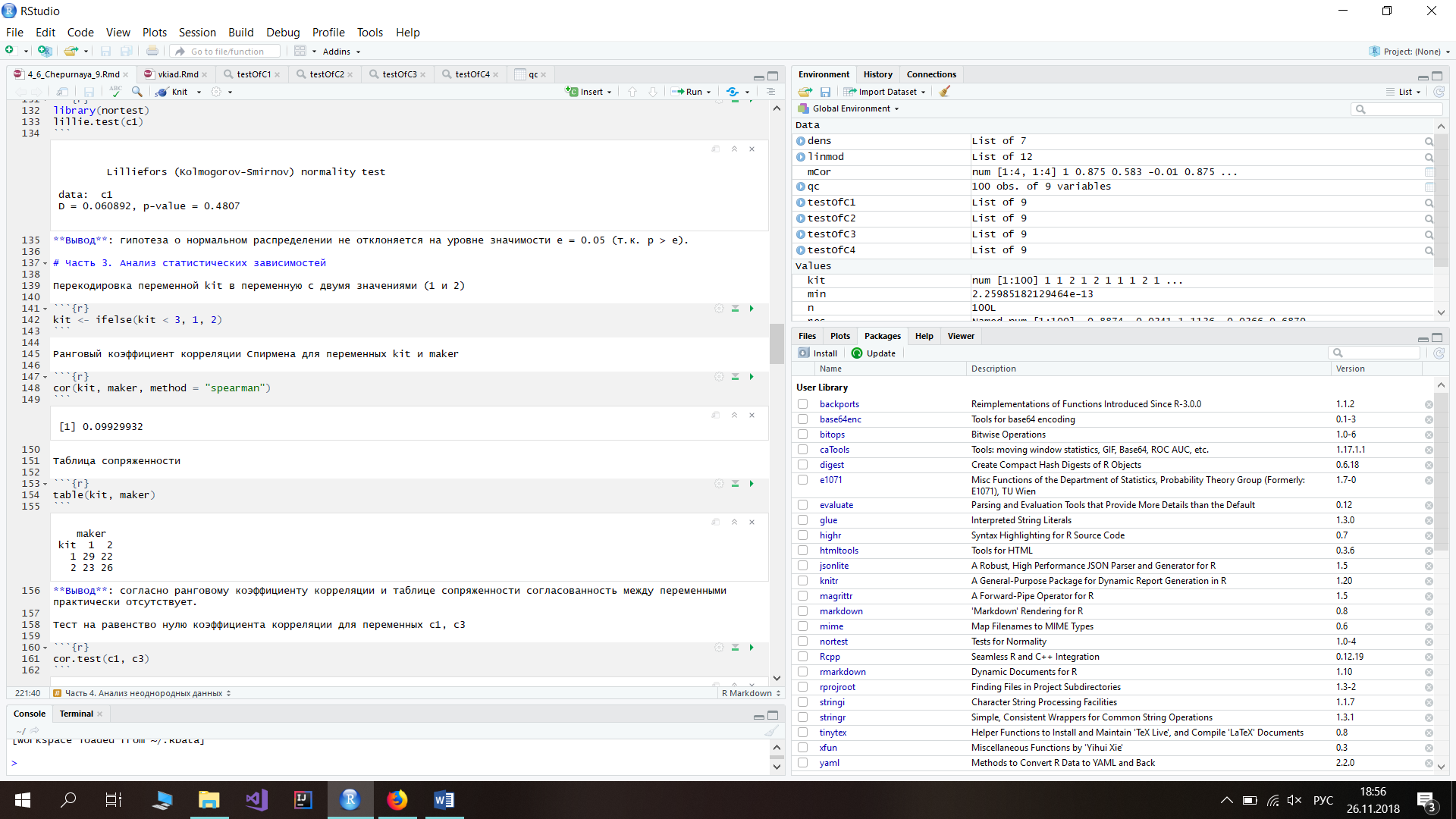
