# Министерство науки и высшего образования

# Российской Федерации

## Федеральное государственное бюджетное образовательное учреждение высшего образования

## «Новосибирский государственный технический университет»



## 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.

1. **Ход работы**

Исходные данные:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | Х1 | Х2 | Х3 | Х4 | 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. Определитель информационной матрицы XTX



* 1. Минимально собственной число



* 1. Мера обусловленности матрицы по XTX по Нейману-Голдстейну



* 1. Максимальная парная сопряжённость

Построим матрицу

Тогда показателем мультиколлинеарности может выступать



* 1. Максимальная сопряжённость

Показателем мультиколлинеарности может выступать можно получить из формулы матрицы обратной к сопряжённой R.



**Вывод по мультиколлинеарности**: исходя из результатов тестов отсутствует.

1. Проверка на гетероскетастичность
   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:

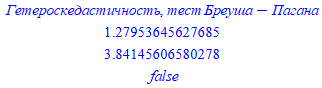
Вектор остатков:

Получим дисперсию:

Построим регрессию по и вычислим ESS.

)

Гипотеза о гомоскедастичности принимается, если



1.27953645627685 < 3.84145606580278 следовательно, гипотеза принимается.

* 1. Тест Глодфельда-Квандтона

Предположим, что источник нарушения гомоскедастичности взят в форме .

Упорядочим последовательность наблюдений в соответствии с величиной отклика.

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

Оценим RSS для первых и последних наблюдений. Гипотеза о гомоскедастичности будет принята, если 1.955

значит, гипотеза не отвергается.

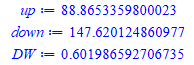
**Вывод по гетероскедастичности:** исходя из результатов двух тестов, можно считать, что гетероскедастичности нет.

1. Проверка данных на автокорреляцию

Тест Дарбина-Уотсона

Статистика , гипотеза

DW := 0.60198

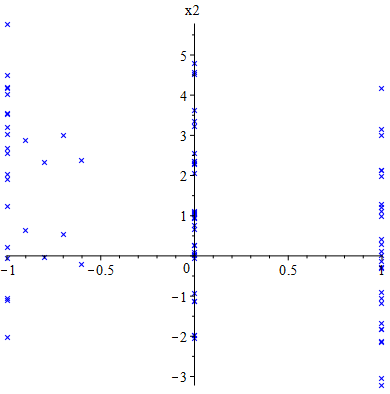
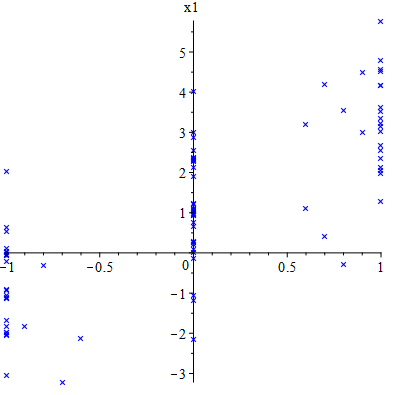


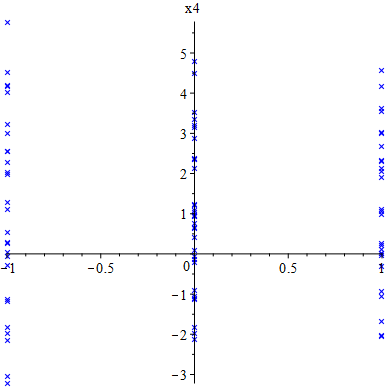
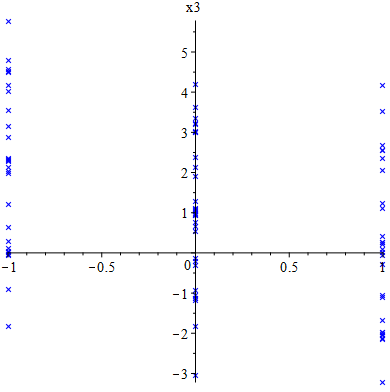
Гипотеза отвергается, так как статистика близка к нулю.

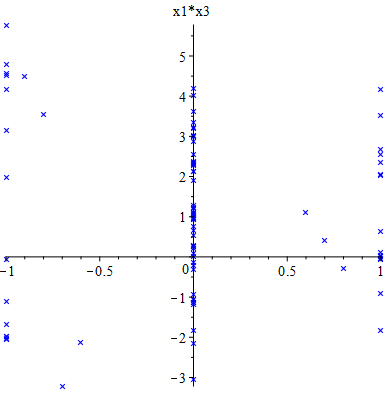
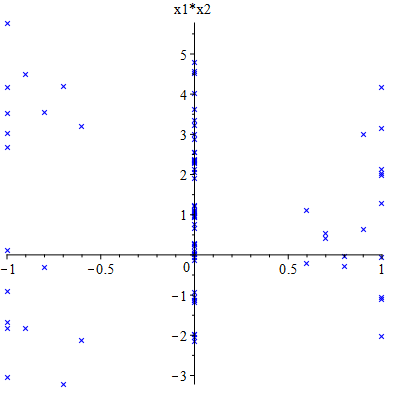
**Выводы по автокорреляции:** так как статистика близка к нулю, присутствует положительная корреляция.

