**Липецкий государственный технический университет**

Факультет автоматизации и информатики

Кафедра­ автоматизированных систем управления

ЛАБОРАТОРНАЯ РАБОТА №2

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПРИКЛАДНЫХ ЗАДАЧАХ

РЕГРЕССИОННЫЙ АНАЛИЗ

|  |  |  |
| --- | --- | --- |
| Студент | \_\_\_\_\_\_\_\_\_\_\_\_  подпись | Станиславчук С.М.  Коновалов К. А. |
| Группа АС-21-1 |  |  |
| Руководитель |  |  |
| Доцент | \_\_\_\_\_\_\_\_\_\_\_\_  подпись | Рыжкова Д.В. |
|  |  |  |

Липецк 2023 г.

Цель работы

На реальных данных (массив данных - независимая Х и зависимая Y, 86 строк), несущих смысловую нагрузку, провести регрессионный анализ, а именно:

1. Проверить выборку данных на нормальность.

2. Вывести уравнение линейной регрессии y=α+βx, оценки коэффициентов получить с помощью методов: МНК, метод Бартлетта-Кенуя.

3. Оценка статистической значимости выборочной регрессии.

4. Определение доверительных областей, включающих в истинную регрессию с заданной вероятностью.

5. Анализ регрессионных остатков.

6. Анализ наличия грубых отклонений от регрессии (выбросов).

7. Построение толерантных границ для регрессии.

Краткая теоретическая справка

Регрессионный анализ – статистический метод исследования влияния одной или нескольких независимых переменных X на зависимую переменную Y.

Линейный регрессионный анализ исходит из наличия зависимости y=α+βx, где α и β – неизвестные коэффициенты регрессии.

Ход выполнения лабораторной работы

Был сформирован массив из реальных данных. Y – % людей пенсионного возраста, а X – потребление алкогольных напитков на душу населения, литр. Данные представлены на рис. 1.

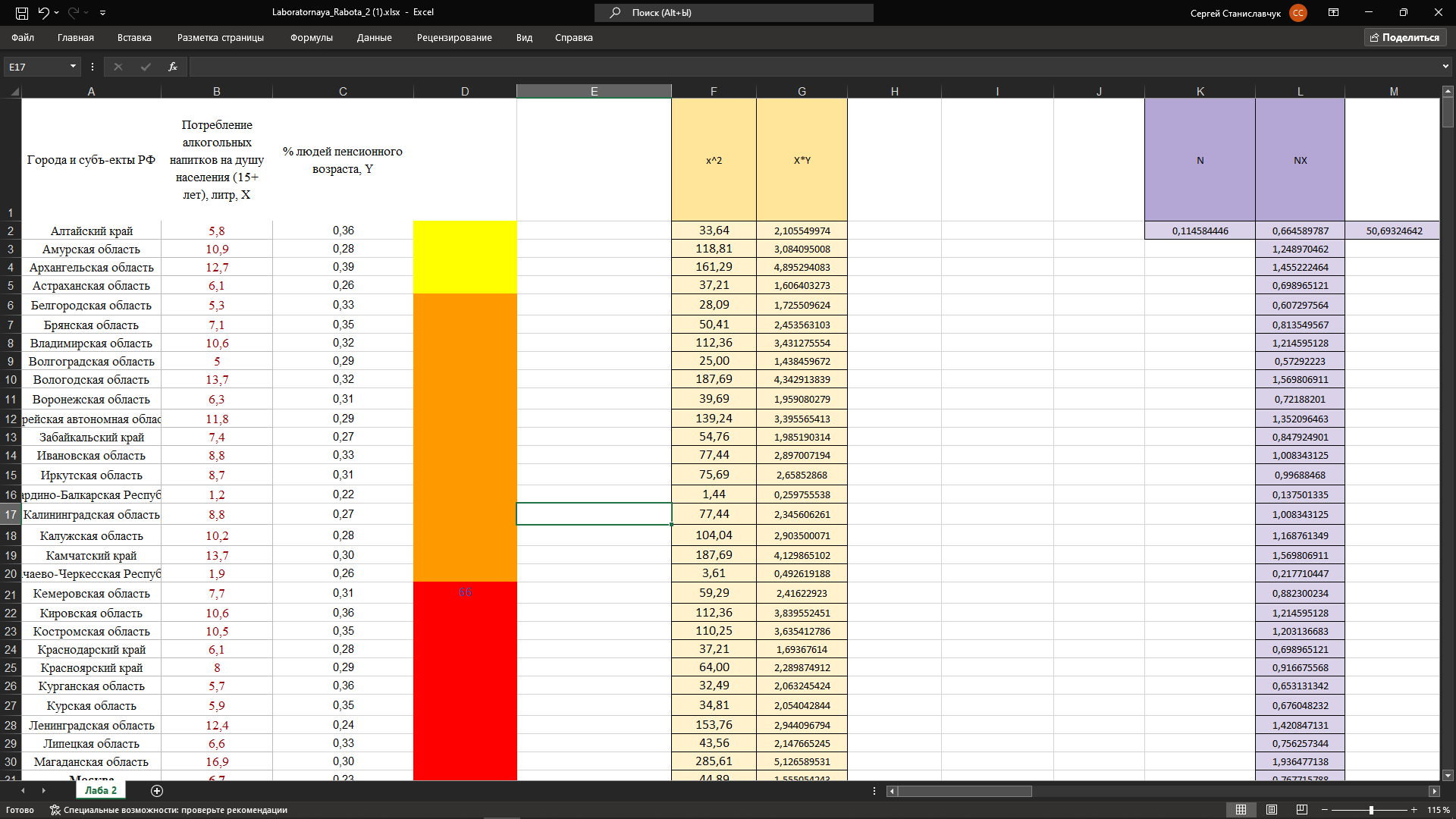


Рисунок 1 – Исходные данные

## Проверка выборки данных на нормальность.

Для проверки используем критерий «хи»-квадрат. Расчет по формуле:

Проверка на нормальность осуществлялась для x и y. Для начала были определены: xср, yср, количество интервалов M, минимальные/максимальные значения переменных, а также шаг ∆/M, изображенные на рис. 2.

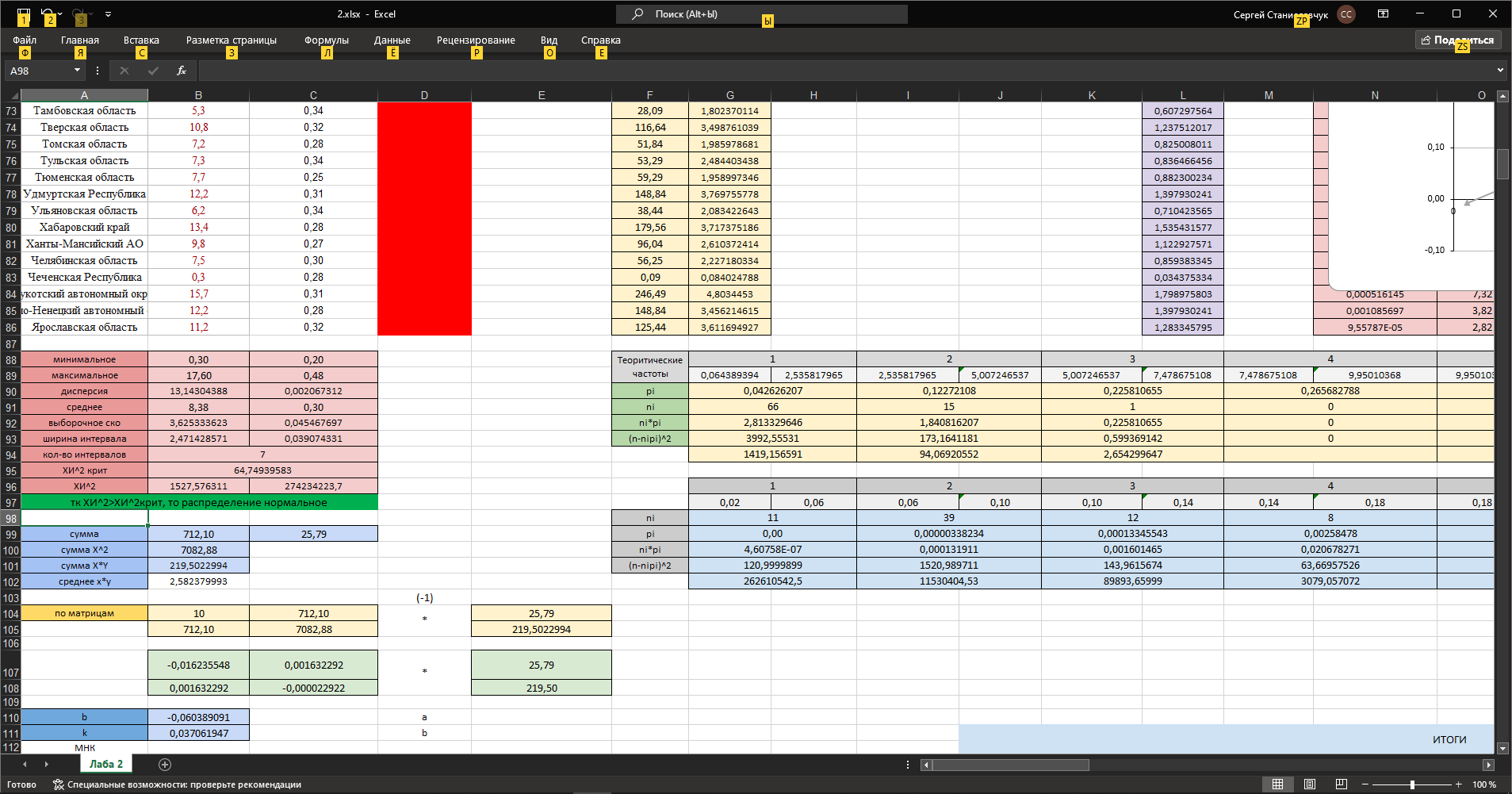


Рисунок 2 – Поясняющие данные для X и Y

На рис. 3 представлены расчеты для Х. χ2экс был посчитан для всех значений с помощью суммирования полученных результатов:

χ2экс = 1527,576311.

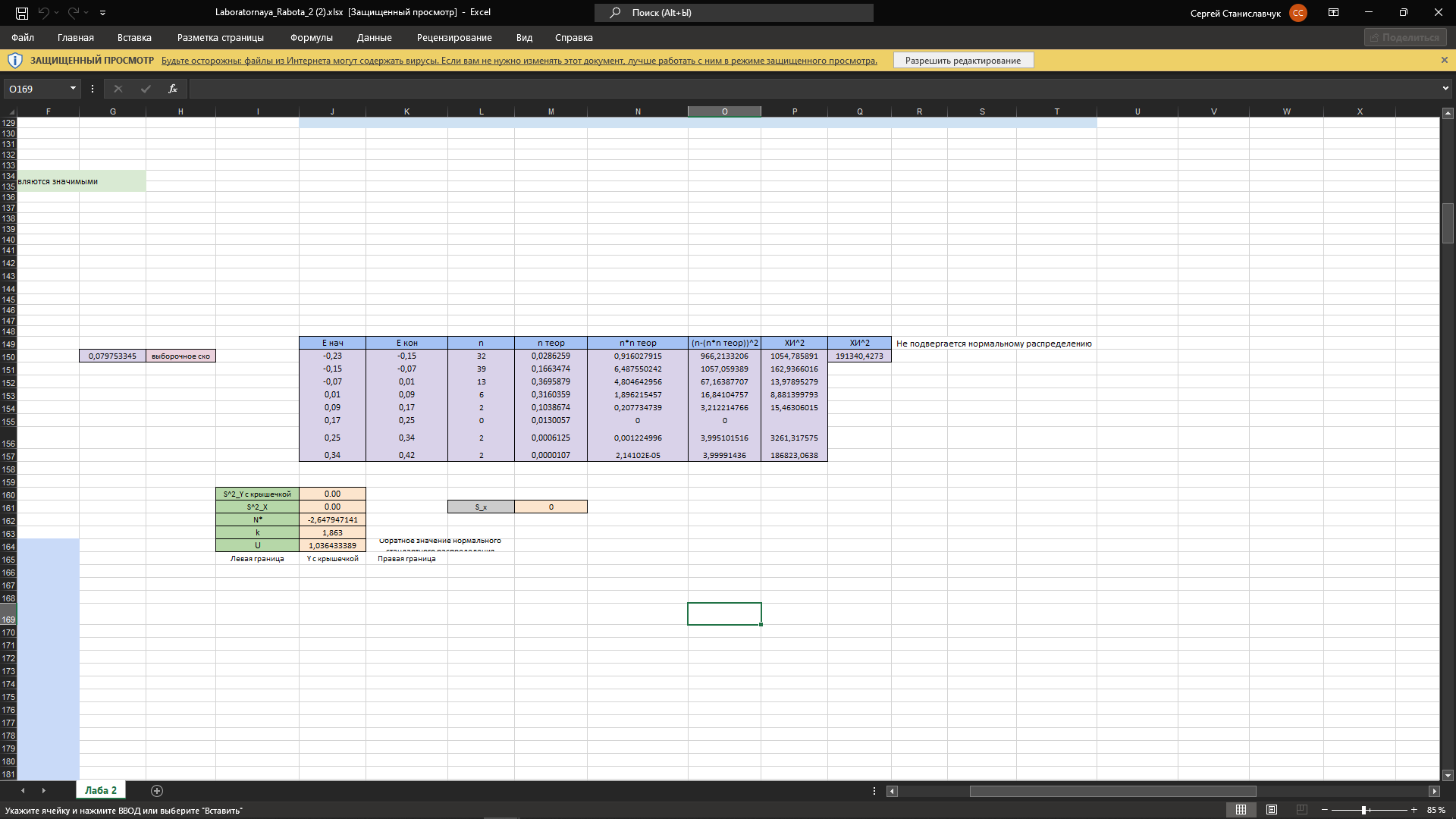


Рисунок 3 – Проверка на нормальность для Y

χ2экс для Y составил 274234223,7.