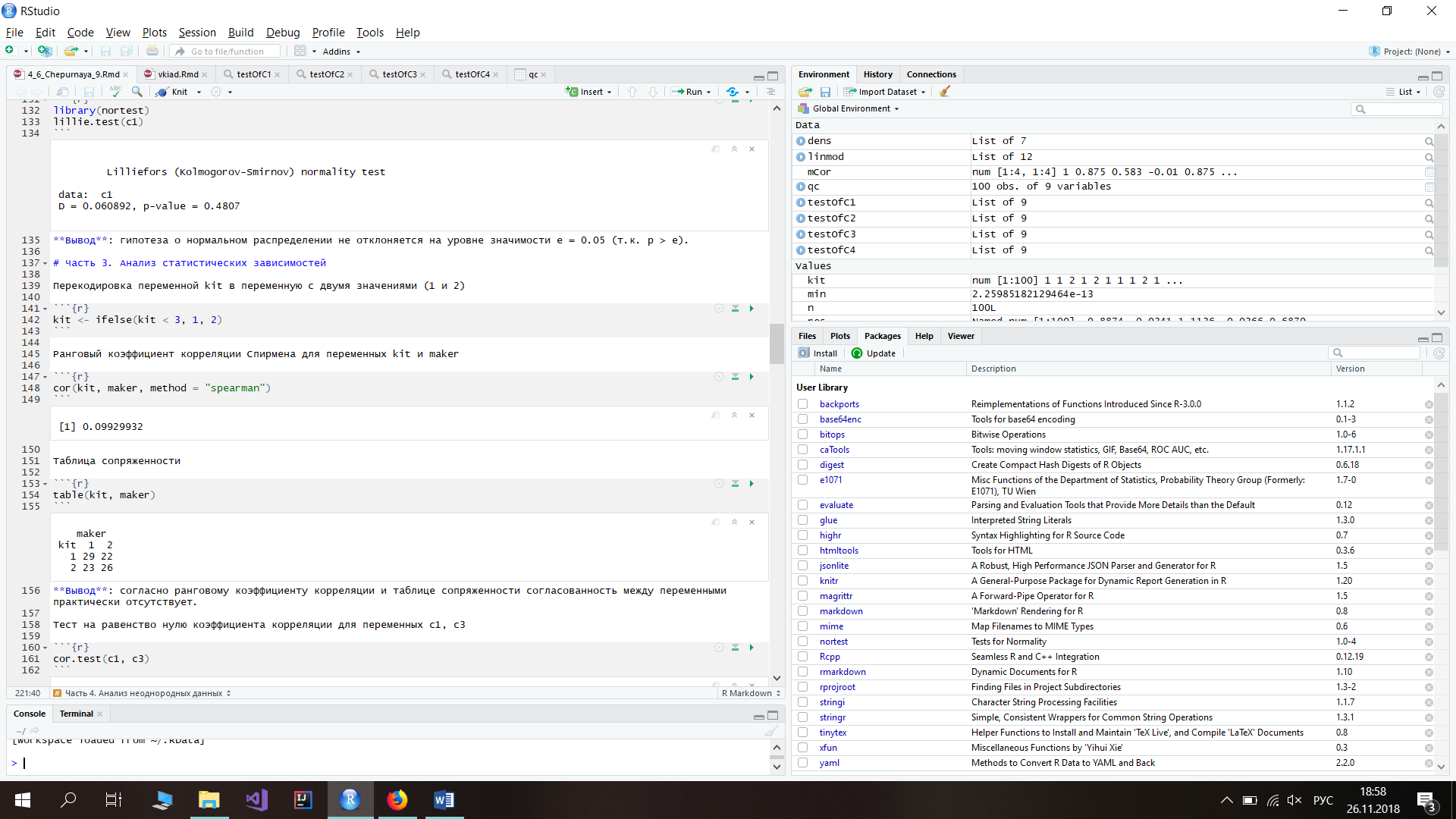
***Задание 1***