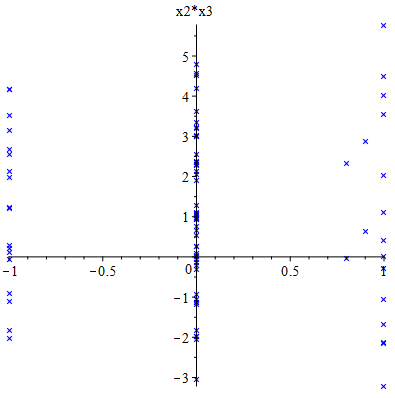
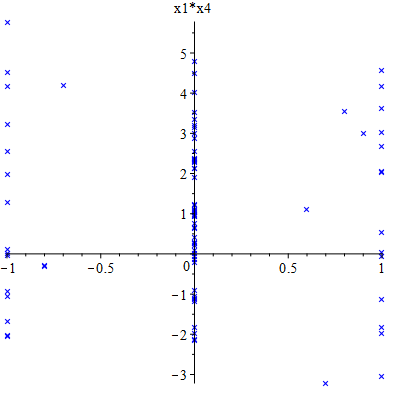
1. Выбор предварительного состава регрессоров с использованием корреляционных полей. В качестве регрессоров-кандидатов предположительно могут выступать: свободный член, сами факторы, их взаимодействия (двух-трех факторов), квадраты факторов

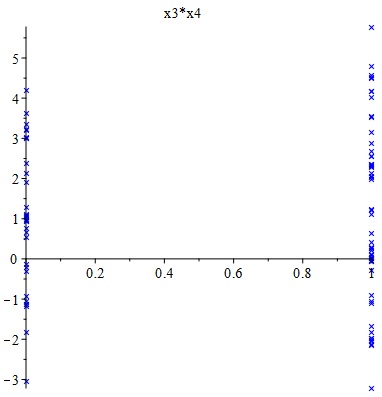
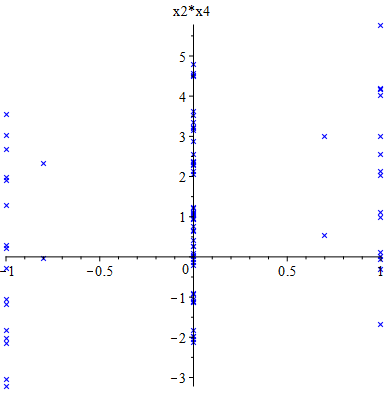
Корреляционные поля

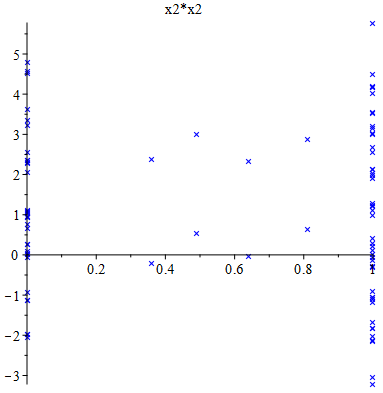
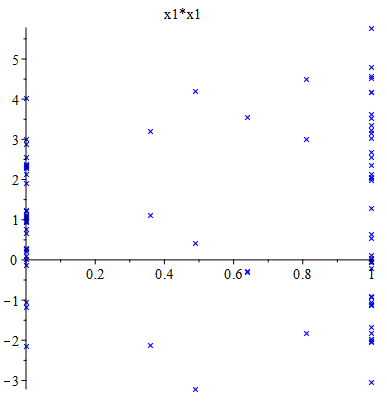


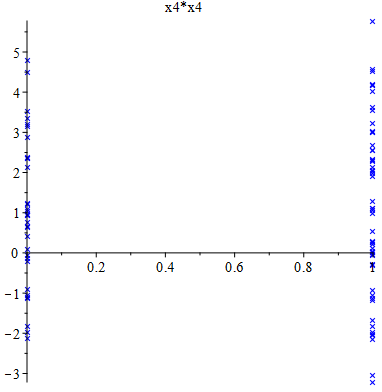
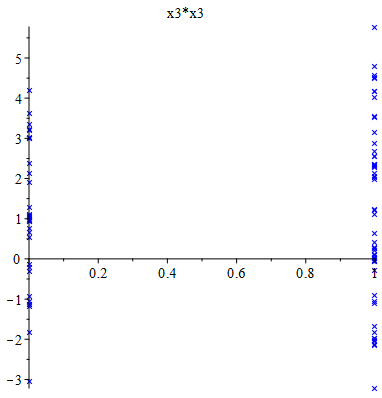












Исходя из данный в корреляционных полях, можно предположить, что предварительная модель имеет вид:

1. Выбор модели оптимальной сложности

Пусть модель среднего это базовая модель.

