**Фролов Максим 21712**

**Задача 1**

**1а) Укажите переменные с левой (правой) асимметрией. Обоснуйте.**

Выборочный коэффициент ассиметрии был рассчитан как среднее куба стандартизованной величины ((величина - минус среднее) /стандартное отклонение), выборочный коэф. куртозиса был рассчитан как среднее 4 степени стандартизованной величины минус 3 (так как коэф. куртозиса = 3 у нормального распределения). Подробные формулы в R показаны ниже:

mean(scale(x)^3), mean(scale(x)^4-3)

Таблица для 5 переменных:

|  |  |  |
| --- | --- | --- |
|  | asy | kurt |
| x1 | 0,543 | 0,217 |
| x2 | 1,994 | 5,119 |
| x3 | 2,166 | 32,224 |
| x4 | 0,049 | -1,199 |
| x5 | 0,013 | 0,029 |

Правая ассиметрия (коэф. >0): x1, x2, x3, x4, x5

Переменных с левой симметрией не наблюдается.

**1б) Укажите переменные с маленьким (большим) куртозисом. Обоснуйте.**

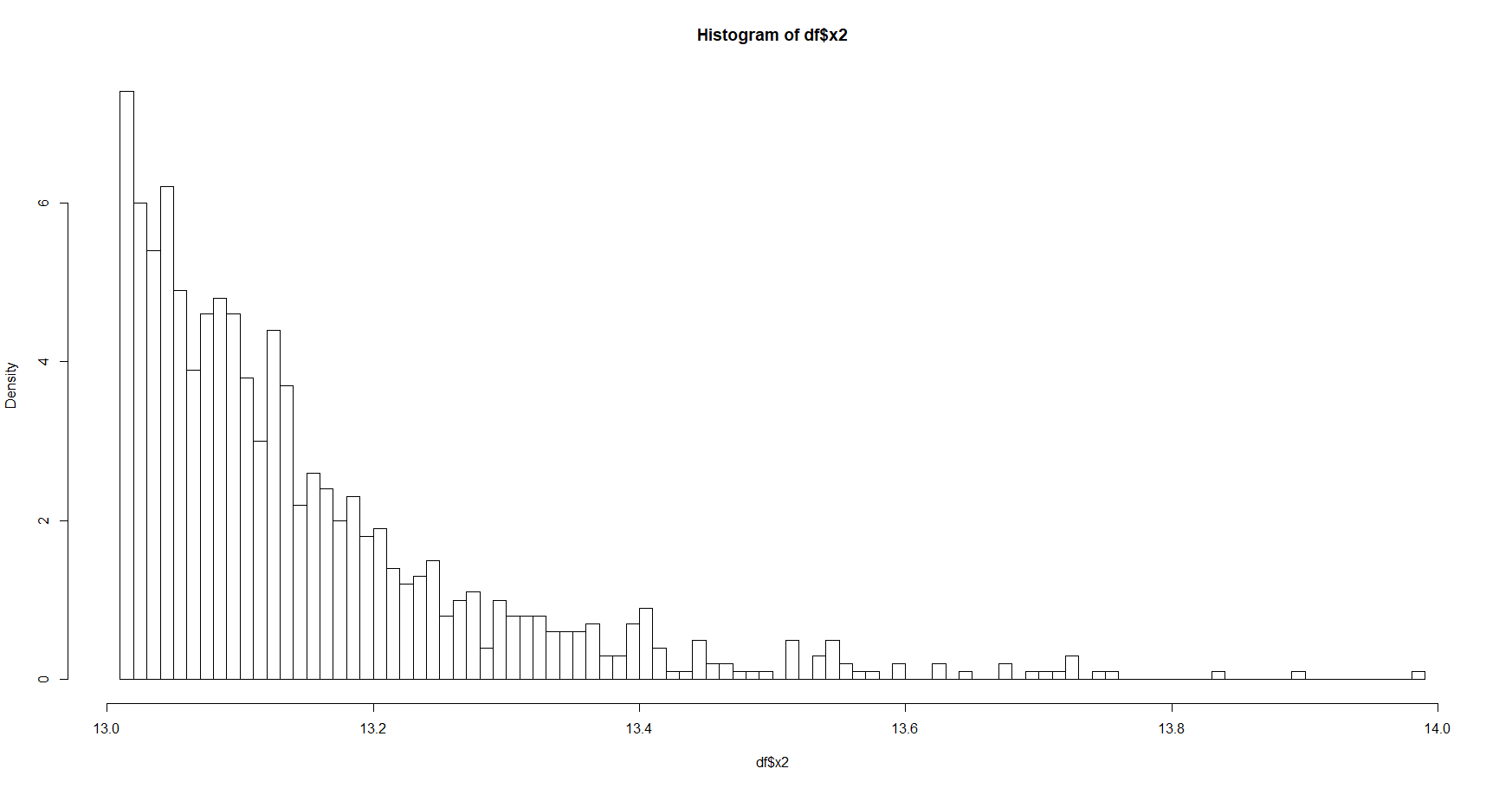
X3, x2 имеют большой коэф. куртозиса по сравнению с нормальным, хвосты распределений этих величин толще, чем у нормального. Тяжелые хвосты определяют вероятность редких событий.