Ifelse работает аналогично С++

А именно «если kit<3, то перекодируем его в 1, иначе – в 2»

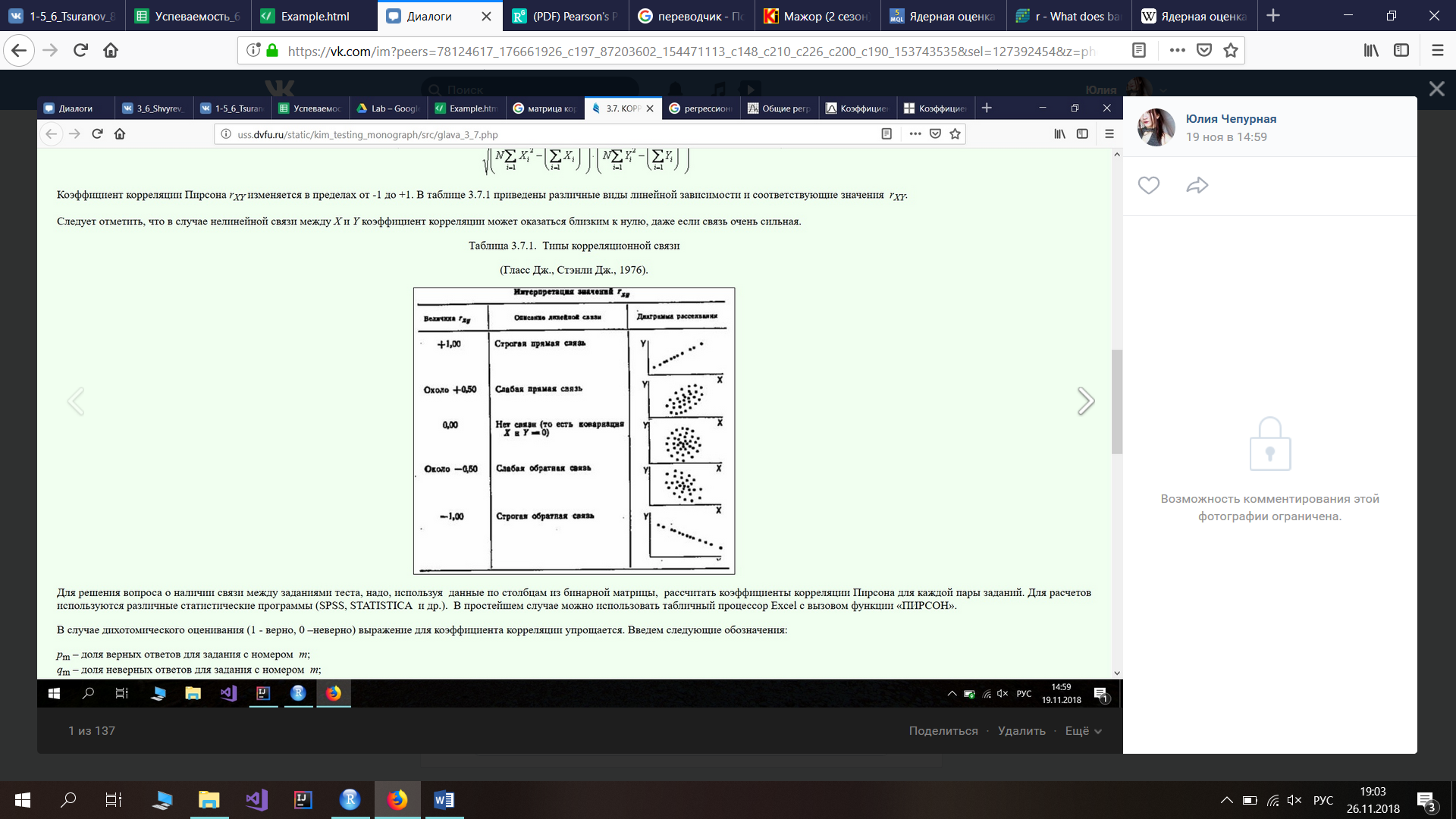
***Задание 2***



Что такое коэффициент корреляции?

Это коэффициент, который показывает взаимосвязь между двумя переменными

Он изменяется от -1 до 1

если показатель около 0, то связи между переменными нет

если она около 0.5, то связь прямая, но слабая (т.е. чем больше одно, тем больше другое и наоборот, но слабо)

если оно около 1, то связь сильная прямая.  
если 1, то чаще всего это одна и та же переменная (у нас в таблице это видно будет)

если она около -0.5, то связь обратная, но слабая (т.е. чем больше одно, тем меньше другое и наоборот, но слабо)

если оно около -1, то связь сильная обратная.

У нас в задании коэффициент 0.0992…, он почти равен нулю, следовательно, мы уже можем сказать, что связи между переменными kit u maker практически нет. В этом мы убедимся в следующем задании.

***Задание 3***

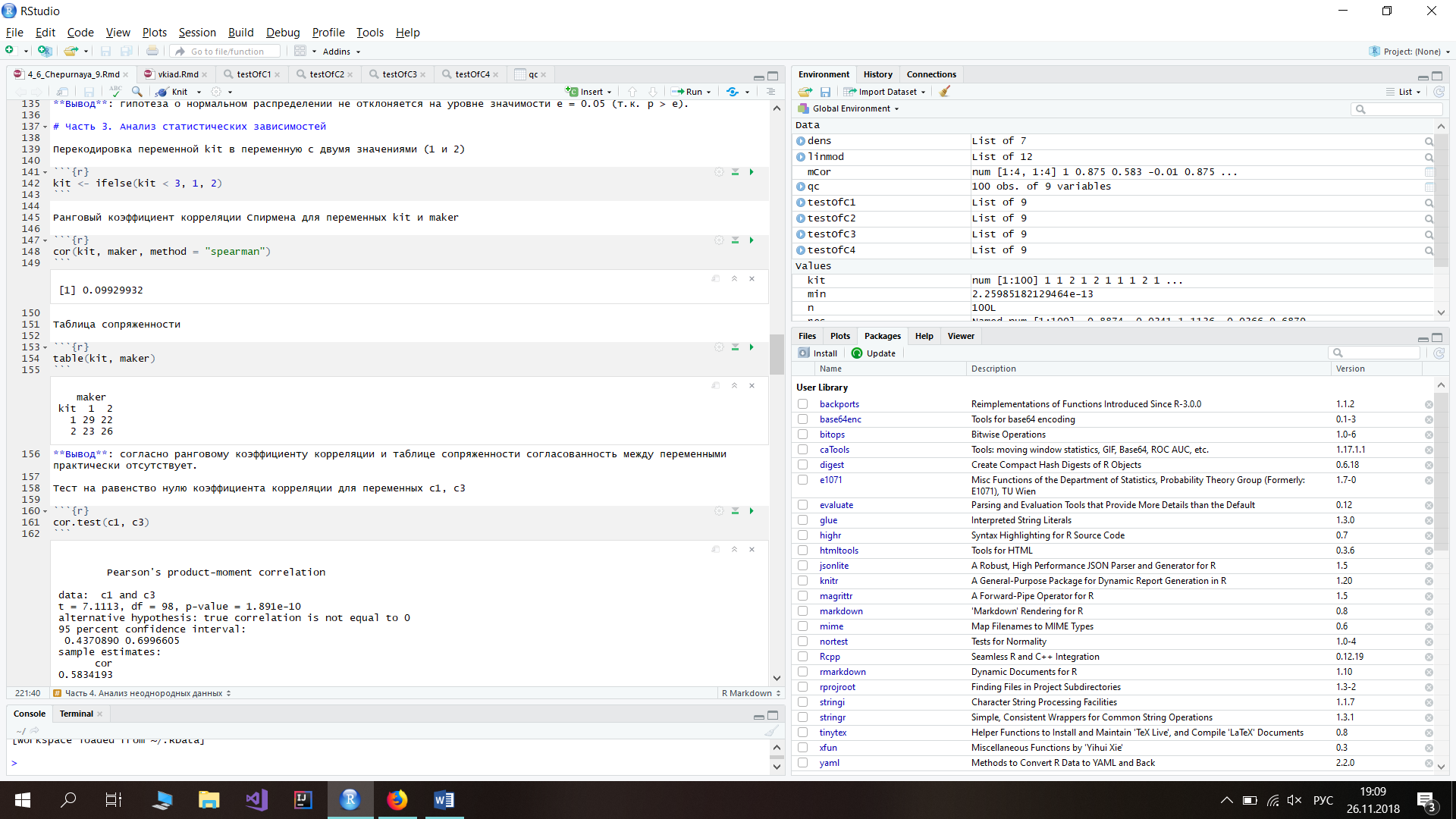


Таблица сопряженности показывает, сколько у нас объектов с теми или иными значениями переменных kit u maker.