В качестве полной модели возьмём

Воспользуемся критериями:

Статистика Мэлоуса:

Множественный критерий корреляции:

MSEP-критерий:

AEV-критерий: , где ,

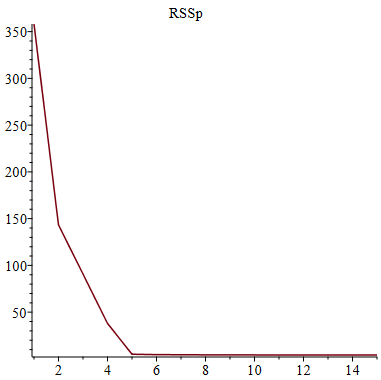
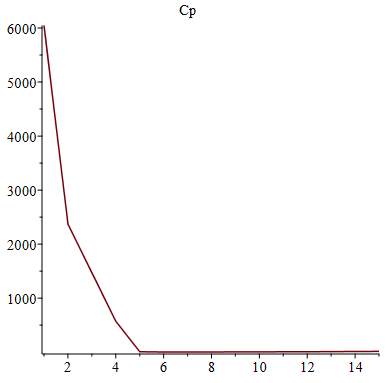
( - число регрессоров полной модели)

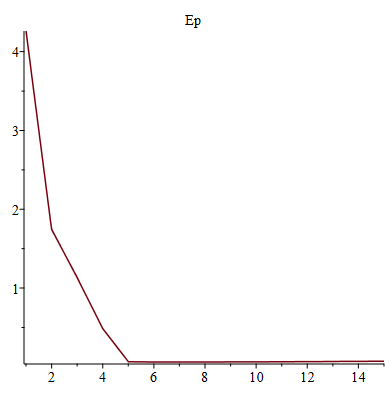
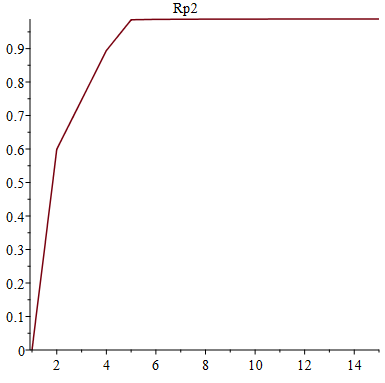
Для подсчета - критериев воспользуемся формулой:

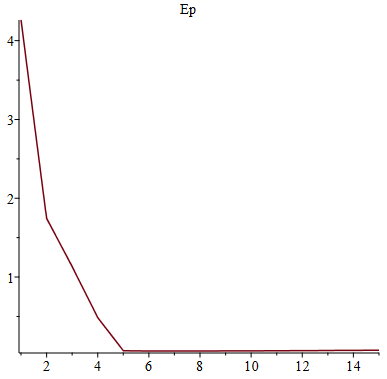
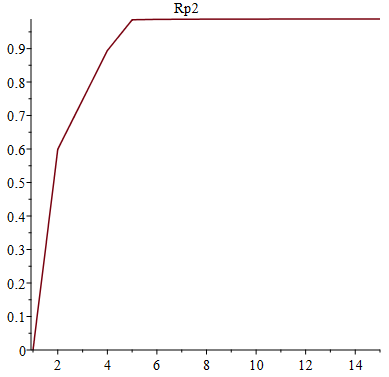
(для алгоритма включения , )

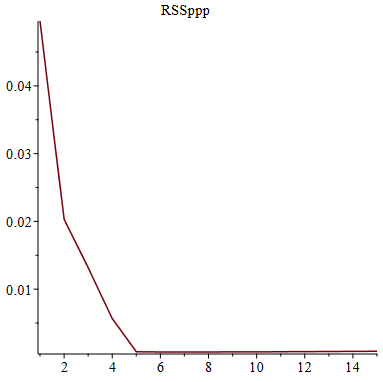
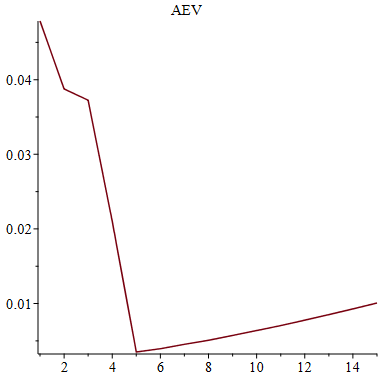
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | Модель |
| 1 | 358.136 | 6048.89 | ≈0 | 4.2624 | 0.0478 |  |
| 2 | 143.328 | 2371.83 | 0.5998 | 1.746 | 0.0387 |  |
| 3 | 90.725 | 1472.88 | 0.746 | 1.132 | 0.0372 |  |
| 4 | 37.977 | 571.455 | 0.894 | 0.485 | 0.021 |  |
| 5 | 4.9958 | 8.5658 | 0.986 | 0.0655 | 0.0035 |  |
| 6 | 4.6338 | 4.3657 | 0.987 | 0.06226 | 0.0039 |  |
| 7 | 4.506 | 4.172 | 0.987 | 0.06208 | 0.0045 |  |

