	3,00
32	0	-0,6	0	0	2,38
33	0	-1	0	1	1,89
34	0	-1	1	-1	2,55
35	0	-1	1	0	1,24
36	0	-1	1	1	0,22
37	0	0	-1	-1	2,27
38	0	0	-1	0	2,36
39	0	0	-1	1	2,29
40	0	0	0	-1	1,11
41	0	0	0	0	1,07

42	0	0	0	1	1,05
43	0	0	1	-1	0,26
44	0	0	1	0	0,08
45	0	0	1	1	0,25
46	0	1	-1	-1	0,29
47	0	1	-1	0	1,21
48	0	1	-1	1	2,11
49	0	1	0	-1	-1,18
50	0	1	0	0	-0,14
51	0	1	0	1	0,98
52	0	1	1	-1	-2,16
53	0	1	1	0	-1,06
54	0	1	1	1	0,01
55	1	-1	-1	-1	5,77
56	0,9	-1	-1	0	4,48
57	0,8	-1	-1	1	3,55
58	0,7	-1	0	-1	4,19
59	0,6	-1	0	0	3,20
60	1	-1	0	1	3,03
61	1	-1	1	-1	4,16
62	1	-1	1	0	3,52
63	1	-1	1	1	2,67
64	1	0	-1	-1	4,52
65	1	0	-1	0	4,80
66	1	0	-1	1	4,56
67	1	0	0	-1	3,21
68	1	0	0	0	3,35
69	1	0	0	1	3,62
70	1	0	1	-1	2,55
71	1	0	1	0	2,35
72	1	0	1	1	2,05
73	1	1	-1	-1	1,99
74	1	1	-1	0	3,16
75	1	1	-1	1	4,17
76	1	1	0	-1	1,28
77	1	1	0	0	2,12
78	0,9	1	0	1	2,99
79	0,8	1	1	-1	-0,29
80	0,7	1	1	0	0,42
81	0,6	1	1	1	1,10
82	0	0	0	0	0,94
83	0	0	0	0	0,95
84	0	0	0	0	1,03
85	0	0	0	0	0,94
86	0	0	0	0	0,77
87	0	0	0	0	0,66

1. Проверка на мультиколлинеарность
   1.1. Определитель информационной матрицы X^TX

Определитель информационной матрицы

$detInf := 7.21384600000000\ 10^6$

1.2. Минимально собственной число

Минимальное собственное число

$\lambda min := 48.8804669684940762$

1.3. Мера обусловленности матрицы по X^TX по Нейману-Голдстейну

Мера обусловленности матрицы по Нейману – Голдстейну

1.107064622

1.4. Максимальная парная сопряжённость

Построим матрицу $R = \begin{pmatrix} 1 & r_{1,2} & \dots & r_{1,m} \\ r_{2,1} & 1 & \dots & r_{2,m} \\ \dots & \dots & \dots & \dots \\ r_{m,1} & r_{m,2} & \dots & 1 \end{pmatrix}$, где $r_{i,j} = cov(\underline{x_i}, \underline{x_j})$

Тогда показателем мультиколлинеарности может выступать $\max_{i,j}|r_{i,j}|$ при условии $i \neq j$

Максимальная парная сопряженность

Rmax1 = 0.01147834114

1.5. Максимальная сопряжённость

Показателем мультиколлинеарности может выступать $\max_i|R_i|$, где R_i можно получить из формулы $R_i^2 = 1 - \frac{1}{R_{ii}^{-1}}$, где R_{ii}^{-1} – это элемент i, i (диагональный) матрицы обратной к сопряжённой R.

Максимальная сопряженность

$Rmax2 := 0.0120992381037185$

Вывод по мультиколлинеарности: исходя из результатов тестов отсутствует.

2. Проверка на гетероскетастичность

2.1. Тест Бреуша-Пагана

Оценивание исходного уравнения по МНК, с получением остатков и оценивание дисперсии.

По полученным экспериментально данным, находим по методу наименьших квадратов точечную оценку параметров:$\hat{\theta} = (X^T X)^{-1} X^T y$

Вектор остатков: $e_t = y_t - f(x_t)\hat{\theta}$

Получим дисперсию: $\tilde{\sigma} = \frac{\sum e_t^2}{n}$

Построим регрессию $c_t = \frac{e_t^2}{\tilde{\sigma}^2}$ по z_t и вычислим ESS.

$$Z = \begin{pmatrix} z_1(u_1) & z_2(u_2) \\ \dots & \dots \\ z_1(u_N) & z_1(u_N) \end{pmatrix}, \alpha = (Z^T Z)^{-1} Z^T d, \text{где } d = (\frac{e_1^2}{\hat{\sigma}^2}, \dots, \frac{e_N^2}{\hat{\sigma}^2})^T)$$

$$ESS = \sum_{i=1}^{N} (c_i - \hat{c})^2, \text{где } c_i = z_i^T \alpha$$

Гипотеза о гомоскедастичности принимается, если $ESS < \chi^2_{0.05,1}$

Гетероскедастичность, тест Бреуша – Пагана

1.27953645627685

3.84145606580278

false

1.27953645627685 < 3.84145606580278 следовательно, гипотеза принимается.

2.2. Тест Глодфельда-Квандтона

Предположим, что источник нарушения гомоскедастичности взят в форме $E(\epsilon_i^2) = \rho(\bar{x})$.

Упорядочим последовательность наблюдений в соответствии с величиной отклика.

Опустим n_c=n/3=87/3=29 наблюдений в середине выборки.

Оценим RSS для первых $\frac{(n-n_c)}{2}$ и последних $\frac{(n-n_c)}{2}$ наблюдений. Гипотеза о гомоскедастичности будет принята, если $\frac{RSS_2}{RSS_1} < F_{\alpha,\frac{(n-n_c-2m)}{2},\frac{(n-n_c-2m)}{2}} = F_{0.05,25,25} \approx 1.955$

$\frac{RSS_2}{RSS_1} = \frac{0.002284}{5.6227} \approx 0{,}0004062 < 1.955$ значит, гипотеза не отвергается.

Вывод по гетероскедастичности: исходя из результатов двух тестов, можно считать, что гетероскедастичности нет.

3. Проверка данных на автокорреляцию

Тест Дарбина-Уотсона

Статистика $DW = \frac{\sum_{i=2}^{n}(\hat{e}_{i-1}-\hat{e}_i)^2}{\sum_{i=1}^{n}\hat{e}_i^{\,2}} = 2(1-\hat{\rho})$, гипотеза $H_0: \rho = 0$

DW := 0.60198

$up := 88.8653359800023$

$down := 147.620124860977$

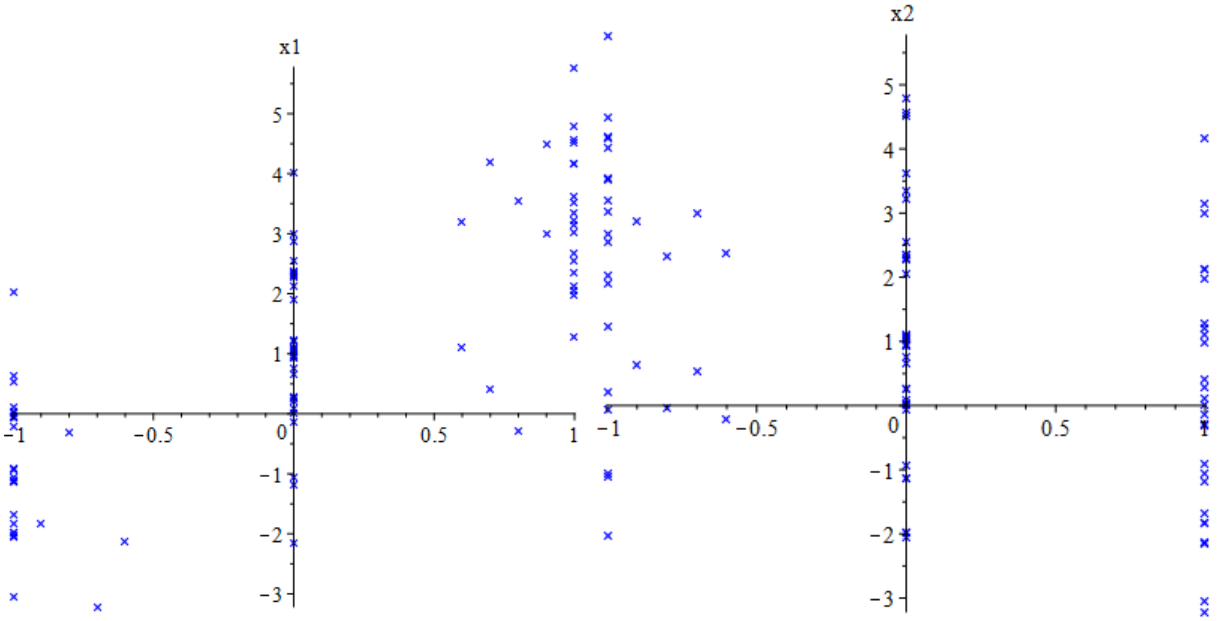
$DW := 0.601986592706735$

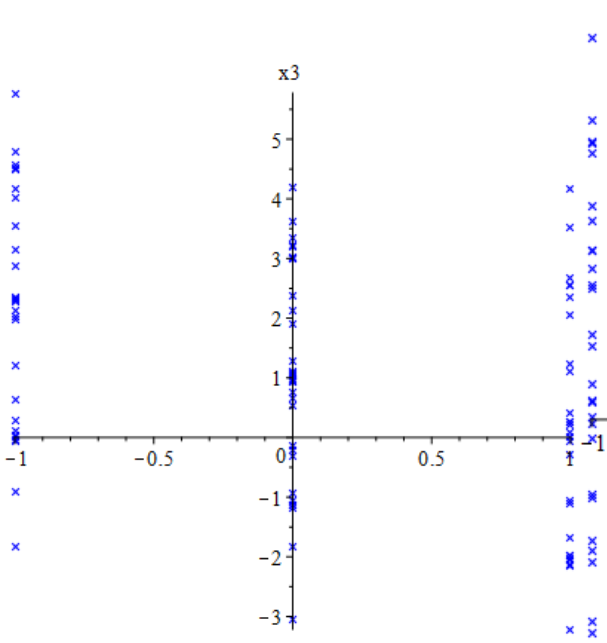
Гипотеза H_0 отвергается, так как статистика близка к нулю.

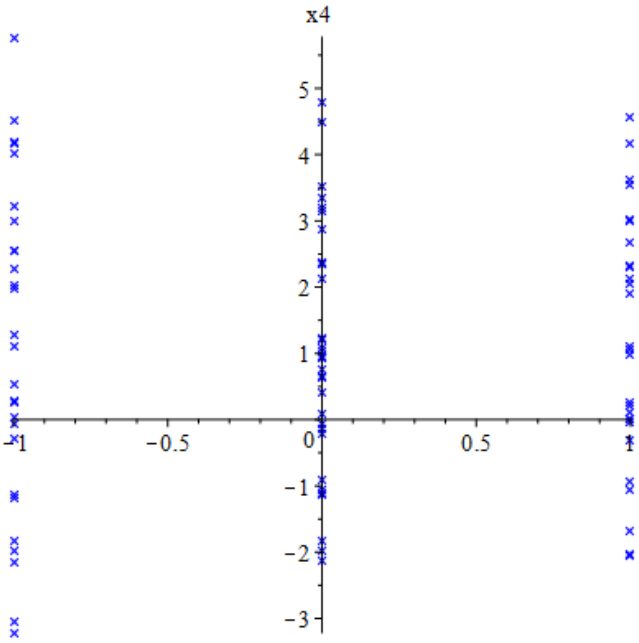
Выводы по автокорреляции: так как статистика близка к нулю, присутствует положительная корреляция.

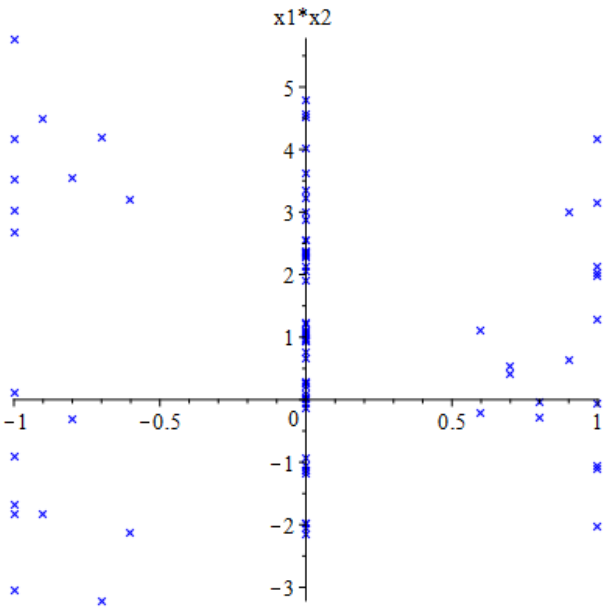
4. Выбор предварительного состава регрессоров с использованием корреляционных полей. В качестве регрессоров-кандидатов предположительно могут выступать: свободный член, сами факторы, их взаимодействия (двух-трех факторов), квадраты факторов

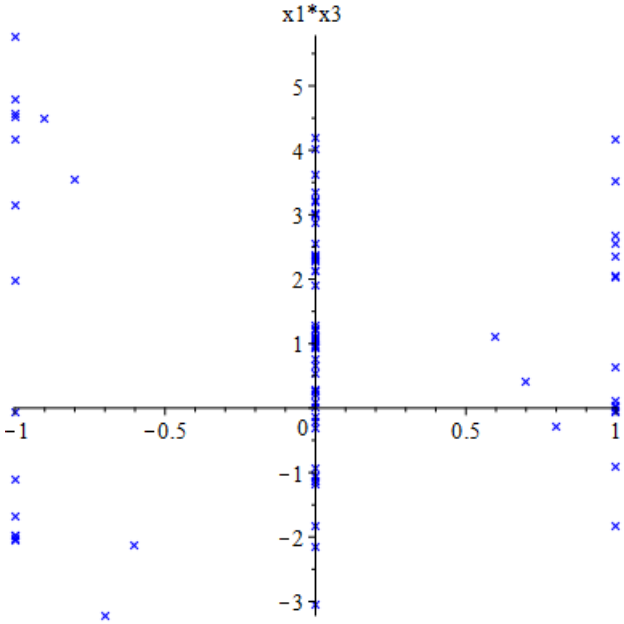
Корреляционные поля

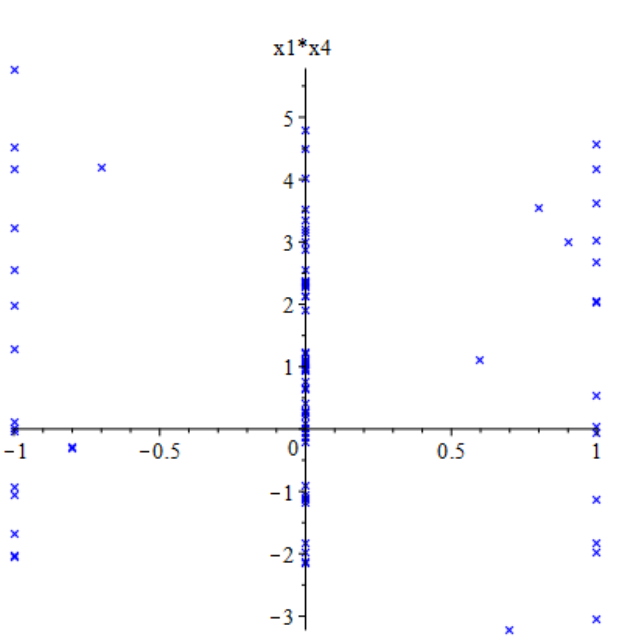


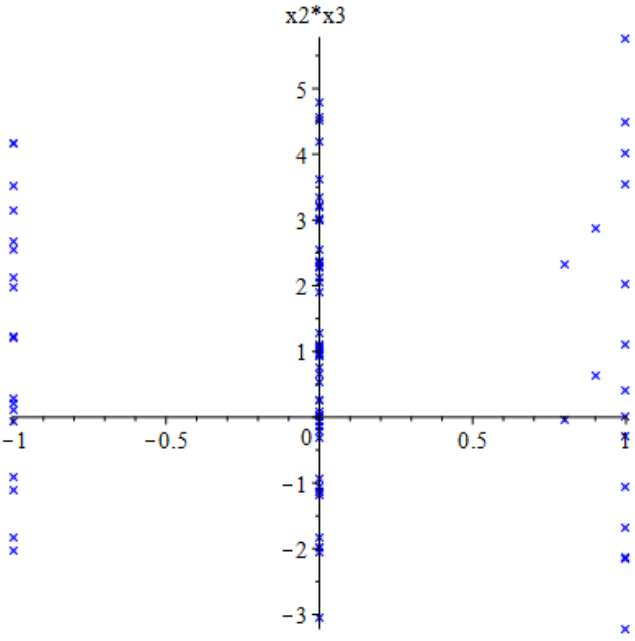

x3
5
4
3
2
1
−1
−2
−3
−1
−0.5
0
0.5
1

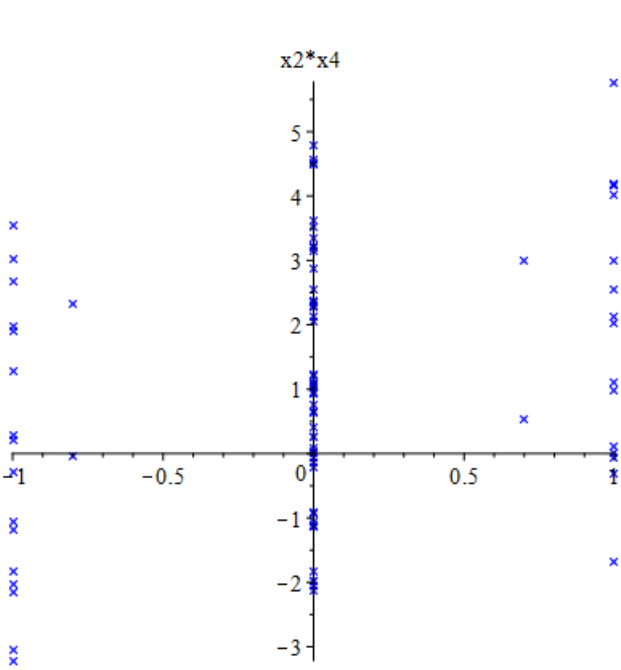

x4
5
4
3
2
1
−1
−2
−3
−1
−0.5
0
0.5
1

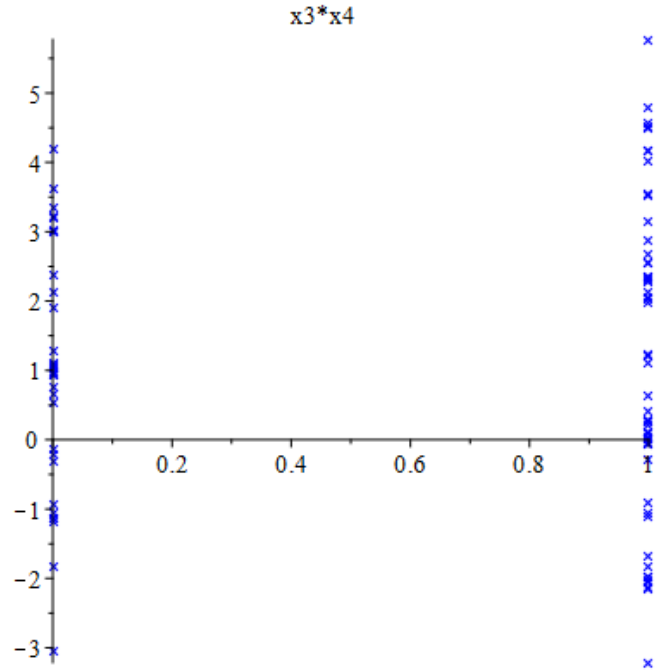

x1*x2
5
4
3
2
1
−1
−2
−3
−1
−0.5
0
0.5
1

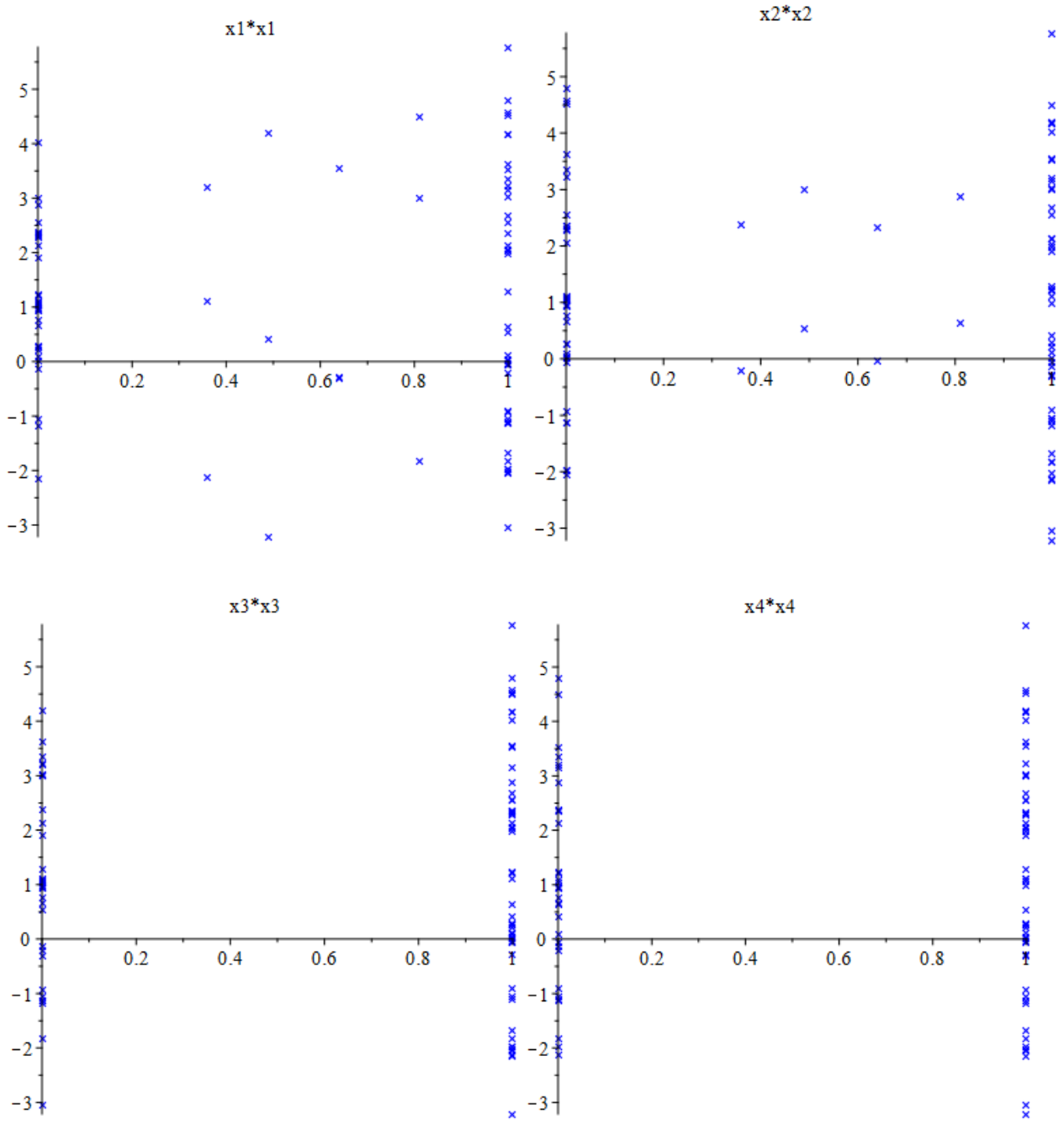

x1*x3
5
4
3
2
1
−1
−2
−3
−1
−0.5
0
0.5
1


x1*x4


x2*x3


x2*x4


x3*x4



Исходя из данный в корреляционных полях, можно предположить, что предварительная модель имеет вид: $f(x) = (1, x_1, x_2, x_3, x_1 * x_2, x_2 * x_4, x_3 * x_4)^T$

5. Выбор модели оптимальной сложности

Пусть модель среднего $f_1(x) = (1)^T$ это базовая модель.
В качестве полной модели возьмём $f_{\text{полная}}(x) = (1, x_1, x_2, x_3, x_4, x_1 * x_2, x_1 * x_3, x_1 * x_4, x_2 * x_3, x_2 * x_4, x_3 * x_4, x_1 * x_1, x_2 * x_2, x_3 * x_3, x_4 * x_4)^T$

Воспользуемся критериями:

Статистика Мэлоуса: $C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n \to min$

Множественный критерий корреляции: $R_p^2 = \frac{\sum(\hat{y}_{ip}-\bar{\hat{y}}_p)^2}{\sum(y_i-\bar{y})^2} \to 1$

MSEP-критерий: $E_p = \frac{RSS_p}{n(n-p)}(1 + n + \frac{p(n+1)}{n-p-2}) \to min$

AEV-критерий: $AEV_p = \frac{p \cdot RSS_p}{n(n-p)} \to min$, где $RSS_p = (y - \hat{y}_p)^T(y - \hat{y}_p)$, $\hat{\sigma}^2 = \frac{RSS}{n-m}$

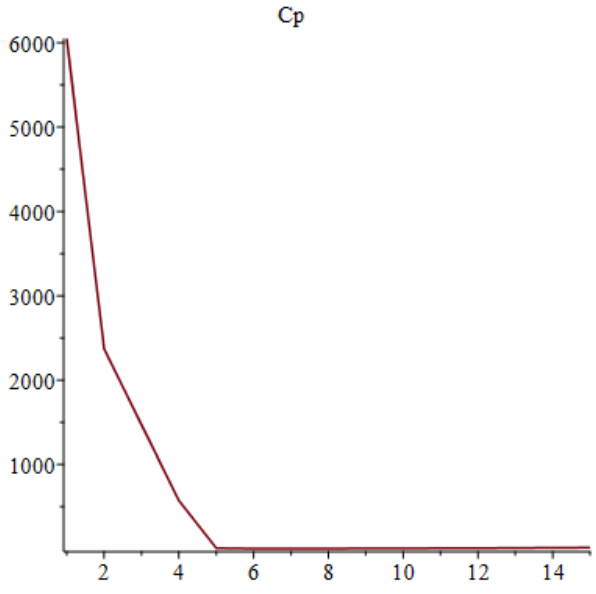
(m - число регрессоров полной модели)

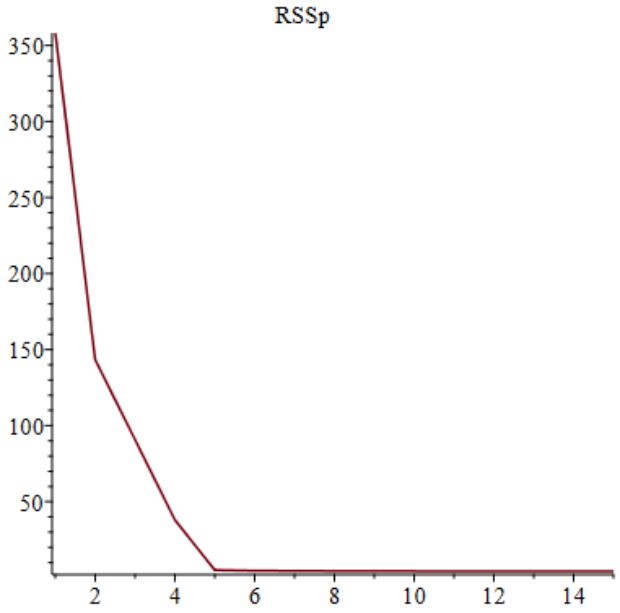
Для подсчета F- критериев воспользуемся формулой:

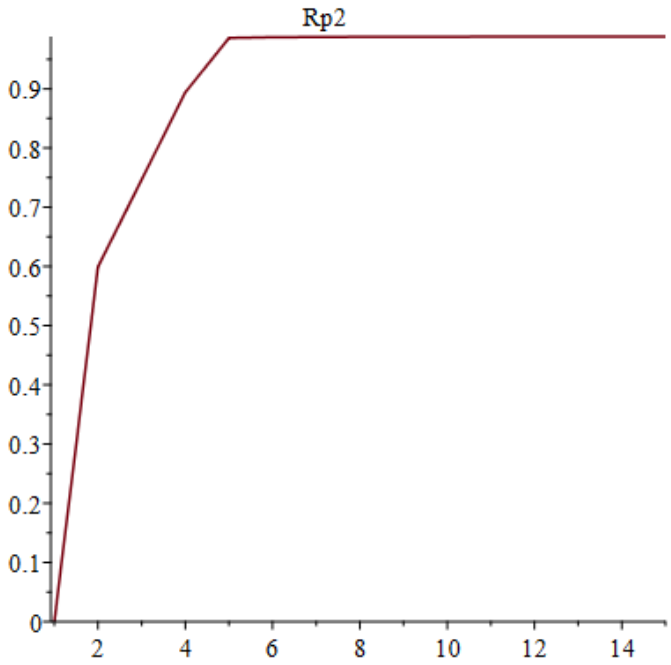
$F_{ij} = \frac{v_2}{v_1} \cdot \frac{RSS_{ij}-RSS_j}{RSS_{i,j}}$ (для алгоритма включения $v_1 = 1, v_2 = n - m$)

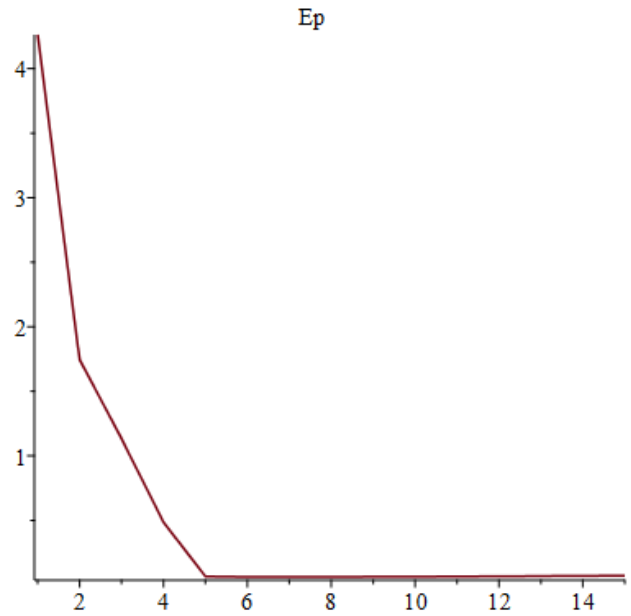
p	RSS_p	C_p	R_p^2	E_p	AEV_p	Модель
1	358.136	6048.89	≈0	4.2624	0.0478	$f(x) = (1)^T$
2	143.328	2371.83	0.5998	1.746	0.0387	$f(x) = (1, x_1)^T$
3	90.725	1472.88	0.746	1.132	0.0372	$f(x) = (1, x_1, x_3)^T$
4	37.977	571.455	0.894	0.485	0.021	$f(x) = (1, x_1, x_3, x_2)^T$
5	4.9958	8.5658	0.986	0.0655	0.0035	$f(x) = (1, x_1, x_3, x_2, x_2 * x_4)^T$
6	4.6338	4.3657	0.987	0.06226	0.0039	$f(x) = (1, x_1, x_3, x_2, x_2 * x_4, x_2 * x_3)^T$
7	4.506	4.172	0.987	0.06208	0.0045	$f(x) = (1, x_1, x_3, x_2, x_2 * x_4, x_2 * x_3, x_4 * x_4)^T$

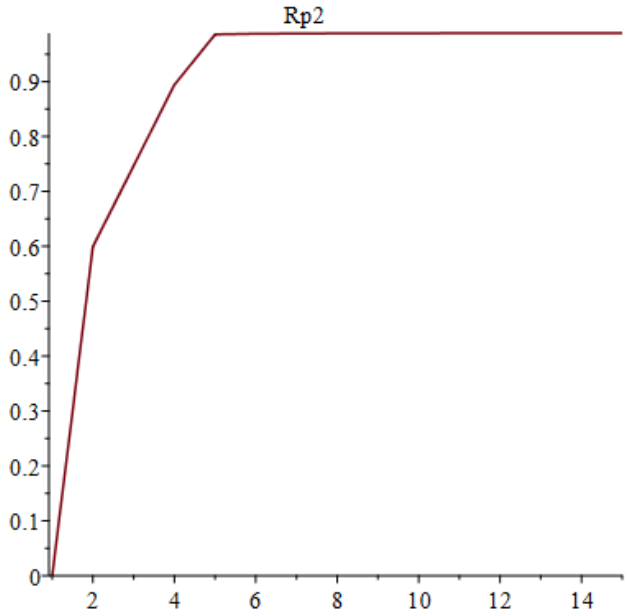
Дальнейшее исследование не имеет смысла из-за слабого изменения параметров.

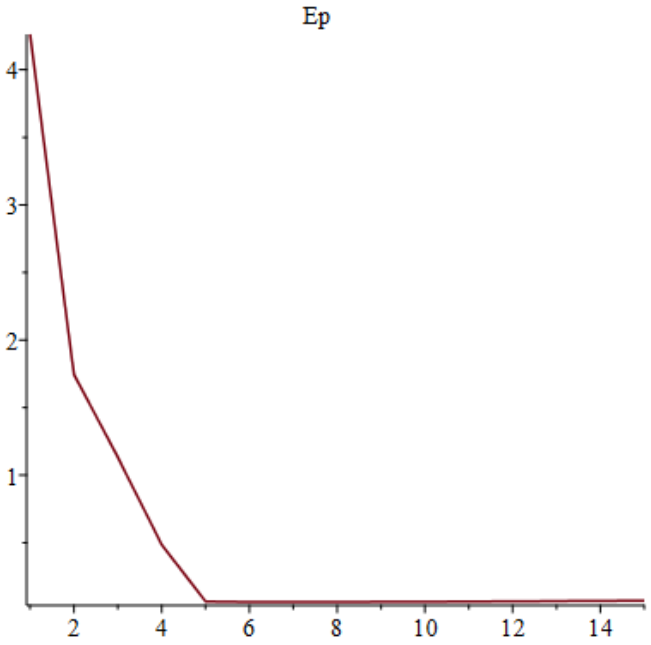


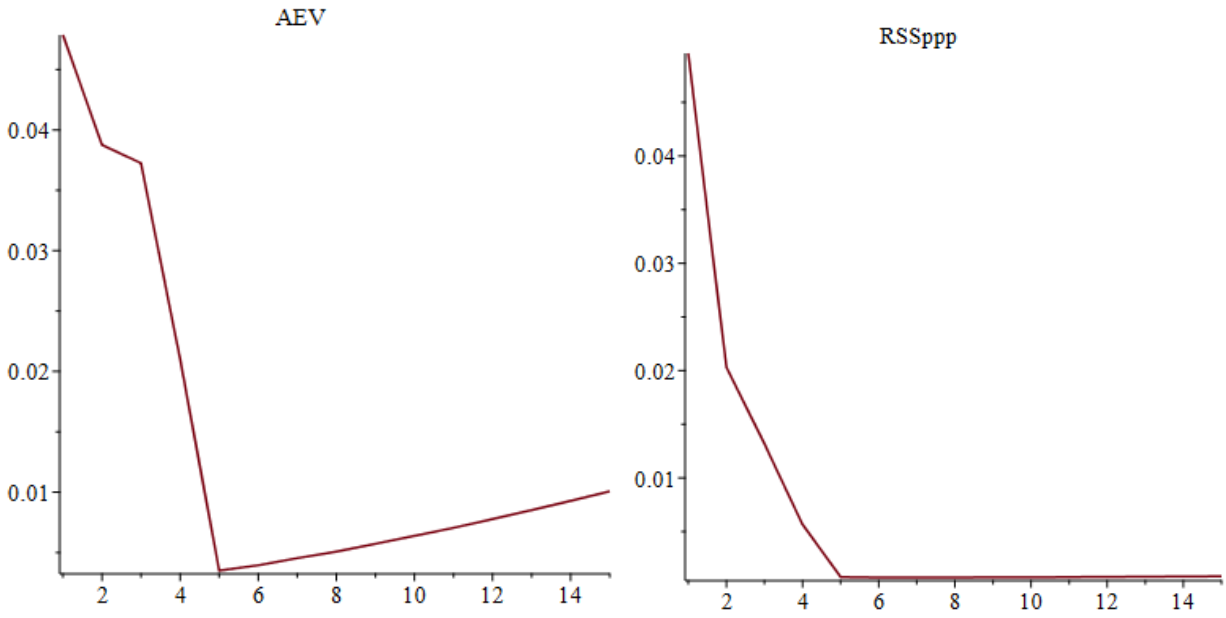



Rp2
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2
4
6
8
10
12
14


Ep
4
3
2
1
2
4
6
8
10
12
14


Rp2
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2
4
6
8
10
12
14


Ep
4
3
2
1
2
4
6
8
10
12
14



Исходя из результатов критериев, следует выбрать пятую модель: $f(x) = (1, x_1, x_3, x_2, x_2 * x_4)^T$
Предварительная модель ($f(x) = (1, x_1, x_2, x_3, x_1 * x_2, x_2 * x_4, x_3 * x_4)^T$) немного хуже.

6. Проверка адекватности полученной модели

Разобьем выборку на 2 части: 81 наблюдение и 6 наблюдений (последних).

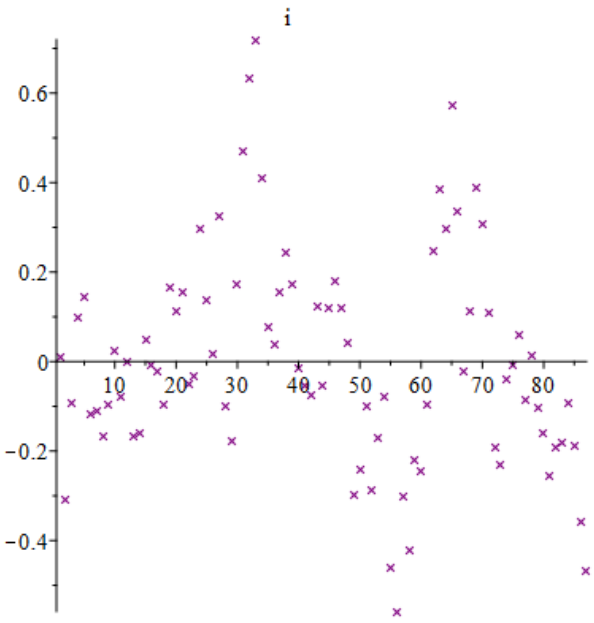
$\hat{\sigma}_E^2 = \frac{\hat{e}^T\hat{e}}{n-m} = \frac{\hat{e}^T\hat{e}}{81-3} = 0.06573526122,$

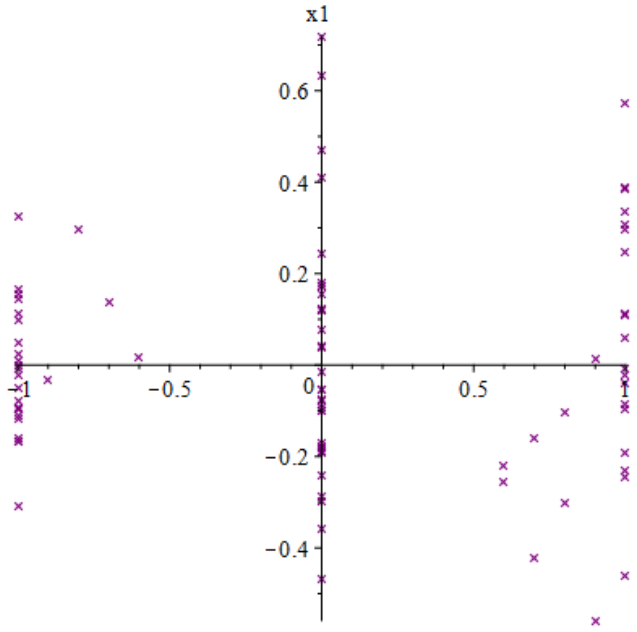
$\hat{\sigma}_{LF}^2 = \frac{\hat{e}^T\hat{e}}{n-m} = \frac{\hat{e}^T\hat{e}}{6-5} = 4.99587985301679183,$

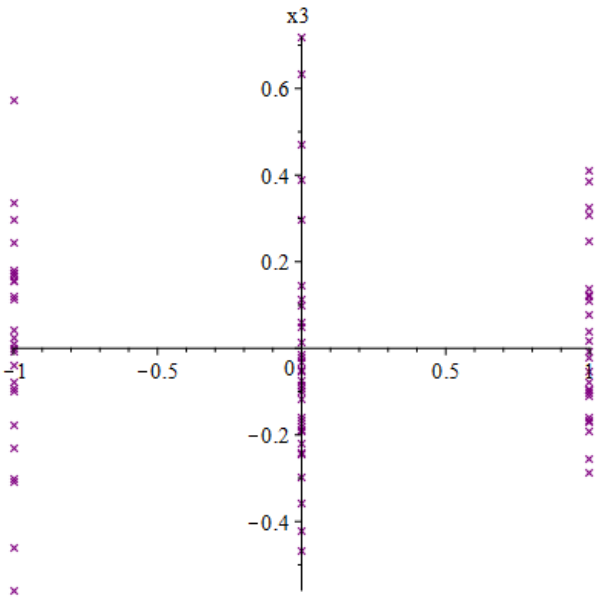
$F_{\alpha,f_{LF},f_E} = F_{0.05,76,1} = 252.640291165918$

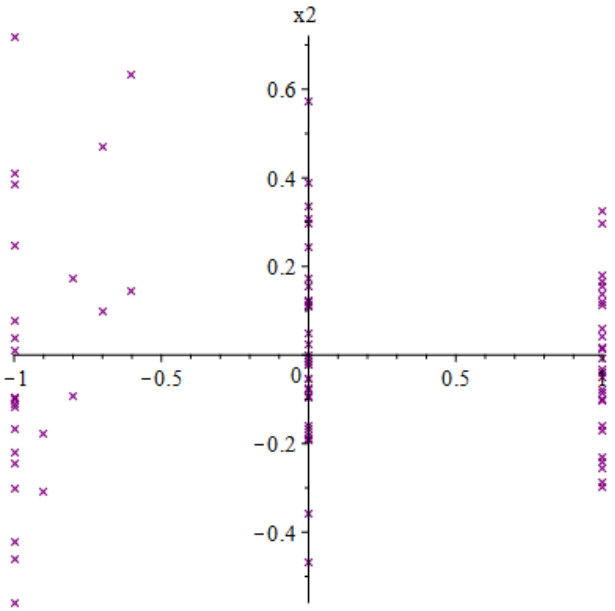
$F = \frac{\hat{\sigma}_{LF}^2}{\hat{\sigma}_E^2} = \frac{4.99587985301679183}{0.06573526122} \approx 76 < 252.640291165918$- значит модель адекватная

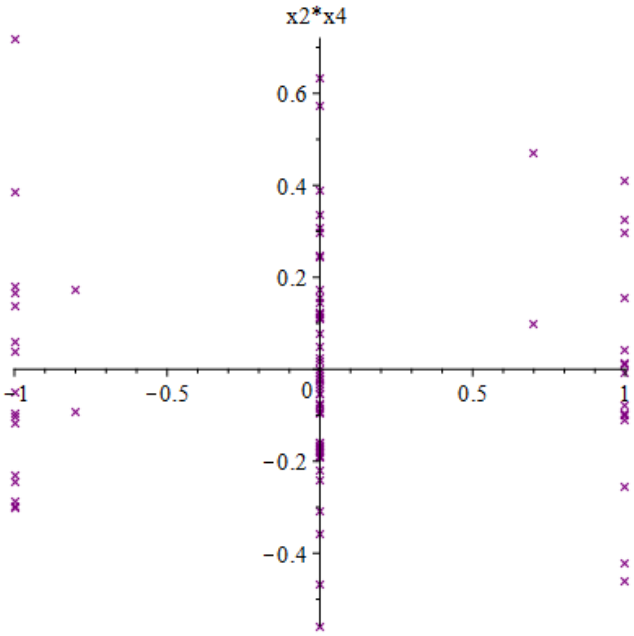
7. Построение графиков остатков











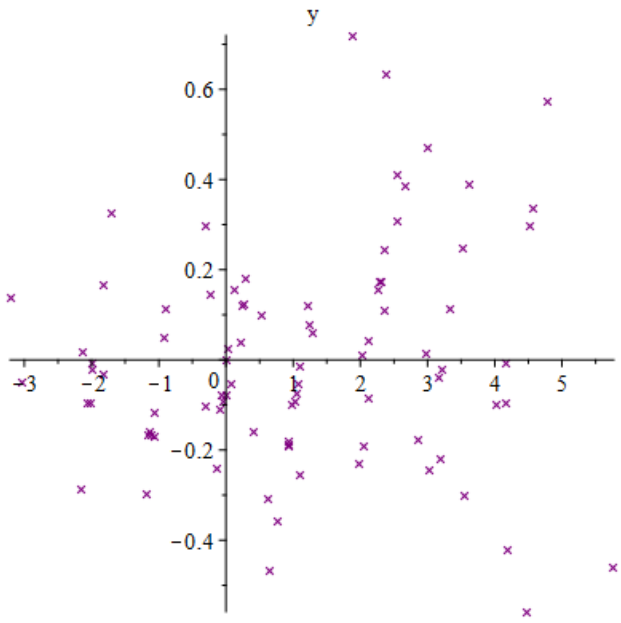


График остатков показывают, что наличие некоторой гетероскедастичности возможно. Тот факт, что графики не криволинейны говорит об адекватности модели.

8. Определение точки в факторном пространстве, имеющей максимальное математическое ожидание отклика и построение доверительного интервала

Для определения точки будем максимизировать следующий функционал:

$$\hat{y}_{max} = \max_{x} \eta(x, \hat{\theta})$$

Полученный результат: $x_{max} = (1, -1, -1, -1)^T$, $\hat{y}_{max} \approx 5.771$

Доверительный интервал: $\eta(x, \hat{\theta}) - t_{\alpha/2, f_R}\sigma(\eta(x, \hat{\theta})) \leq \eta(x, \theta) \leq \eta(x, \hat{\theta}) + t_{\alpha/2, f_R}\sigma(\eta(x, \hat{\theta}))$,
где $\sigma^2(\hat{y}(x, \hat{\theta})) = \hat{\sigma}^2(1 + f^T(x)(X^T X)^{-1} f(x)) \approx 1.916735686$, $t_{\alpha/2, f_R} = t_{0.05/2,84} \approx 1.988959$

Таким образом: $y_{max} \in [1.95897729531954;\ 9.58359767268046]$