χ2крит подсчитывалось по встроенной формуле =ХИ2.ОБР(). Так как χ2крит зависит от α(0.05) и степеней свободы, то для X и Y этот параметр будет одинаков. χ2крит = 64,74939583.

Для X: 1527,576311 > 64,74939583 и для Y: 274234223,7 > 64,74939583

Вывод: выборка X и выборка Y подчиняются нормальному закону распределения.

## 2. Вывод уравнения линейной регрессии

Линейное парное уравнение регрессии имеет вид: y=α+β\*xi, i=1, …,n , где n – объем совокупности (число наблюдений).

Оценки параметров линейной регрессии (a и b) могут быть найдены разными методами, наиболее распространенным является метод наименьших квадратов. Данный метод позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака yi от расчетных (теоретических) значений yi\* (рассчитанных по уравнению регрессии) минимальна.

Формулы для расчета:

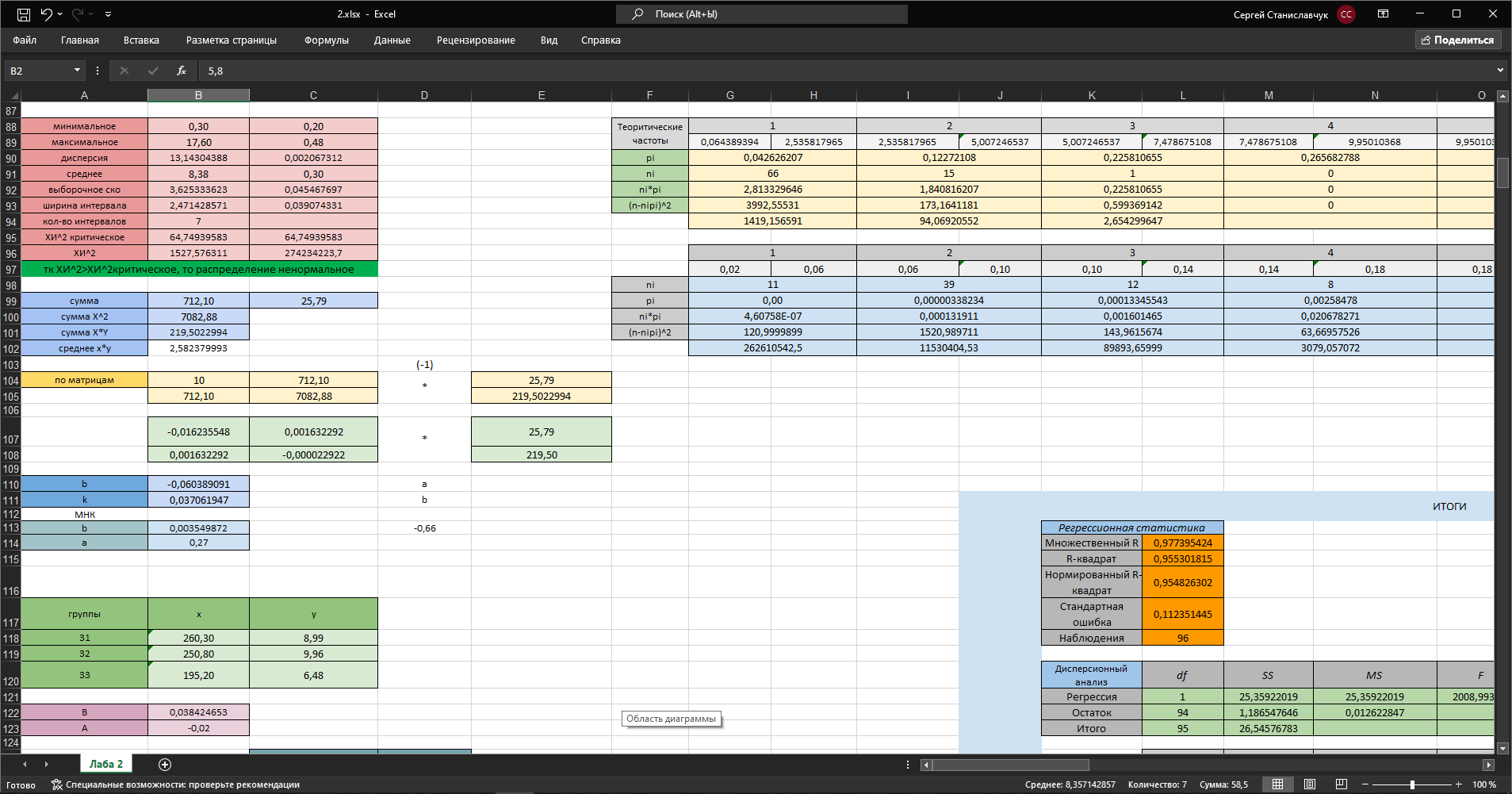


Рисунок 4 – МНК

Пары наблюдений (yi, xi) упорядочиваются по x и разбиваются на 3 примерно равные группы, так чтобы первая и третья группа были обязательно разного объема.

В каждой группы находятся суммы  и 

Тогда коэффициенты регрессии находятся с помощью соотношений:



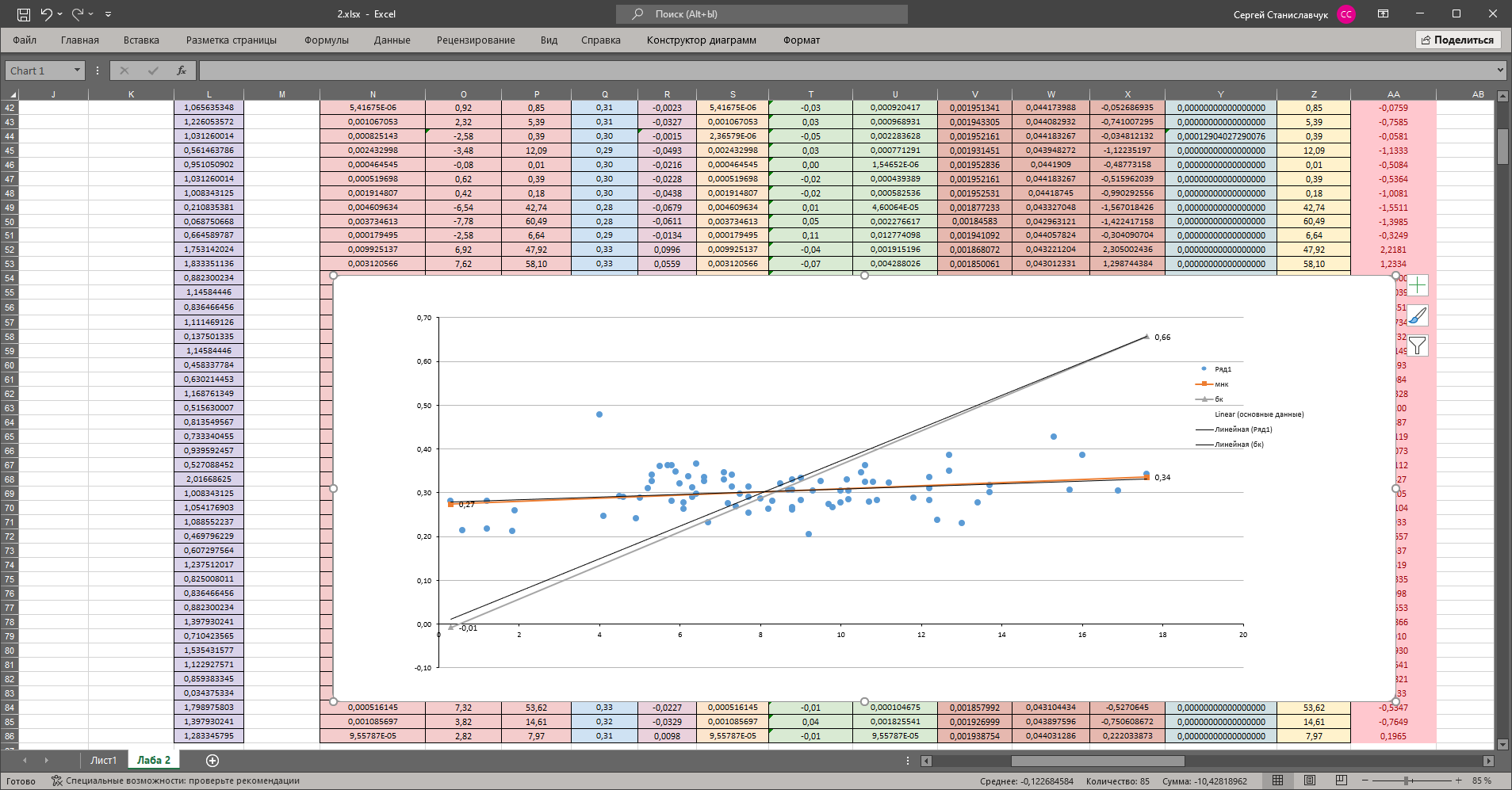


Рисунок 5 – График

Был построен график (рис. 5), на котором отображен разброс начальных данных, и прямые, полученные с помощью МНК и метода Бартлетта-Кенуя.

Как видно на рис. 5, линия тренда для X,Y совпала с прямой МНК, а значит значения были получены верно.

## 3. Оценка статистической значимости выборочной регрессии

Статистические выводы относительно коэффициента β регрессии y=α+βx получаются с помощью статистики:



Значение коэффициента β регрессии является значимым, если выполняется 

Для коэффициента α используется статистика: 

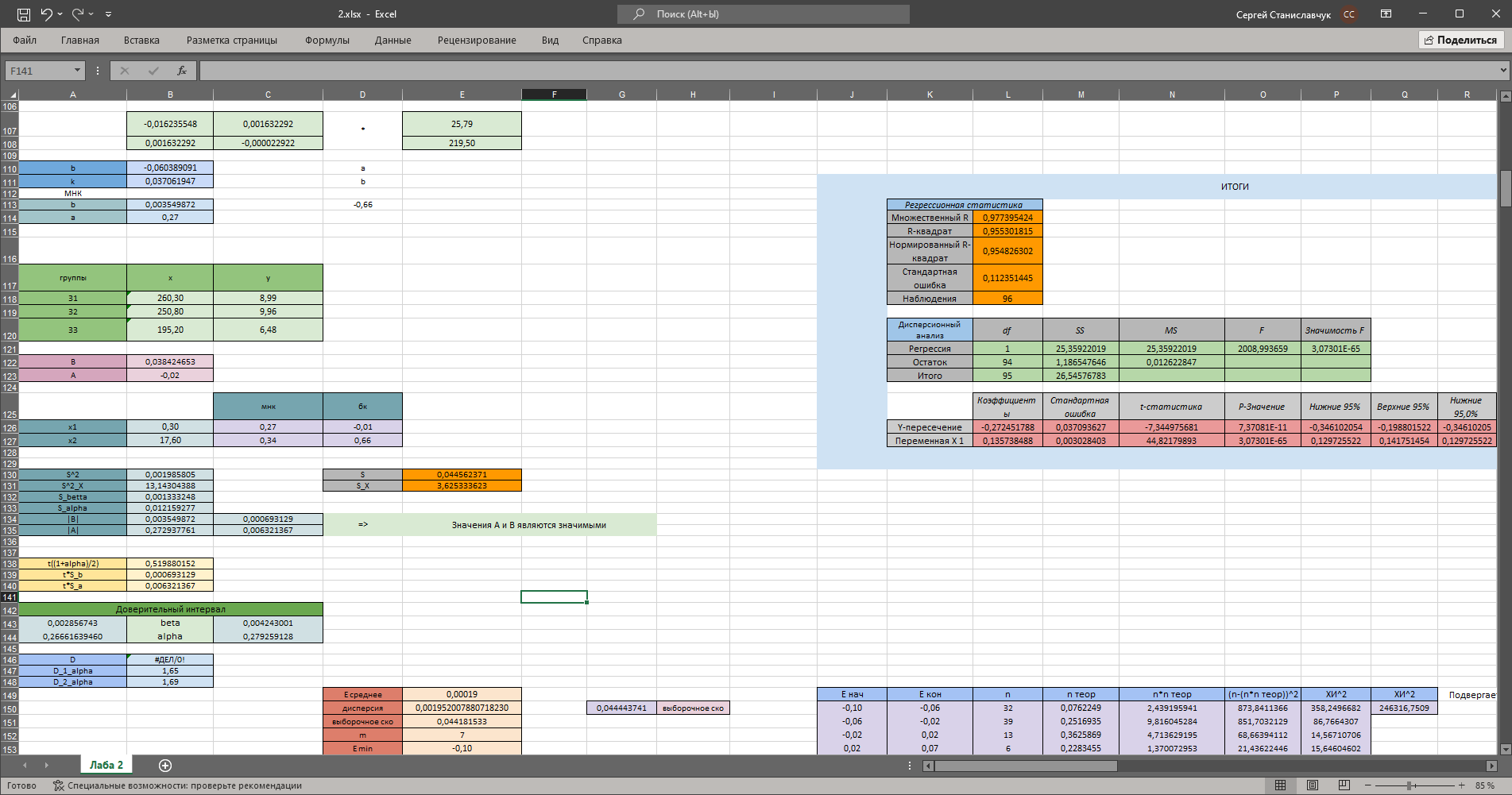


Рисунок 6 – Оценка значимости

Исходя из расчетов (рис. 6), можно сделать вывод о том, что коэффициенты a и b являются значимыми (t\*S\_b < |B| и t\*S\_a < |A|)

## 4. Определение доверительных интервалов

Двусторонний доверительный интервал для β имеет вид:



Двусторонний доверительный интервал для α имеет вид:



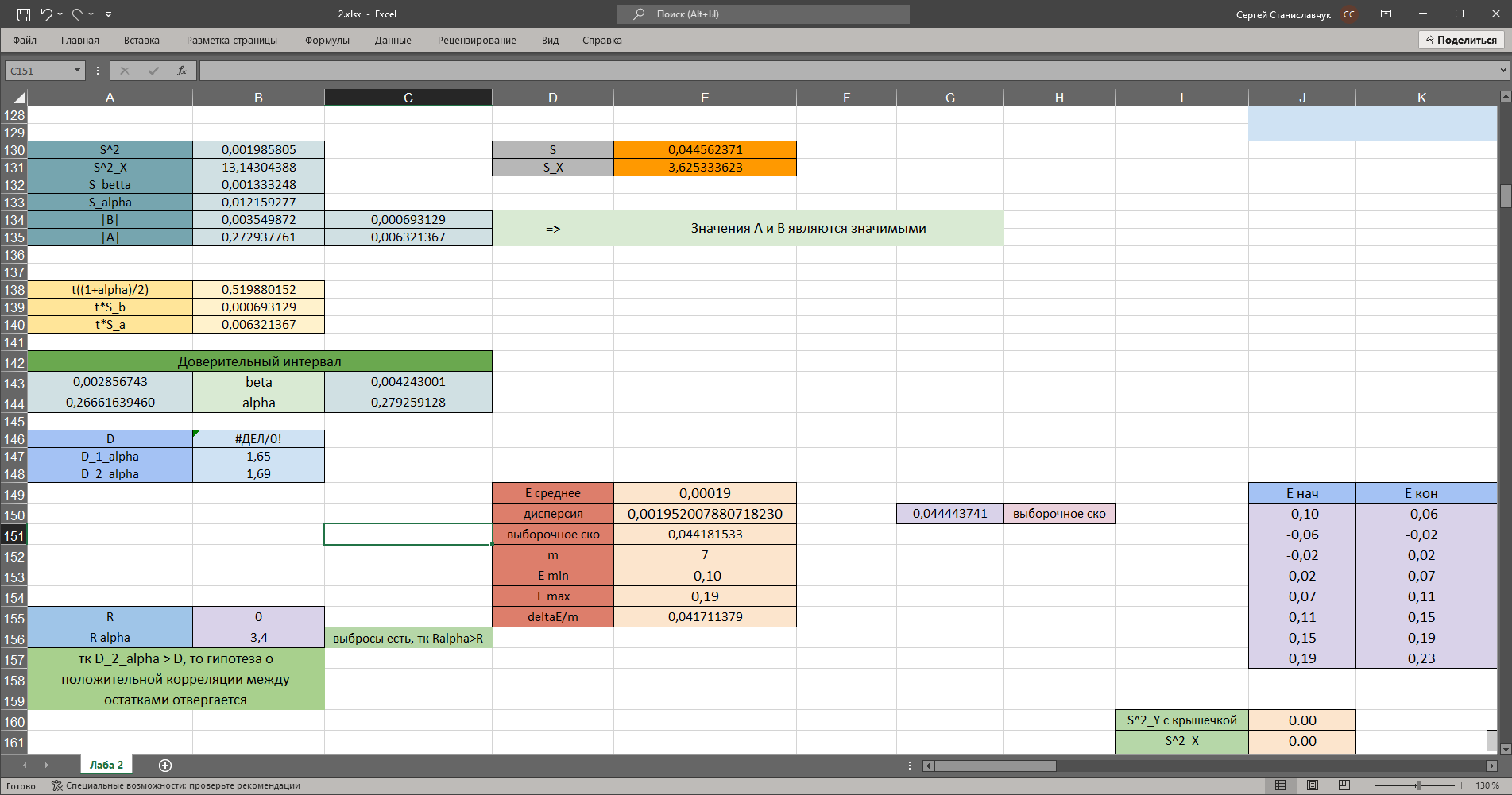


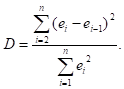
Рисунок 7 – Определение доверительных интервалов для значений a и b.

## 5. Анализ регрессионных остатков

Значительную информацию об адекватности уравнения регрессии дает анализ ее остатков. Остатки вычисляются по формуле где 

Если уравнение регрессии хорошо описывает исходные данные, то остатки должны быть распределены по нормальному закону. Поэтому исследование остатков должно включать в себя проверку гипотезы о нормальности распределения.

Независимость остатков может быть проверена с помощью статистики Дарбина-Уотсона:



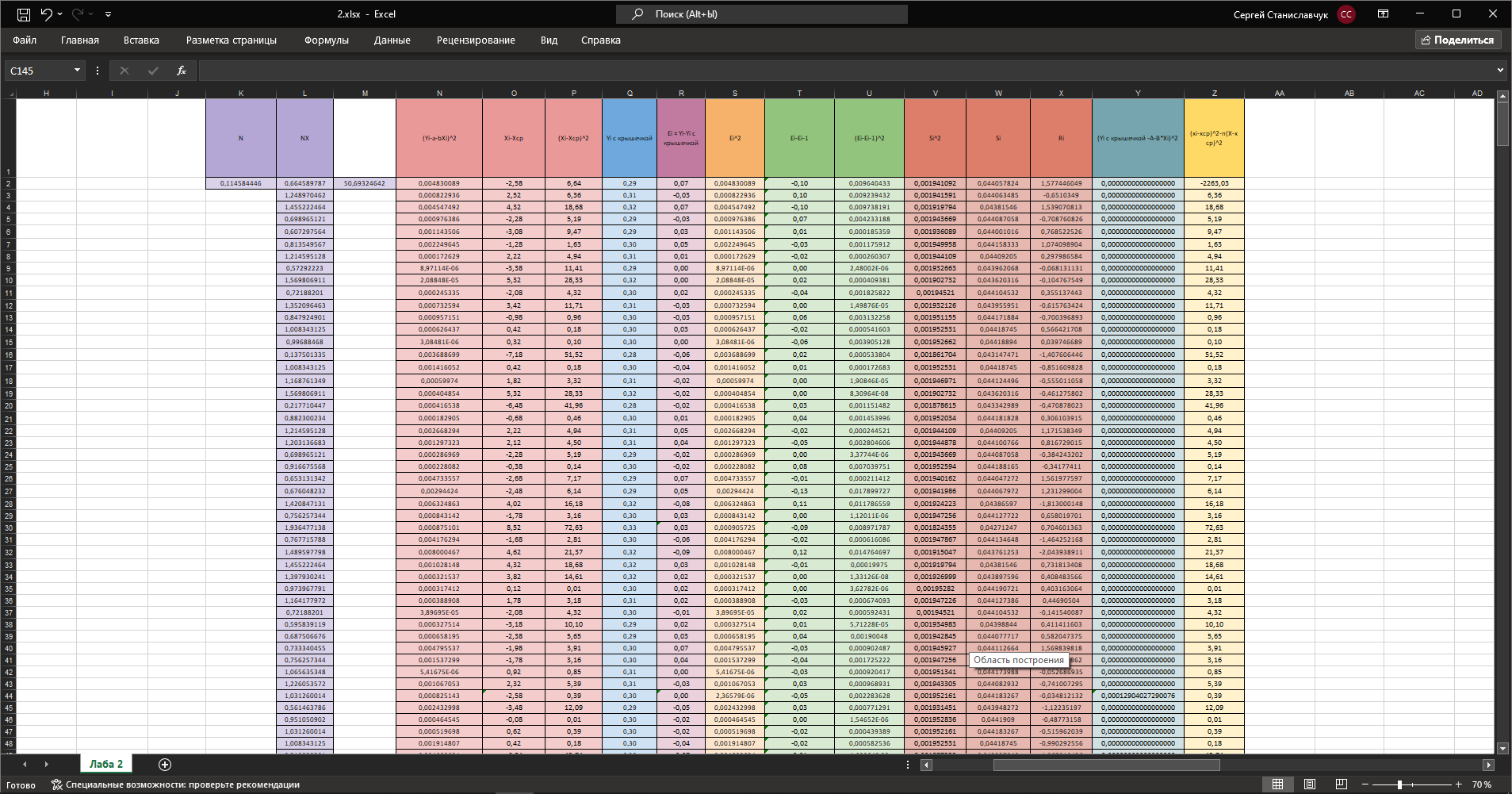


Рисунок 8 – Анализ регрессионных остатков

Отдельно посчитаны ei,ei2, (ei - ei-1)2.

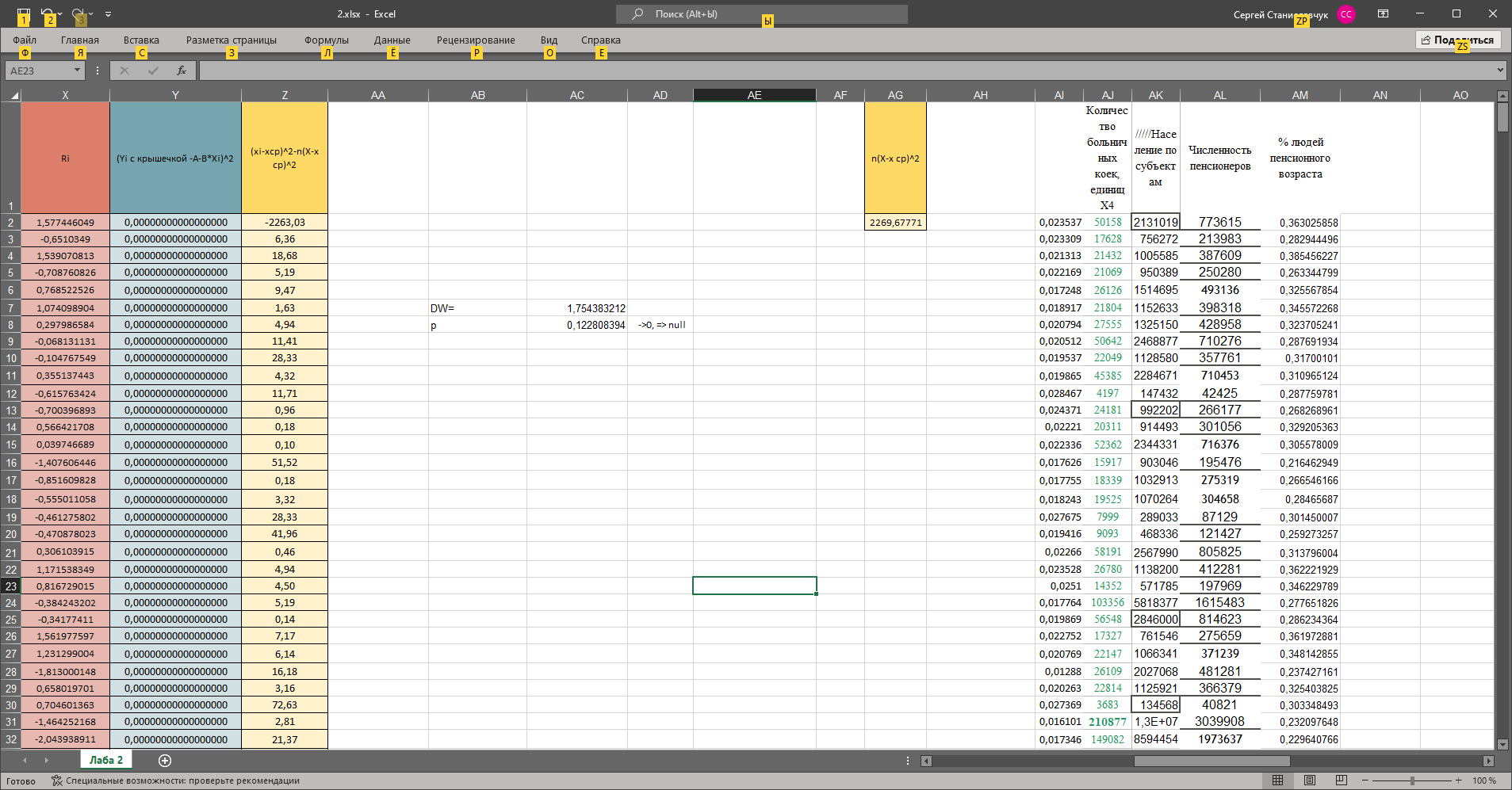


Рисунок 9.

На рис. 9 видно, что выполняется условие DW -> 2, p->0, это говорит о том, что автокорреляция отсутствует.

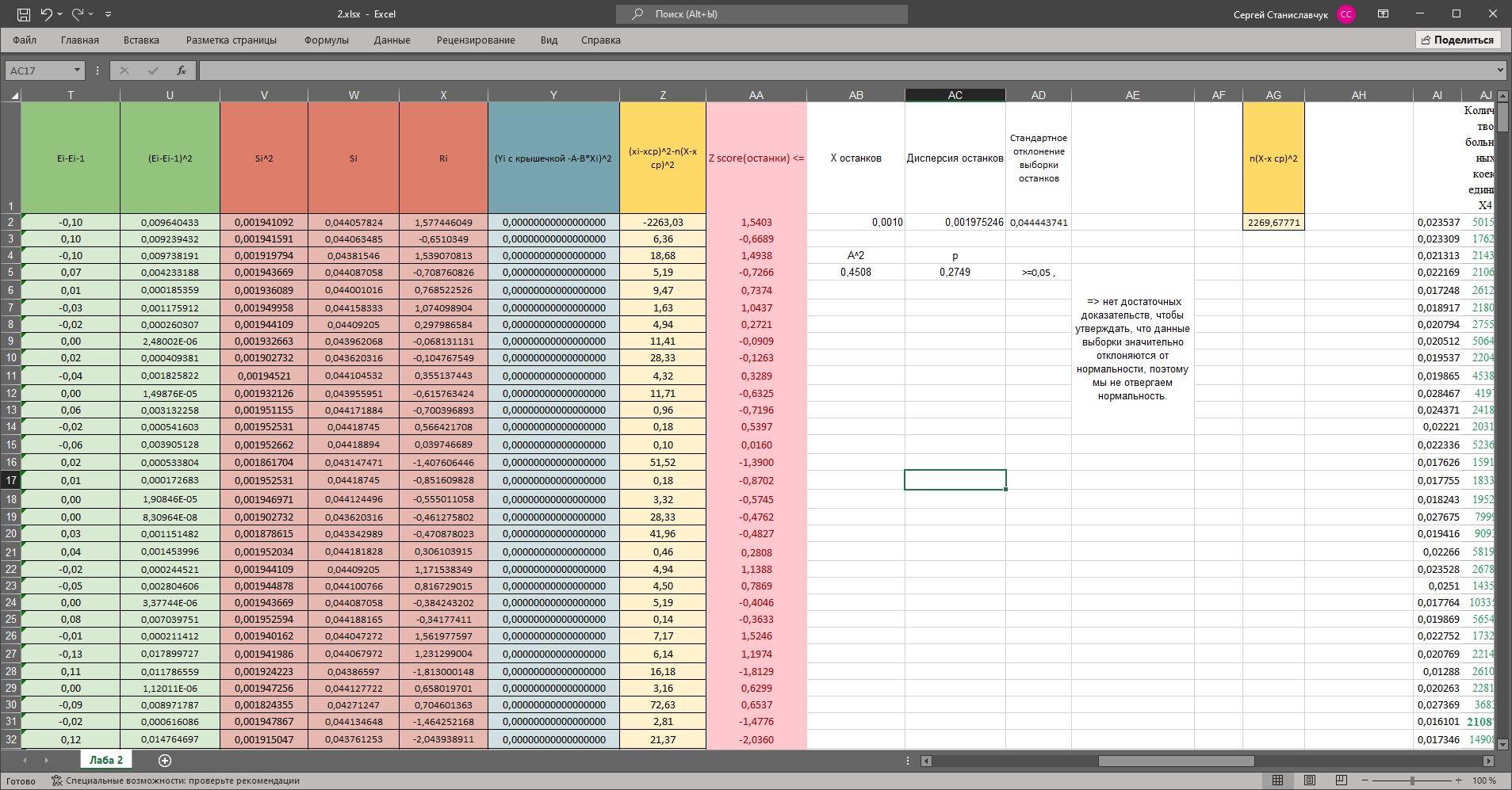


Рисунок 10. Проверка гипотезы нормальности – Андерсон-Дарлинг

Для проверки на нормальность так же использовался критерий «хи»-квадрат, который показал, что выборка подчиняется нормальному закону распределения.

## 6. Анализ наличия грубых отклонений от регрессии (выбросов)

## Способ - межквартильный диапазон.

Межквартильный размах (IQR) — это разница между 75-м процентилем (Q3) и 25-м процентилем (Q1) в наборе данных. Он измеряет разброс средних 50% значений.

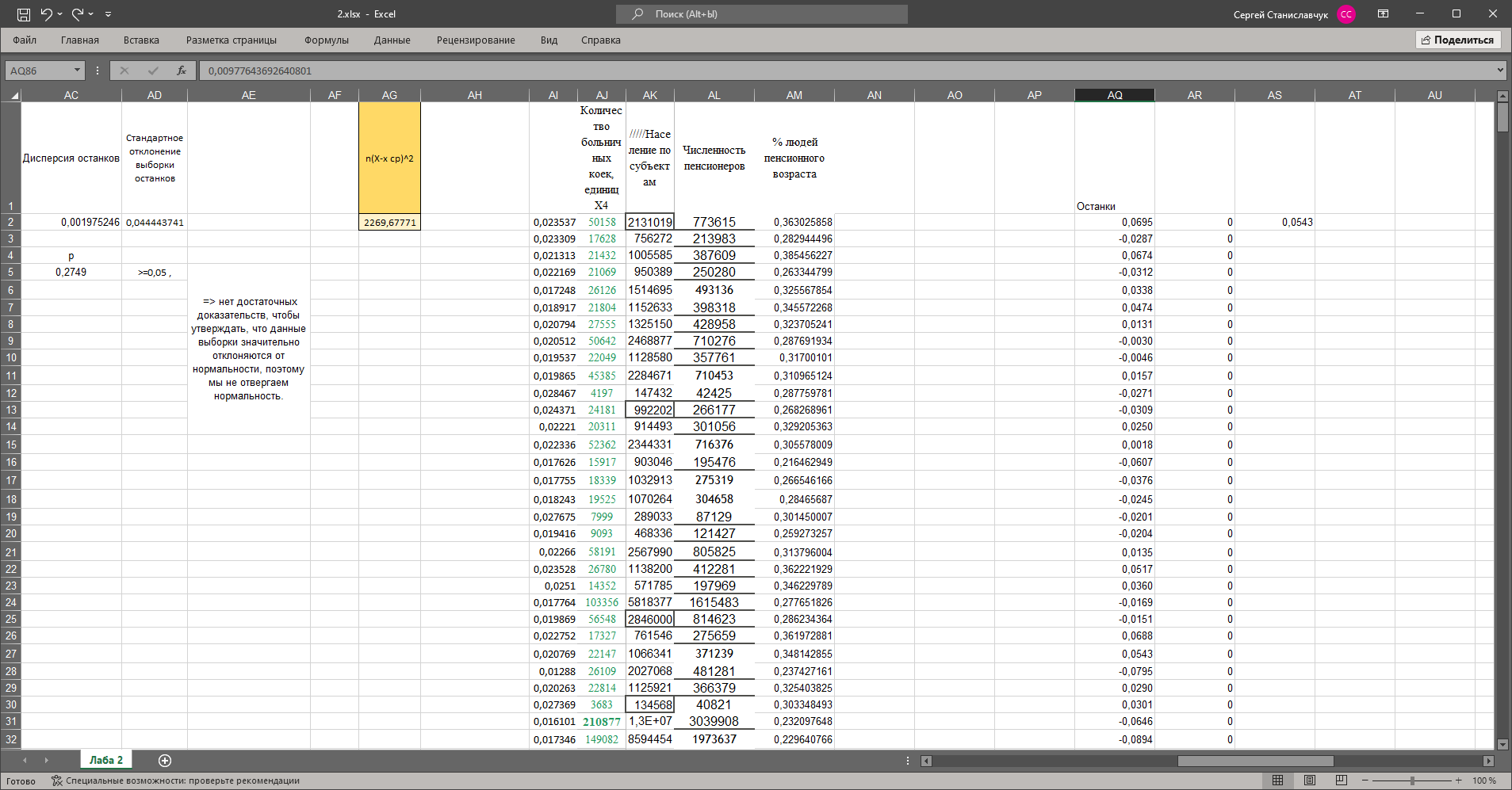


Рисунок 12. Анализ наличия выбросов (IQR)

## Мы можем определить наблюдение как выброс, если оно в 1,5 раза превышает межквартильный размах, превышающий третий квартиль (Q3), или в 1,5 раза превышает межквартильный размах меньше, чем первый квартиль (Q1)

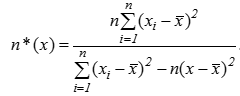
## В нашем случае выбросов среди остатков не обнаружено.

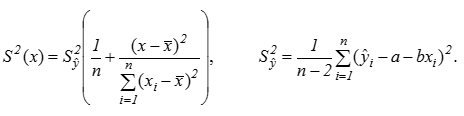
## 7. Построение толерантных границ для регрессии

Для того, чтобы найти толерантный интервал, для которого можно утверждать с вероятностью α, что внутри него лежит β часть значений, необходимо найти доверительные интервалы для каждого значения икса. Искомая толерантная область будет геометрическим местом точек толерантных интервалов для отдельных xi.

Двусторонний толерантный интервал имеет вид:

, где

,



Показанные на рис. 11 и найденные по приведенным выше формулам толерантные границы для регрессии были построены

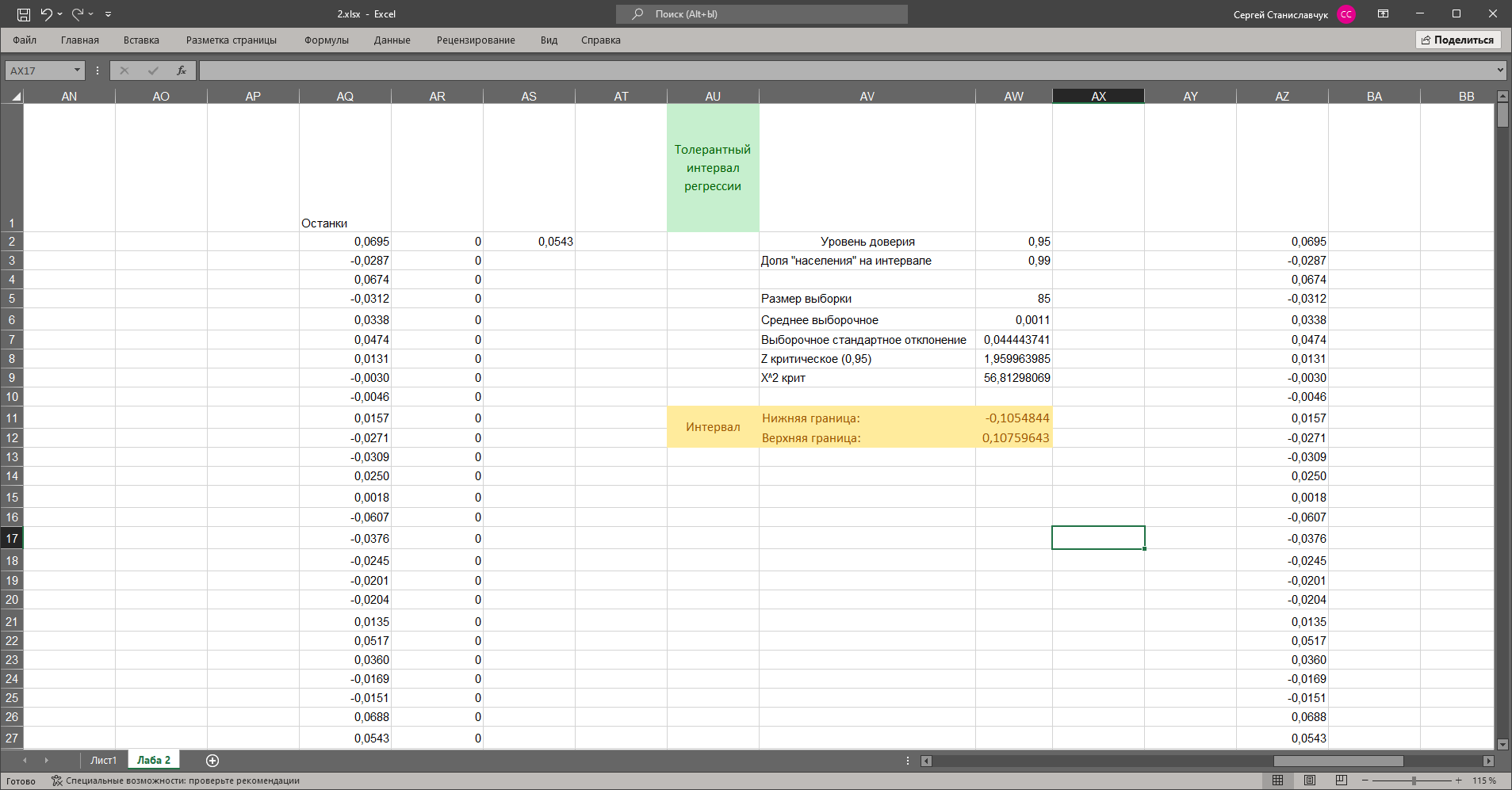


Рисунок 9 – Построение толерантных границ

Выводы

В ходе лабораторной работы был проведен регрессионный анализ и получены следующие результаты:

1. Выборки данных X и Y подчиняются нормальному закону распределения.

(χ2экс (Х)) 1527,576311 > 64,74939583 (χ2крит)

(χ2экс (Y)) 274234223,7 > 64,74939583 (χ2крит)

=> мы можем применять множество разных методов, рассчитанных на нормальность распределения данных.

2. Получено уравнение линейной регрессии.

МНК: y = 0,27 + 0,003549872x;

Метод Бартлетта-Кенуя: y= -0,02 - 0,038424653x

Исходя из полученного уравнения, видим некоторую положительную зависимость, между X и Y => употребление алкоголя довольно-таки распространено именно среди пенсионеров.

3. Проведена проверка статистической значимости выборочной регрессии. Коэффициенты a и b признаны значимыми, т.к. t\*S\_b < |B| и t\*S\_a < |A|

=> между X и Y есть функциональная связь

4. Определены доверительные интервалы для коэффициентов α и β.

0,002856743 < β < 0,004243001

0,26661639460 < α < 0,279259128

=> с вероятностью 95% доверительный интервал будет содержать наши значения коэффициентов.

\*коэффициенты α и β в нашем случае попали в этот интервал.

5. Проведен анализ регрессионных остатков. Принята гипотеза о отсутствии автокорреляции между остатками (т.е. значения предыдущих остатков никак не влияют на значения последующих), а также проверка на нормальность показала, что выборка подчиняется нормальному закону распределения.

6. Проведен анализ наличия грубых отклонений от регрессии. Выявлено, что выбросов нет => данные подобраны корректно, т.к. находятся близко друг к другу.

7. Построение толерантных границ выполнено успешно: все значения линии регрессии попадают в толерантные интервалы.