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









Исходя из результатов критериев, следует выбрать пятую модель:

Предварительная модель () немного хуже.

1. Проверка адекватности полученной модели

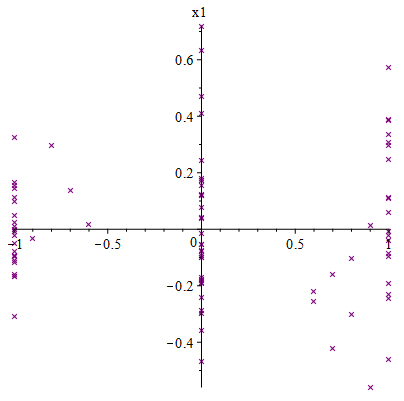
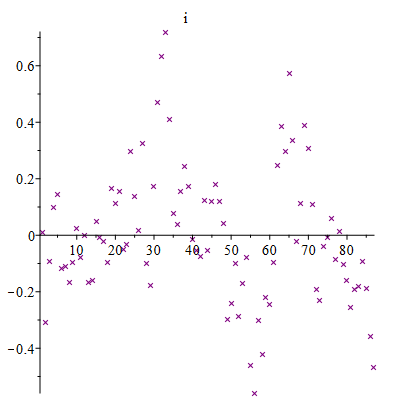
Разобьем выборку на 2 части: 81 наблюдение и 6 наблюдений (последних).

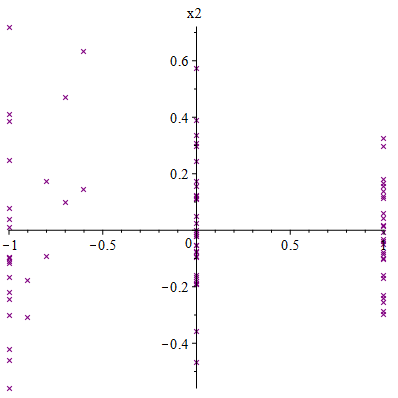
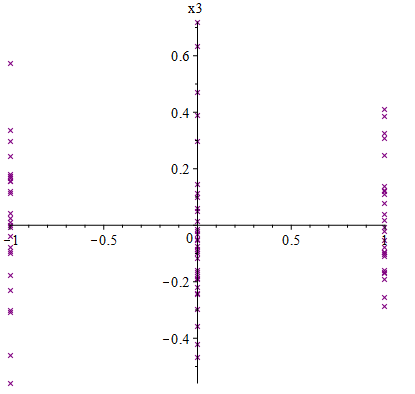
,

,

- значит модель адекватная

1. Построение графиков остатков





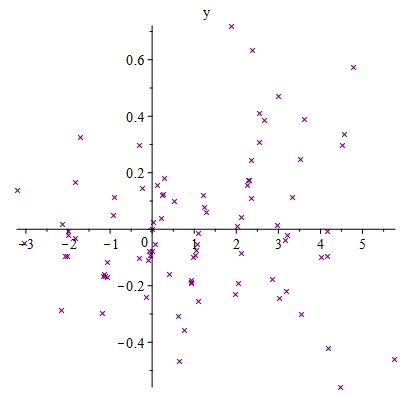
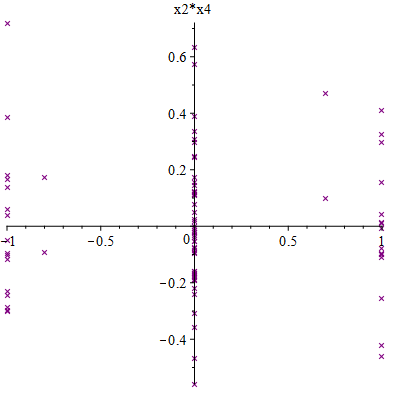


График остатков показывают, что наличие некоторой гетероскедастичности возможно. Тот факт, что графики не криволинейны говорит об адекватности модели.

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

Для определения точки будем максимизировать следующий функционал:

Полученный результат: ,

Доверительный интервал: , где,

Таким образом: