Министерство науки и высшего образования Российской Федерации

Новосибирский Государственный технический университет

Кафедра автоматизированных систем управления



**Отчет по лабораторной работе 1**

**по дисциплине «Методы анализа данных»**

**«Предварительный анализ данных, дисперсионный анализ, корреляционный анализ, регрессионный анализ в среде MS Excel»**

Выполнили

студенты группы АВТ-812:

Антонянц Егор

Гостеева Анна

Амельченко Артем

Преподаватель:

Ганелина Наталья Давидовна,

к.т.н., доцент кафедры АСУ

г. Новосибирск

2021 г.

Содержание

[1 Цель работы 3](#_Toc84350911)

[2 Описание исходных данных 3](#_Toc84350912)

[3 Результаты дескриптивного анализа 3](#_Toc84350913)

[4 Анализ результатов таблиц сопряженности 5](#_Toc84350914)

[5 Результаты корреляционного анализа 6](#_Toc84350915)

[7 Регрессионная модель. Уравнение регрессии, график остатков, гистограмма остатков. Оценка качества построенной модели 8](#_Toc84350916)

[8 Сравнение регрессионных моделей 11](#_Toc84350917)

[9 Формулировка гипотез. Описание градаций фактора. Результаты дисперсионного анализа 12](#_Toc84350918)

[10 Двухфакторный дисперсионный анализ 14](#_Toc84350919)

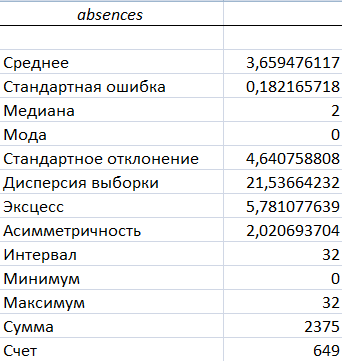
# 1 Цель работы

Ознакомиться с алгоритмами применения дисперсионного, корреляционного и регрессионного анализа данных.

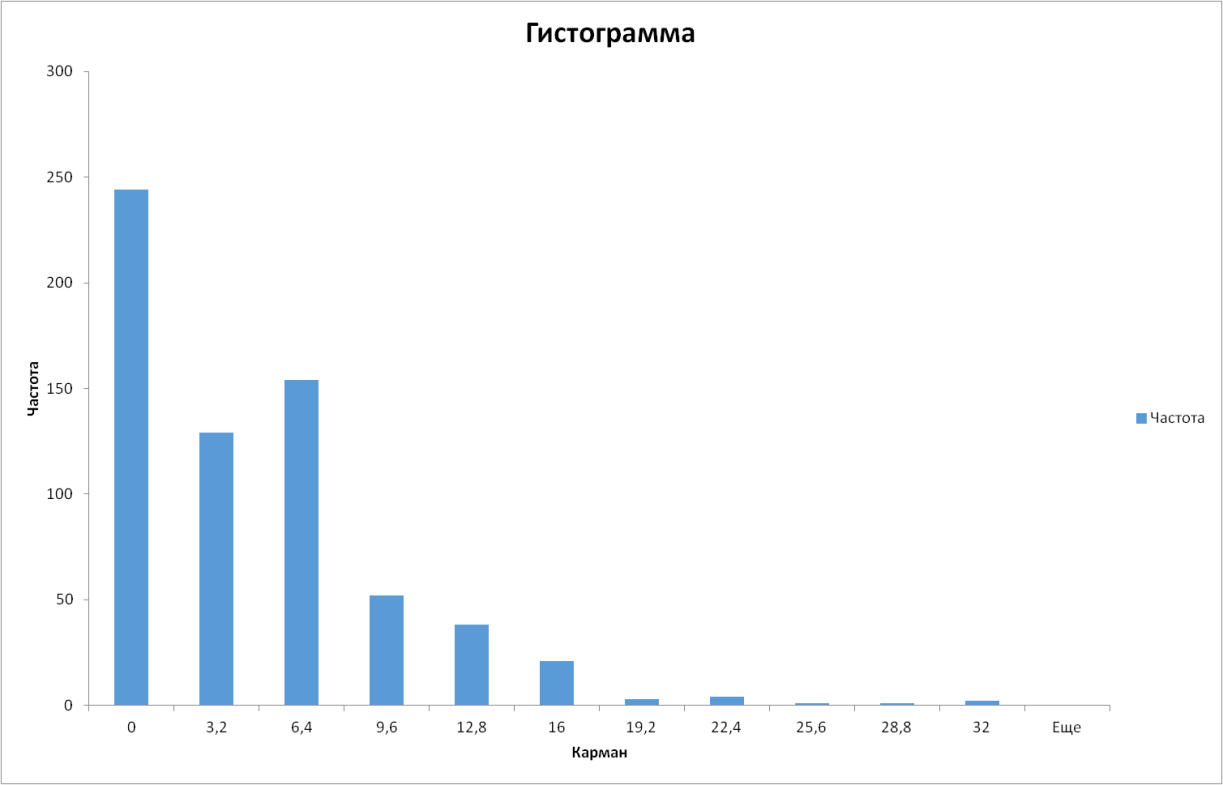
# 2 Описание исходных данных

Был выбран массив данных «Student Performance Data Set». Эти данные относятся к успеваемости учащихся в средних школах португальских школ. Данные включают в себя оценки учащихся, демографические, социальные и школьные характеристики, они были собраны с помощью школьных отчетов и анкет. Зависимые переменные: годовые оценки школьников (0-20). Независимые переменные: количество пропусков в школе (0-32).

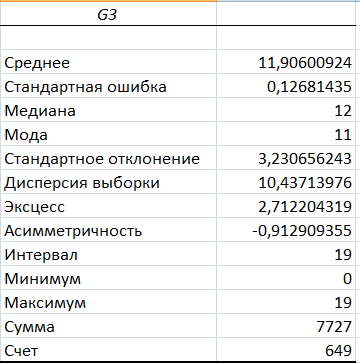
# 3 Результаты дескриптивного анализа



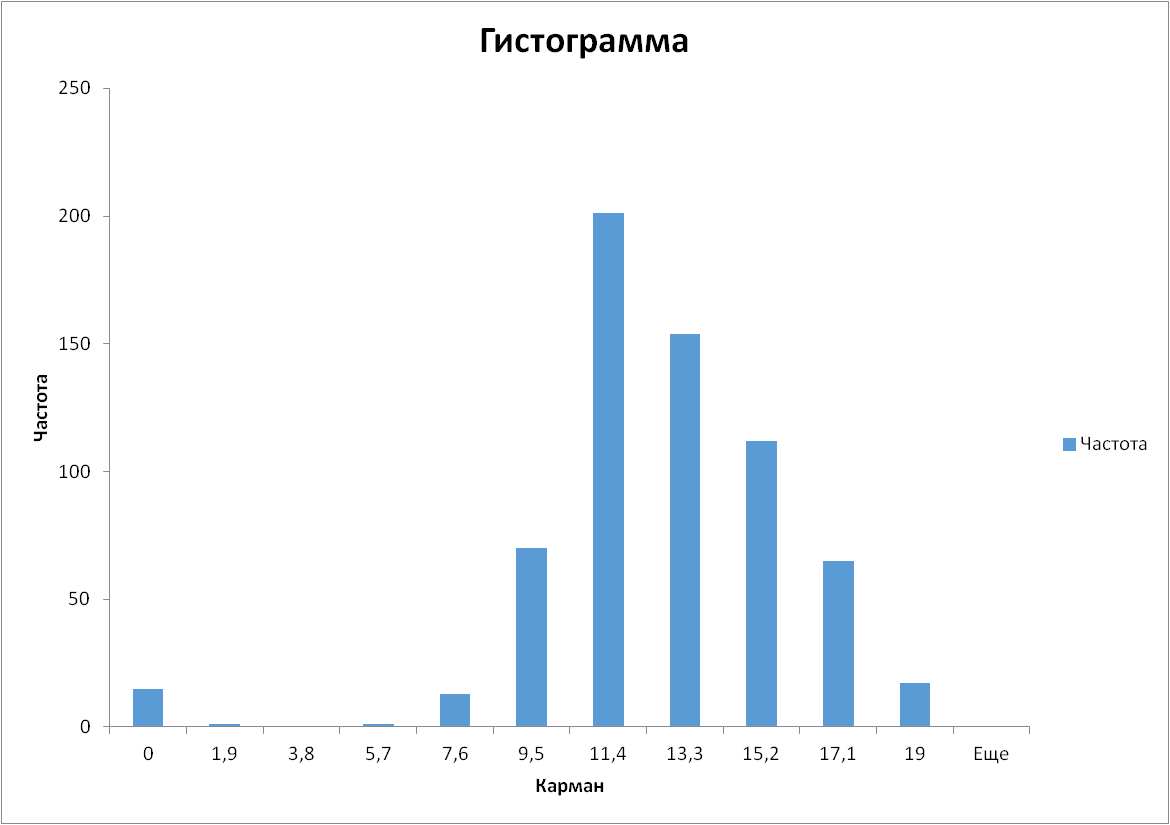
Среднее количество пропусков – 3,659. Максимальное количество пропусков – 32. Исходя из значения моды, можно утверждать, что большинство учеников не пропускают занятия.



Коэффициент эксцесса – положительный, значит, распределение является островершинным. Положительный знак коэффициента асимметрии говорит о наличии правосторонней асимметрии. Выборка не похожа на нормальное распределение.



Средняя годовая оценка школьников – 11,906. Максимальный балл – 19. Исходя из значения моды, можно утверждать, что большинство учеников имеет годовую оценку – 11.

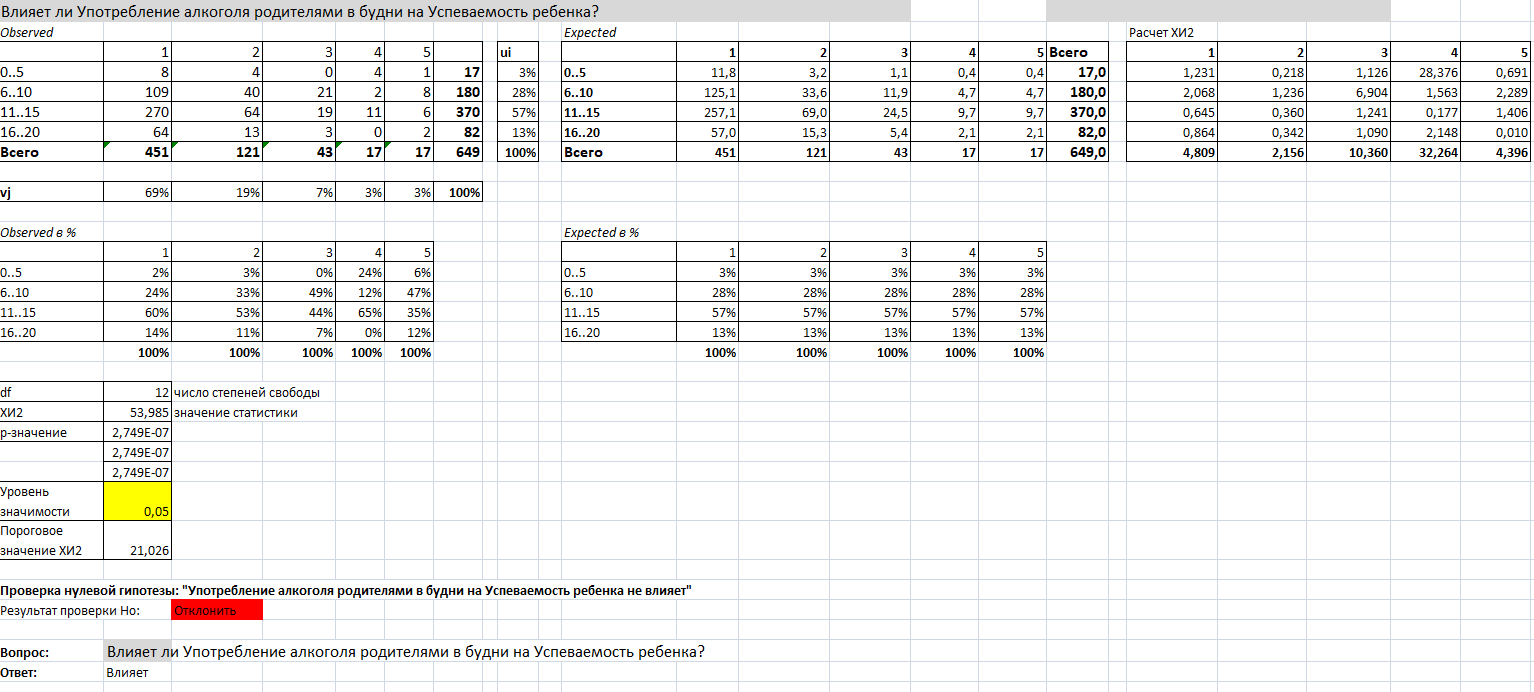


Коэффициент эксцесса – положительный, значит, распределение является островершинным. Отрицательный знак коэффициента асимметрии говорит о наличии левосторонней асимметрии. Эта гистограмма похожа на нормальное распределение.

# 4 Анализ результатов таблиц сопряженности

В качестве категориальных переменных были выбраны потребление алкоголя родителями в будние дни (1-5), а годовые оценки учеников были разделены на 4 категории (0-5, 6-10, 11-15, 16-20). Была сформулирована нулевая гипотеза H0: «Употребление алкоголя родителями в будни на Успеваемость ребенка не влияет».

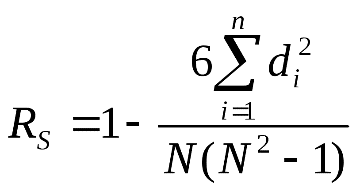
Был проведен анализ с помощью таблиц сопряженности. В результате нулевая гипотеза была отклонена. Отсюда следует, что употребление алкоголя родителями в будни влияет на успеваемость ребенка.



# 5 Результаты корреляционного анализа

Так как рассматриваемые выборки (по годовым оценкам школьников и по количеству пропусков в школе) были далеки от нормального распределения, то был проведен ранговый корреляционный анализ.

По формуле для расчета коэффициента Спирмена вычислили корелляцию между двумя выбранными признаками.



Корреляция равна -0,11557, значит, между выбранными признаками отсутствует линейная зависимость. Коэффициент корреляции отрицательный, следовательно, имеет место обратная зависимость. Связь между признаками – слабая.

Проверим значимость коэффициента корреляции Спирмена при помощи t-критерия, рассчитанного по следующей формуле:

https://medstatistic.ru/pictures/ttest.png

t-статистика = 2,5959

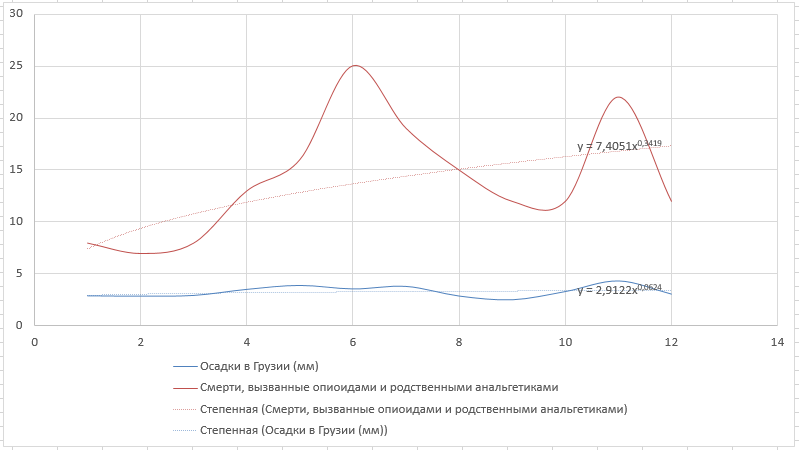
Табличное значение для степеней свободы 647 и значимости 0,05 равно 1,9636. Расчетное значение больше табличного, следовательно, корреляция значима.

**6 Пример и интерпретация ложной корреляции**

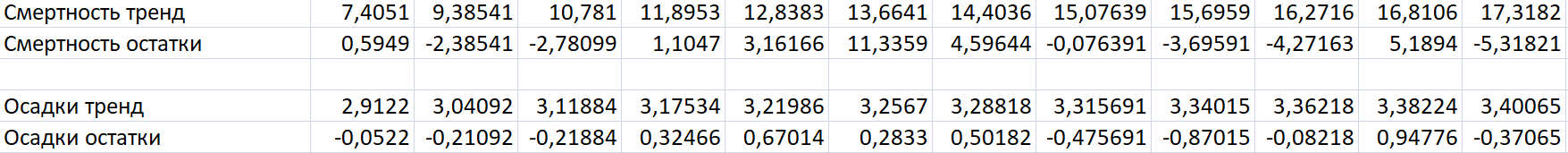
На сайте http://www.tylervigen.com/spurious-correlations выбраны два динамических ряда, построены графики, проведена оценка ложной корреляций между переменными.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Осадки в Грузии (мм) | 2,86 | 2,83 | 2,9 | 3,5 | 3,89 | 3,54 | 3,79 | 2,84 | 2,47 | 3,28 | 4,33 | 3,03 |
| Смерти, вызванные опиоидами и родственными анальгетиками | 8 | 7 | 8 | 13 | 16 | 25 | 19 | 15 | 12 | 12 | 22 | 12 |

|  |
| --- |
| Коэффициент корреляции |
| 0,722099 |



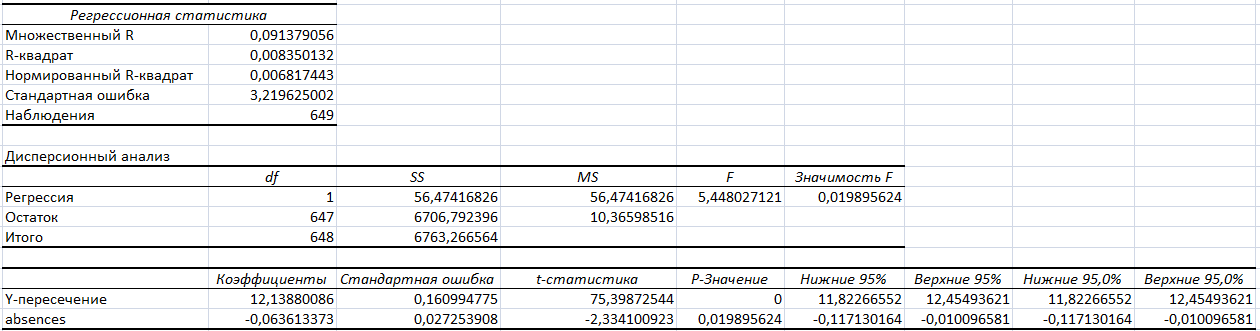
Рассчитаем взаимную корреляцию по остаткам. В данном случае **не удалось избавиться от ложной корреляции** с помощью тренда.

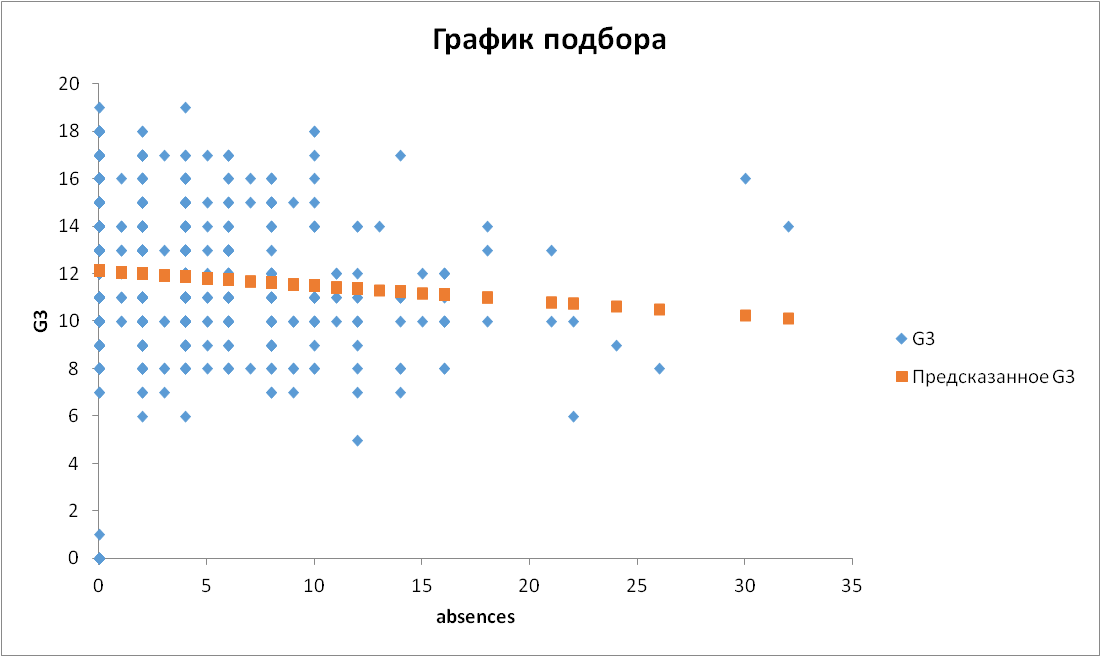


|  |
| --- |
| Коэффициент корреляции |
| 0,69566 |

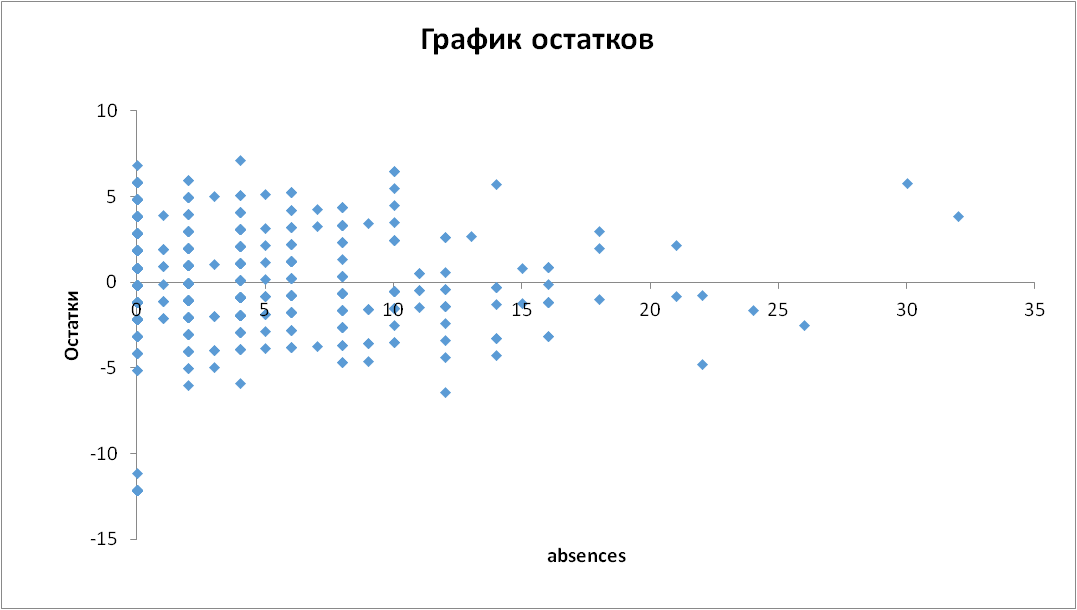
# 7 Регрессионная модель. Уравнение регрессии, график остатков, гистограмма остатков. Оценка качества построенной модели

Построим линейную регрессионную модель (для признаков с низким коэффициентом корреляции). Зависимость годовых оценок школьников от количества пропусков.



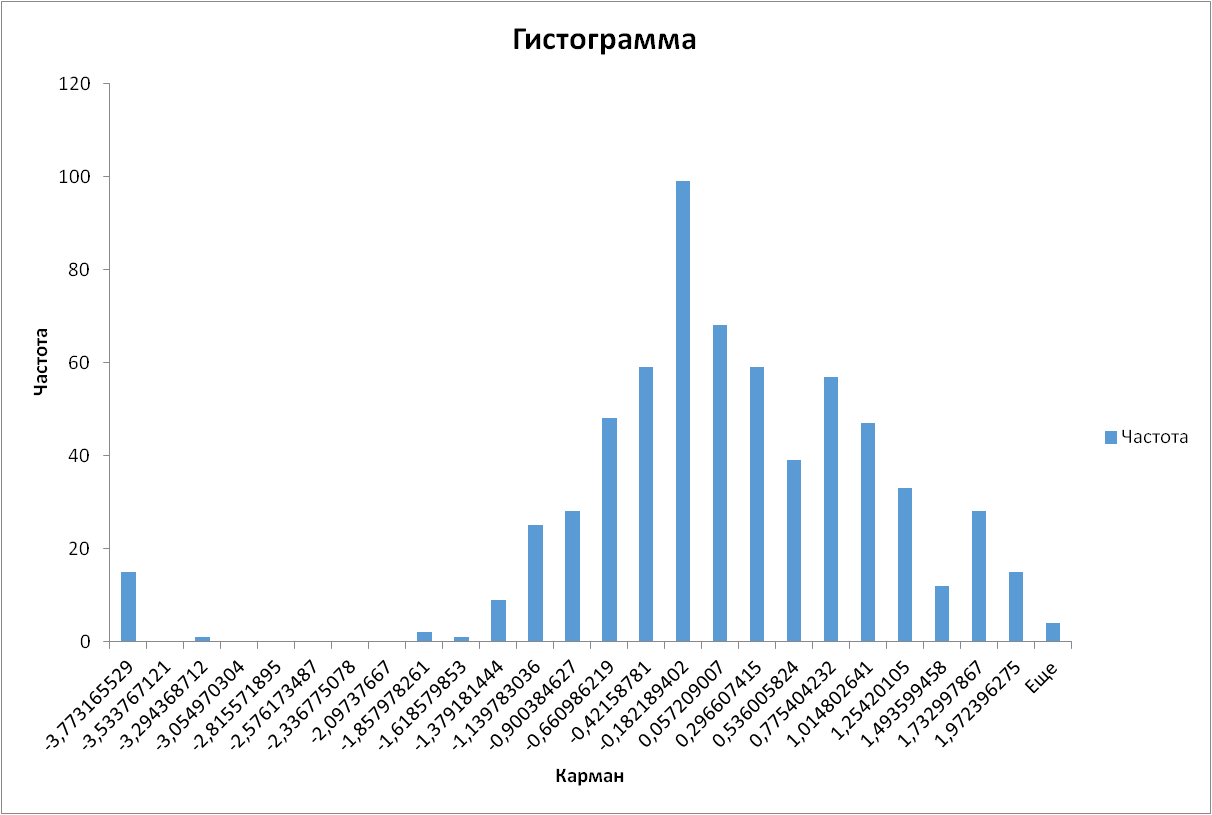


Выведем график остатков.



Оценим постоянство среднего и дисперсии. Среднее остатков – постоянное и равно нулю. Дисперсия (разброс) непостоянна и наблюдаются выбросы.

Построим гистограмму стандартизированных остатков.



Эта гистограмма похожа на нормальное распределение с левосторонней асимметрией. Было посчитано среднее значение стандартизированных остатков -1,89799E-16. Можно считать его нулем.

Запишем уравнение регрессии. Y-пересечение и Пересечение absences – это коэффициенты уравнения регрессии. Таким образом, уравнение будет иметь вид: .

Коэффициент 12,139 показывает, каким будет Y, если все переменные в рассматриваемой модели будут равны 0. То есть на значение анализируемого параметра влияют и другие факторы, не описанные в модели.

Коэффициент -0,064 показывает весомость переменной Х на Y. То есть количество прогулов в пределах данной модели влияет на годовую оценку с весом -0,064 (это небольшая степень влияния). Отрицательный знак указывает на отрицательное влияние: чем больше прогулов, тем меньше годовая оценка.

Оценим значимость уравнения регрессии.

Коэффициент значимости F = 0,019895624. Он меньше, чем 0,05, следовательно, можно считать уравнение регрессии значимым.

Это даёт нам возможность составить прогноз. Допустим, ученик совершил 40 прогулов, и нам нужно примерно представить, каким будет его годовая оценка. Подставляем в уравнение регрессии данное значение и получаем результат . Следовательно, у ученика, совершившего 40 прогулов, будет годовая оценка 9,579.

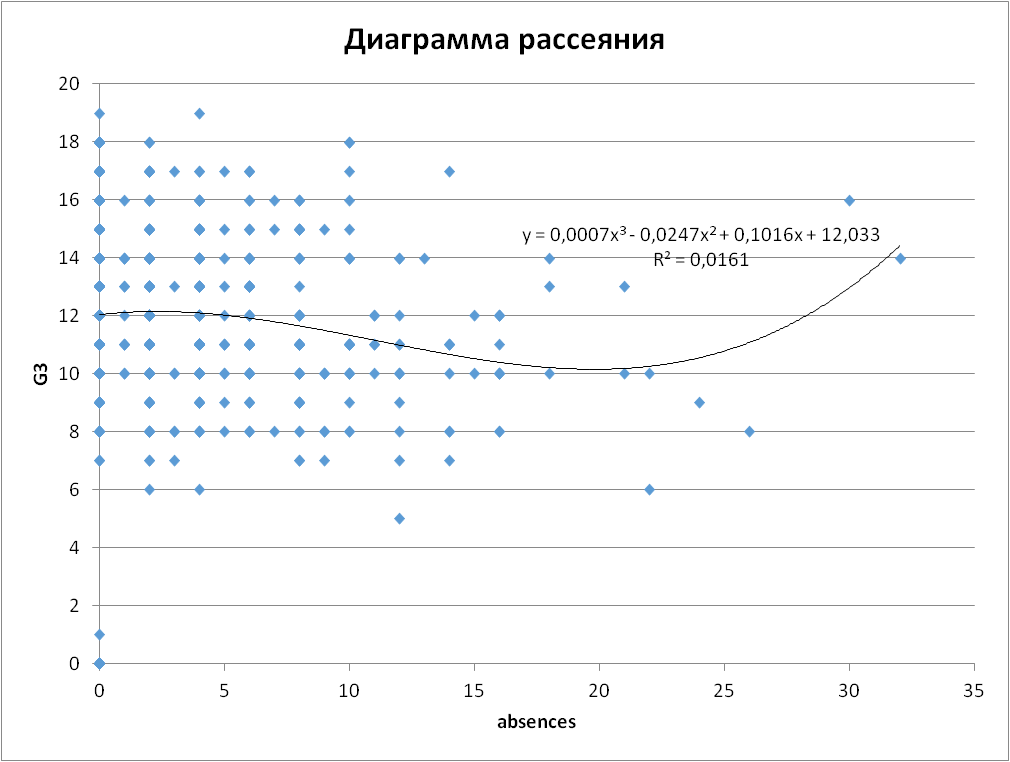
Оценим качество построенной модели. График остатков имеет постоянные дисперсию (разброс) и среднее, наблюдаются выбросы. Гистограмма остатков характеризуется левосторонней асимметрией, распределение является островершинным, но не относится к нормальному.

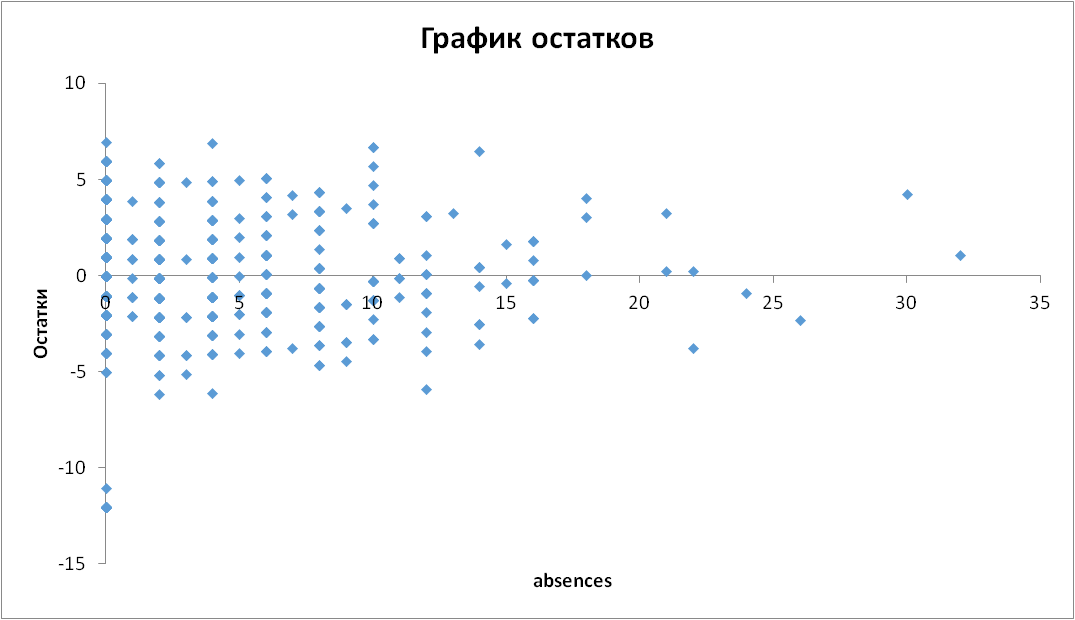
R2 – это коэффициент детерминации. Коэффициент детерминации рассматривают, как правило, в качестве основного показателя, отражающего меру качества регрессионной модели, описывающей связь между зависимой и независимыми переменными модели. В данном примере коэффициент детерминации = 0,00835 (0,835%), т. е. расчетные параметры модели объясняют зависимость между рассматриваемыми переменными на 0,835%. Чем выше значение коэффициента детерминации, тем выбранная модель считается более применимой для конкретной задачи. Считается, что она корректно описывает реальную ситуацию при значении коэффициента детерминации выше 0,8. Следовательно, при коэффициенте 0,00835 смысла в построенной модели нет.

# 8 Сравнение регрессионных моделей

Построим диаграмму рассеяния для тех же переменных, для которых была построена линейная регрессионная модель. Добавим линию тренда на диаграмму рассеяния. Был выбран полиномиальный тренд (степень полинома – 3). Отобразим уравнение и коэффициент детерминации на графике.

Были посчитаны новые значения зависимой переменной (годовая оценка школьников) по полученному уравнению. Далее посчитаны остатки в новой модели (). Затем построен график остатков.

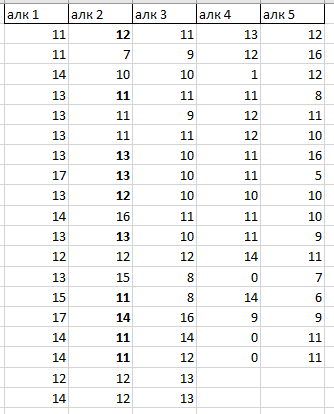




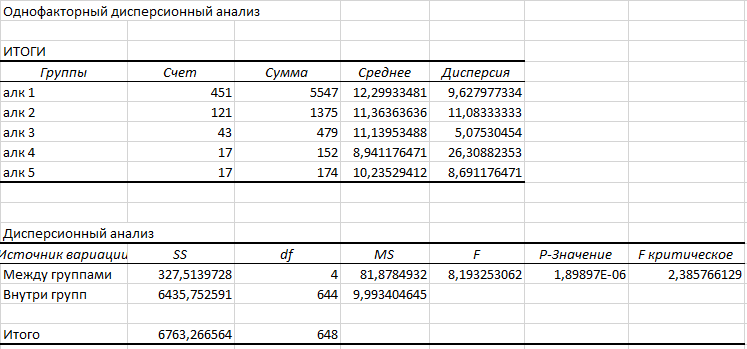
Сравним две регрессионные модели: линейную (из раздела 7) и полиномиальную. Обе модели имеют графики остатков с постоянными средним, однако наблюдаются выбросы. Коэффициент детерминации во второй модели получился выше, чем в первой модели (0,02 и 0,00835 соответственно). Значит, построенные модели получились неудачными.

# 9 Формулировка гипотез. Описание градаций фактора. Результаты дисперсионного анализа

Была сформулирована начальная гипотеза: «Употребление алкоголя родителями в будние дни не влияет на годовые оценки их детей». Для независимой переменной были определены градации (уровни пьянства – 1, 2, 3, 4 и 5 – от очень низкого потребления алкоголя к очень высокому). В качестве зависимой переменной были взяты годовые оценки школьников (от 0 до 20).



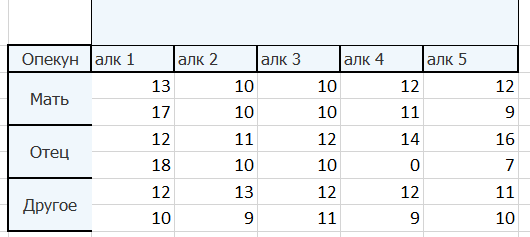
Проведен однофакторный дисперсионный анализ.



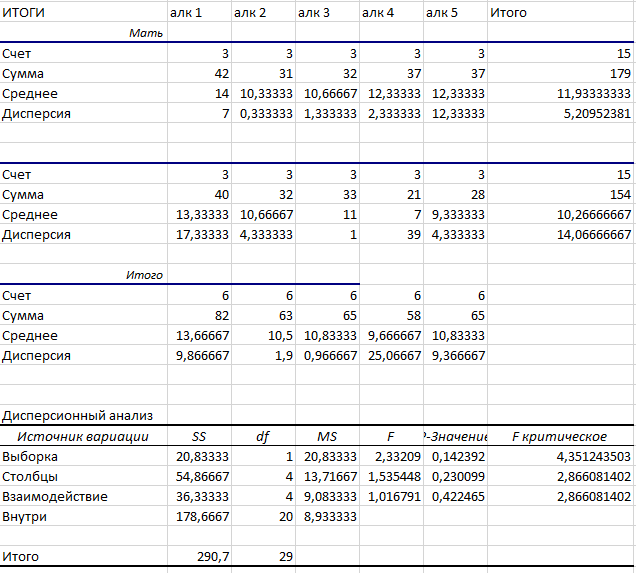
Значение F должно быть не меньше F-критического. В нашем случае F = 8,193 > F-критическое = 2,385, следовательно нулевая гипотеза о независимости отвергается. Значит, уровень употребления алкоголя родителями в будние дни влияет на успеваемость детей.

# 10 Двухфакторный дисперсионный анализ

Для проведения двухфакторного дисперсионного анализа были сформулированы гипотезы: «Употребление алкоголя родителями в будние дни не влияет на годовые оценки школьников», «Тип опекуна не влияет на годовые оценки школьников», «Употребление алкоголя родителями в будние дни и тип опекуна не дают эффекта взаимодействия на годовые оценки школьников». Для независимых переменных были определены градации (уровни пьянства – 1, 2, 3, 4 и 5 – от очень низкого потребления алкоголя к очень высокому; типы опекунов – мать, отец, другой). Значения зависимой переменной были случайным образом взяты из массива значений годовых оценок в соответствии с факторами в количестве 2 штук.



Провели двухфакторный дисперсионный анализ.



Значение F фактора А должно быть не меньше F-критического. В нашем случае F = 2,33 < F-критическое = , следовательно, нулевая гипотеза о независимости принимается. Значит, тип опекуна не влияет на годовые оценки школьников.

Значение F фактора В должно быть не меньше F-критического. В нашем случае F = 1,53 < F-критическое = , значит, нулевая гипотеза о независимости принимается. Значит, потребления алкоголя родителями в будние дни не влияет на годовые оценки школьников.

Значение F фактора взаимодействия должно быть не меньше F-критического. В нашем случае F = 1,01 < F-критическое = , значит, гипотеза об отсутствии эффекта взаимодействия принимается.

Такие результаты нельзя считать верными, потому что 2 значения оценок были выбраны случайным образом из большого числа значений. Поэтому в данном анализе получилось, что уровень употребления алкоголя родителями в будние дни не влияет на успеваемость детей, хотя в однофакторном дисперсионном анализе эта зависимость была выявлена.