X1, x5 имеют практически куртозис, как у нормального, у них также коэф. ассиметрии небольшой, поэтому их распределение визуально должно быть похожим на нормальное.

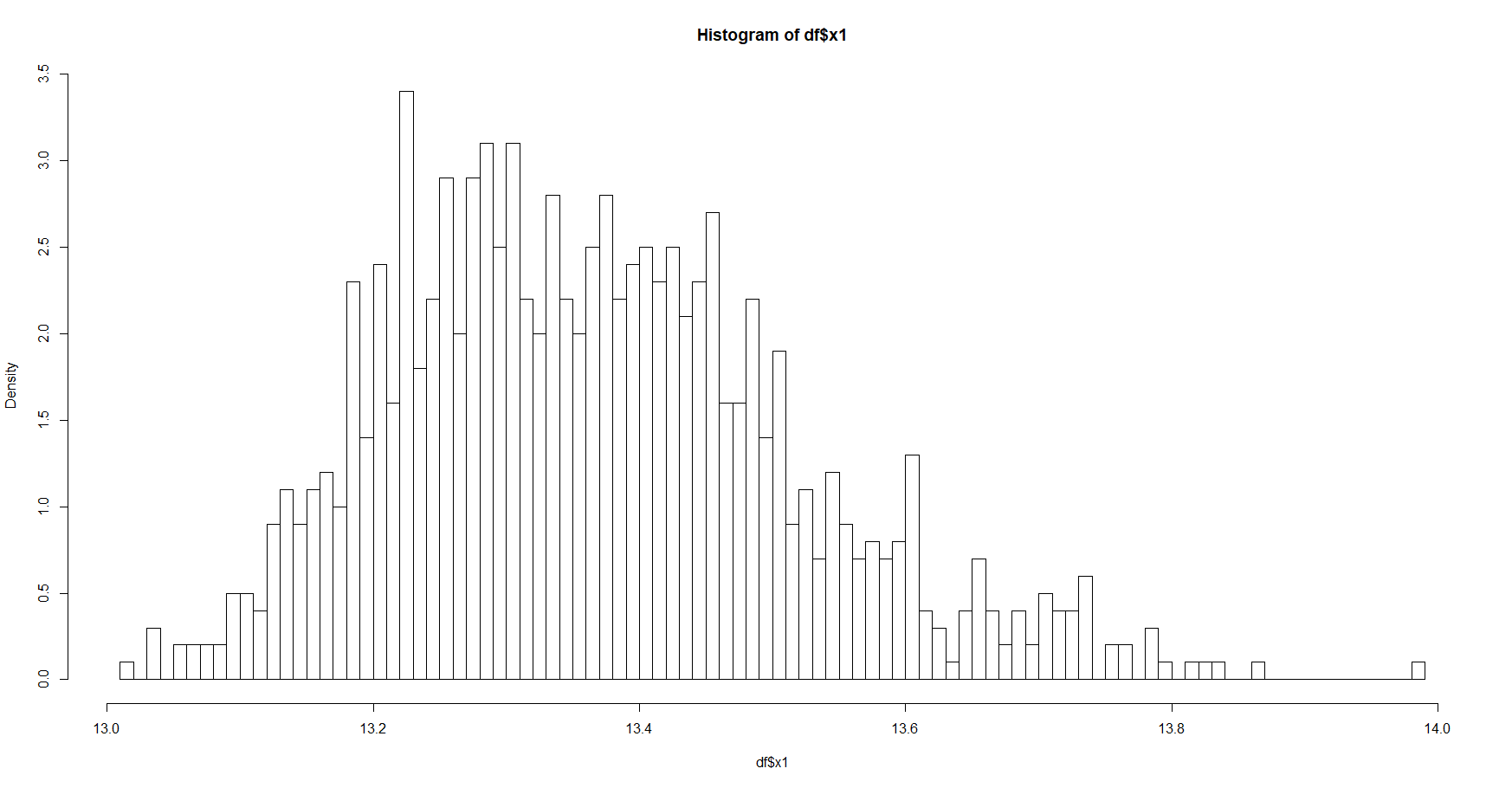
X4 имеет куртозис меньше, чем у нормального, значит у него более вероятность редких событий меньше, чем у нормального.

**1в) С помощью графиков примерно оцените моду 1-й переменной (или 2-й, если у 1-й слишком «плоское» распределение).**

Нашел для x1 и x2 переменных.



Для x2 моду найти просто, 13.0 является модой.



X1 имеет визуально распределение, напоминающее нормальное. Я пробовал делить на большее количество бинов, но картина практически не менялась. Здесь визуально мода лежит в этом интервале 13.25-13.3.

**Задача 2**

**2а) Найдите пары значимо коррелированных переменных. Какие из пар отрицательно, а какие положительно коррелированы?**

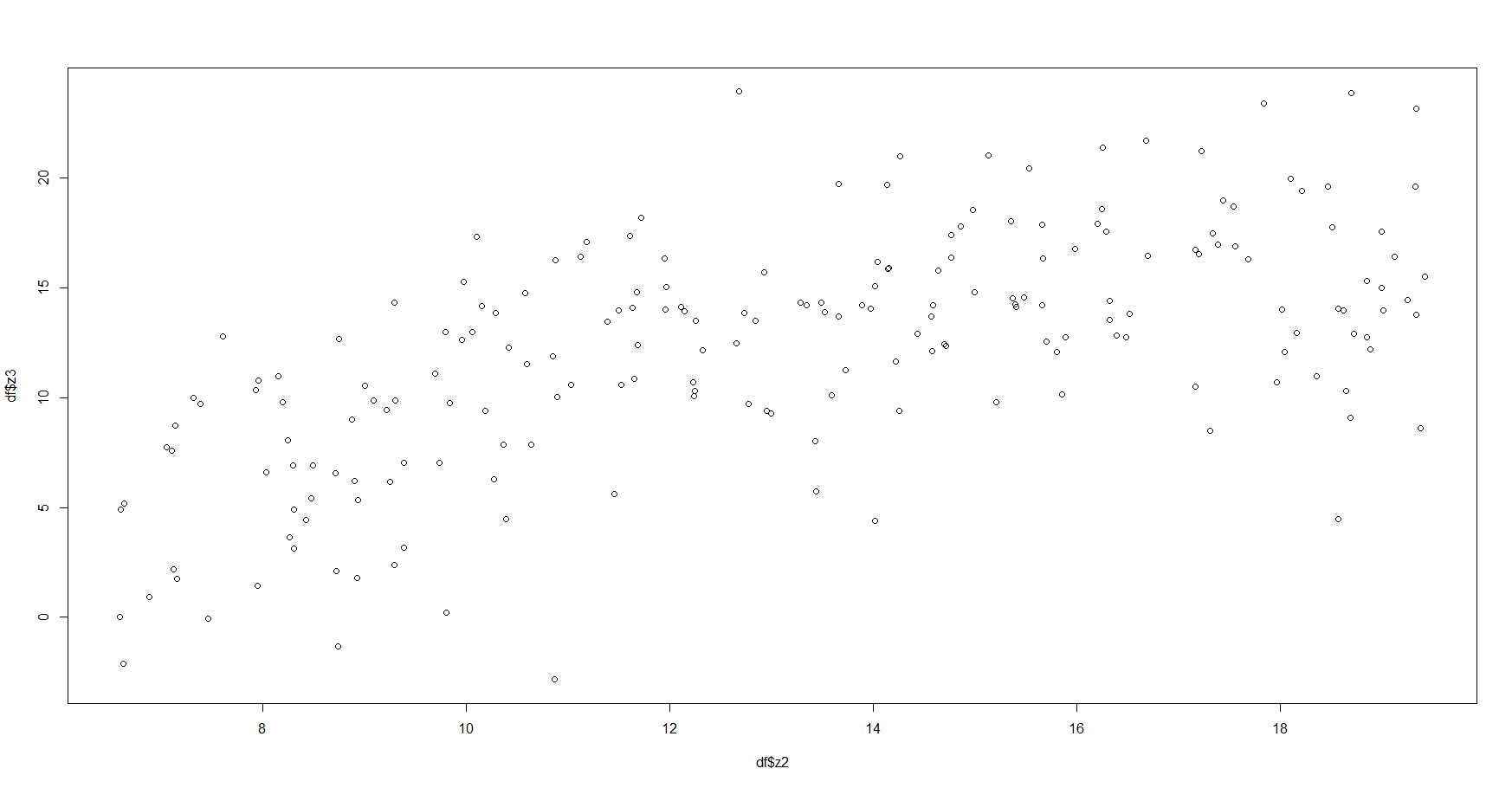
Корреляционная матрица:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| z1 | z2 | z3 | z4 | z5 | z6 | z7 | z8 |  |
| z1 | 1 | 0.014 | -0.028 | 0.031 | -0.004 | -0.111 | 0.115 | 0.904 |
| z2 | 0.014 | 1 | 0.62 | -0.093 | 0.009 | -0.077 | -0.013 | 0.014 |
| z3 | -0.028 | 0.62 | 1 | -0.018 | 0.063 | 0.009 | -0.078 | -0.026 |
| z4 | 0.031 | -0.093 | -0.018 | 1 | -0.116 | 0.434 | 0.159 | -0.003 |
| z5 | -0.004 | 0.009 | 0.063 | -0.116 | 1 | -0.064 | -0.687 | -0.031 |
| z6 | -0.111 | -0.077 | 0.009 | 0.434 | -0.064 | 1 | 0.156 | -0.091 |
| z7 | 0.115 | -0.013 | -0.078 | 0.159 | -0.687 | 0.156 | 1 | 0.119 |
| z8 | 0.904 | 0.014 | -0.026 | -0.003 | -0.031 | -0.091 | 0.119 | 1 |

По данным из таблицы обнаружены 4 значимых корреляции: (z8, z1), (z3, z2), (z7, z5), (z6, z4)

**2б) Среди пар найти нелинейную зависимость. Подтвердить тестом.**

Я нашел нелинейную зависимость между z3 и z2. Привел график:

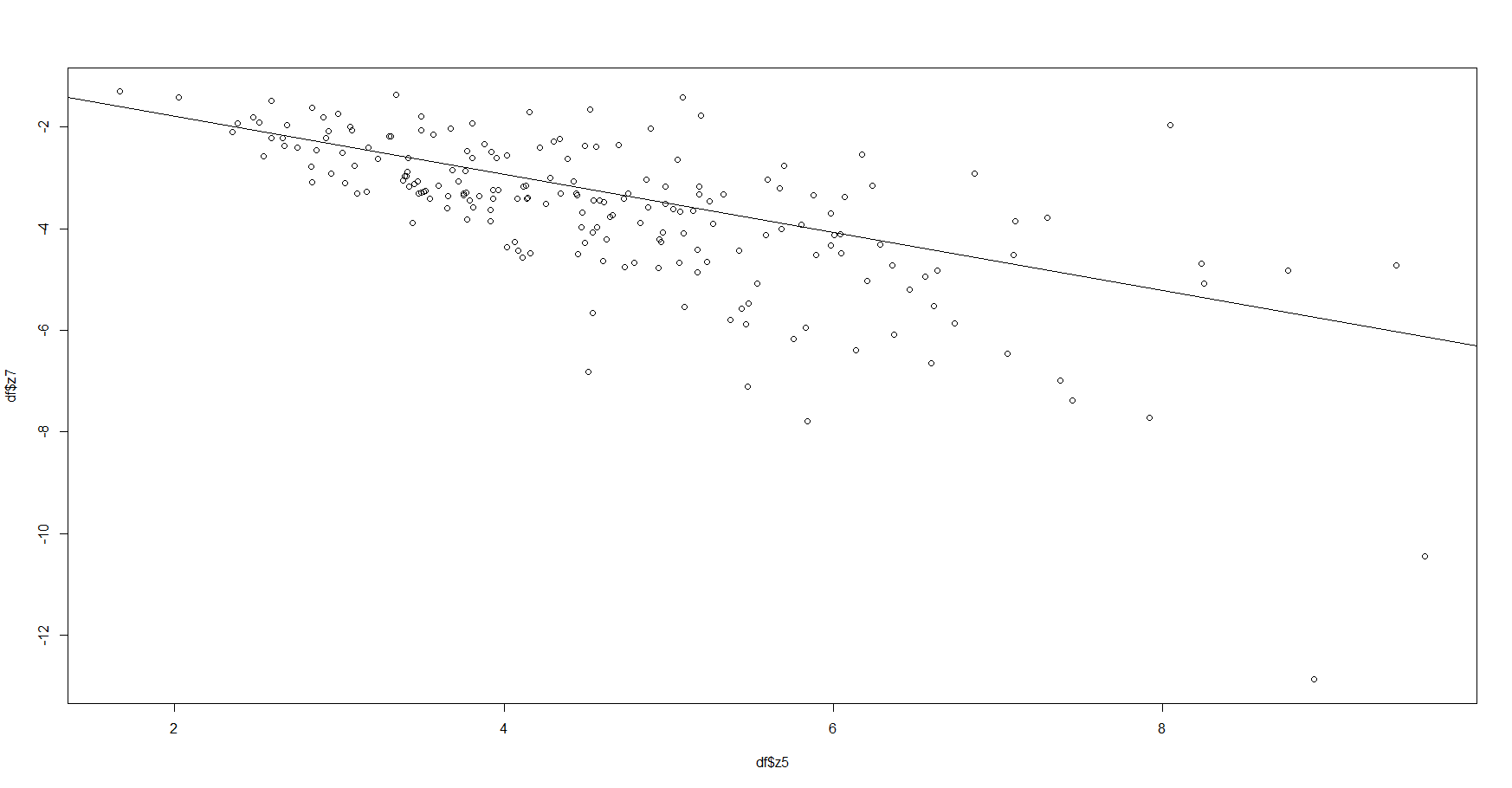


Чтобы проверить гипотезу, сделаем регрессию z3 от z2, после чего сделаем тест на значимость коэффициентов. В качестве теста я выбрал coeftest. В результате теста, линейный коэф. имеет p-value 9.8\*10^10, квадратичный коэф. 1,47\*10^-6. Значит при любом разумном уровне значимости (больше 2\*10^-6) оба коэф. значимы по отдельности, следовательно имеется нелинейность.

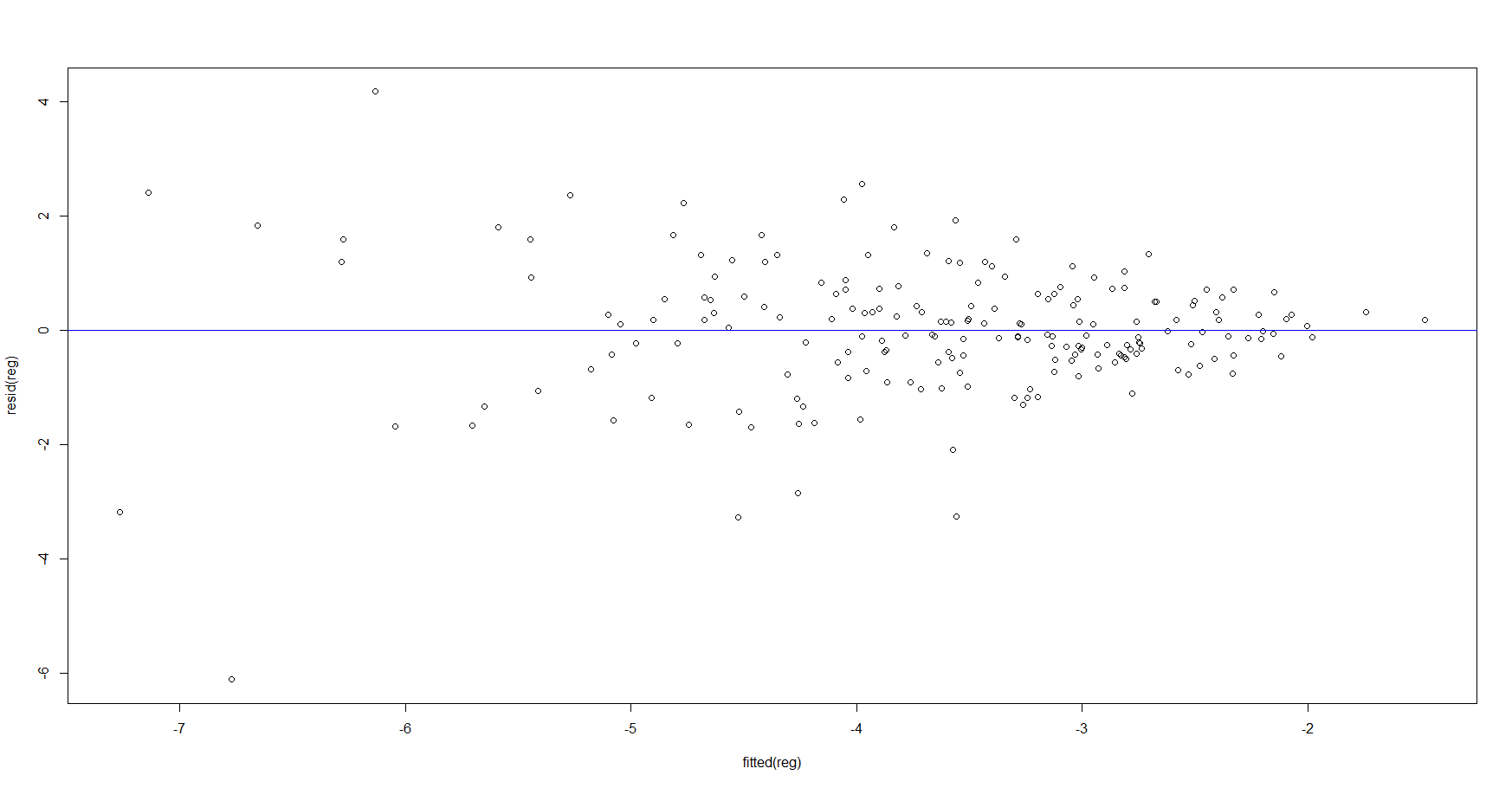
Также я сделал тест с правильной ковариационной матрицей из пакета sanwitch: coeftest(reg, vcov=vcovHC). P-value для квадратичного коэф. получился 2,26\*10^-6. Значит нулевую гипотезу о равенстве нулю коэф. отклоняем при любом разумном уровне значимости.

**2в) Среди пар найти зависимость с гетероскедастичностью. Обосновать наличие гетероскедастичности. Проверить гипотезу, о том, что коэффициент наклона = 1 (или =-1, если корреляция отрицательна) с робастной ковариационной матрицей на уровне значимости 5%.**

Нашел пару z5, z7, построил линейную регрессию. На рисунке разброс данных визуально увеличивается по мере увеличения z5. То есть предположительно дисперсия ошибок будет разная.



Построим график остатков (resid - остатки для регрессии) от расчетных значений (fitted - расчетное значение (y с крышкой)). Проведем линию 0. Визуально наблюдается изменение разброса остатков при изменении расчетного значения. График показан ниже:

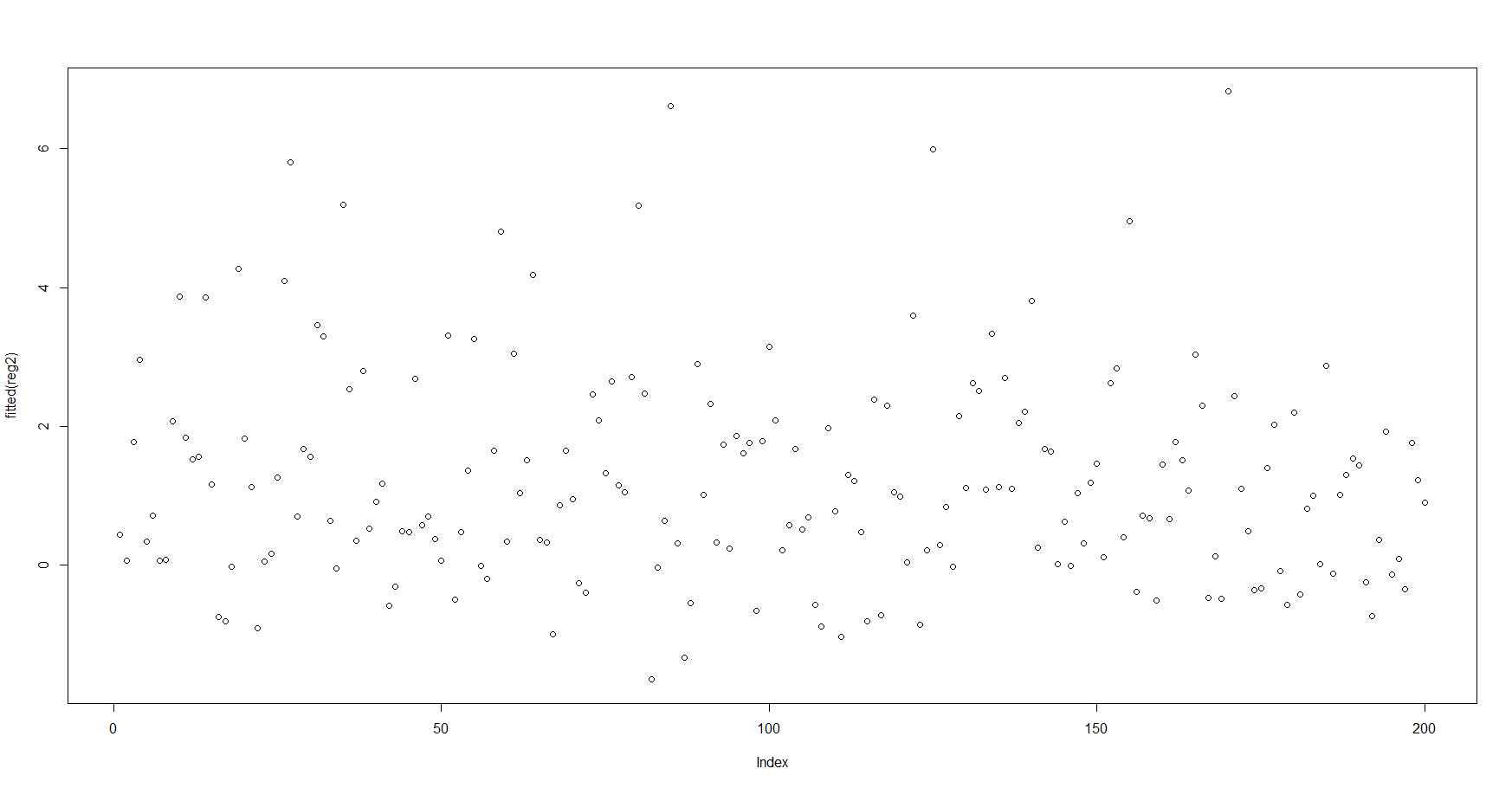


Создадим вспомогательную регрессию: reg2 <- lm(resid(reg)^2 ~ fitted(reg)) # квадраты остатков от расчетных значений. Проведем тест на гетероскедостичность. Нулевая гипотеза – нет гетероскедостичности. Используется F статистика.

Тест Бройша-Пейгана. p-value 9.08\*10^-05, нулевая гипотеза отвергаем.

Через summary: p-value 1.44\*10^-12, отклоняем нулевую гипотезу.

График остатков для регрессии показан ниже:



Дале был построен доверительный интервал для коэффициентов регрессии (coefci функция) с правильной ковариационной матрицей, устойчивой к гетероскедостичности. 95 % доверительный интервал для коэф. наклона получился: -0,932 : -0,527. -1 не попала в него, значит гипотеза о том, что коэф. равен -1 отклоняется при 5% уровне значимости, коэф. значимо отклоняется от минус 1.