Видим, что с kit=1 u maker=1 29 объектов

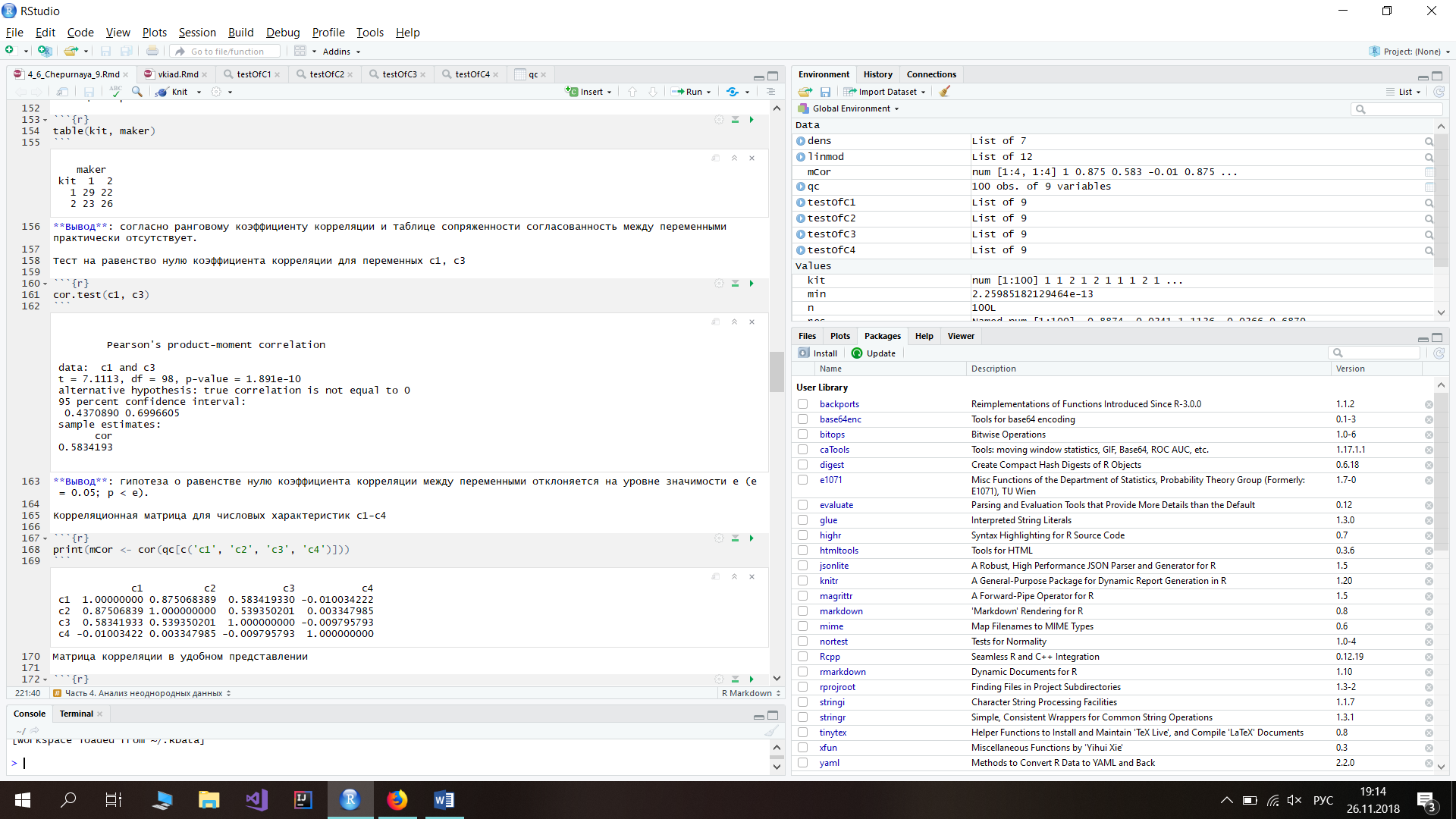
kit=1 u maker=2 22 объектов

kit=2 u maker=1 23 объектов

kit=2 u maker=2 26 объектов

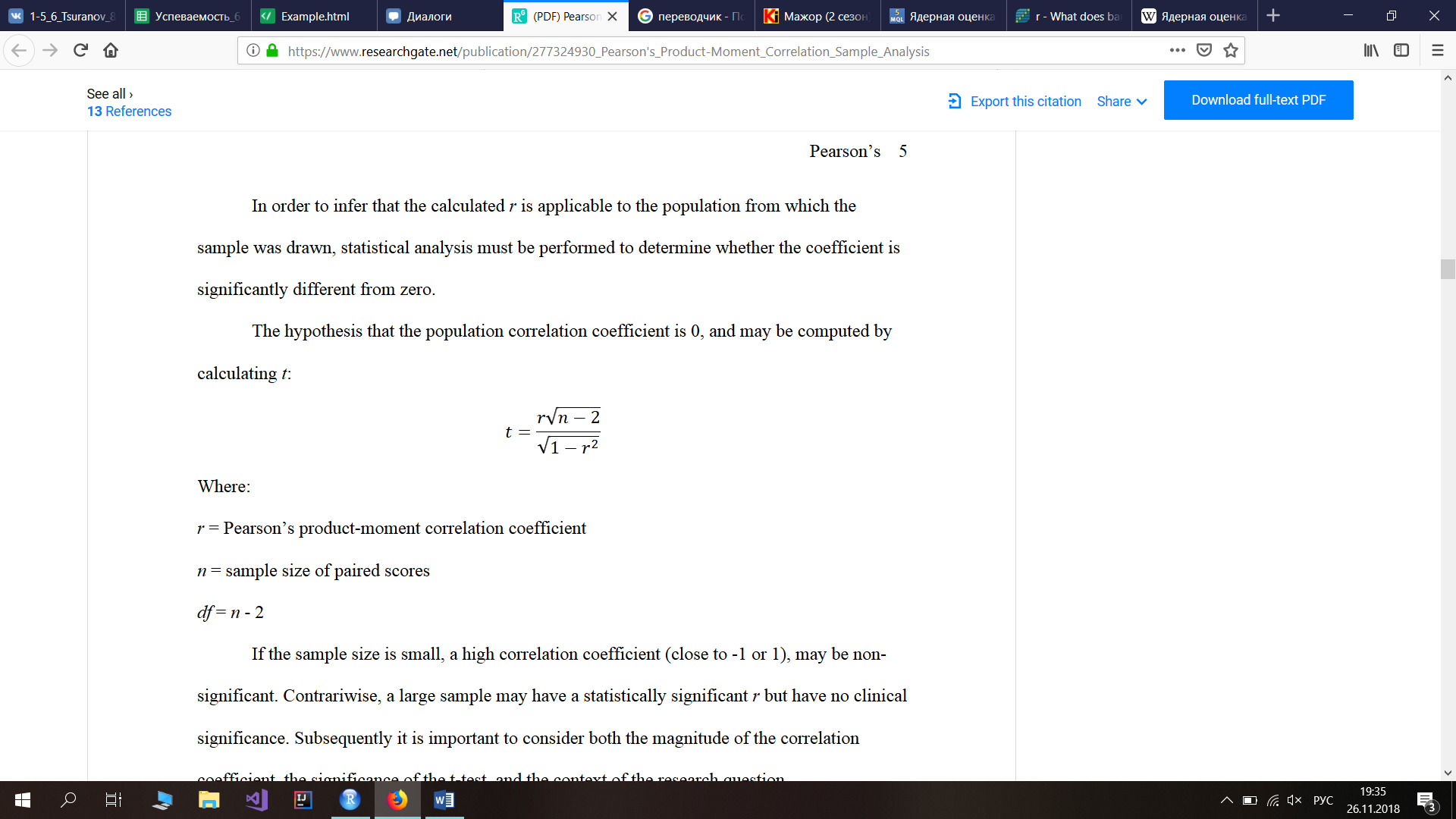
тут видно, что независимо от того, какие у нас kit u maker, кол-во объектов почти одинаковое. Поэтому это еще раз подтверждает коэффициент корреляции. Поэтому и пишем такой вывод.

***Задание 4***



Данный тест отвечает нам на вопрос: «Независимы ли переменные с1 и с3?» или «Равен ли коэффициент корреляции нулю?»

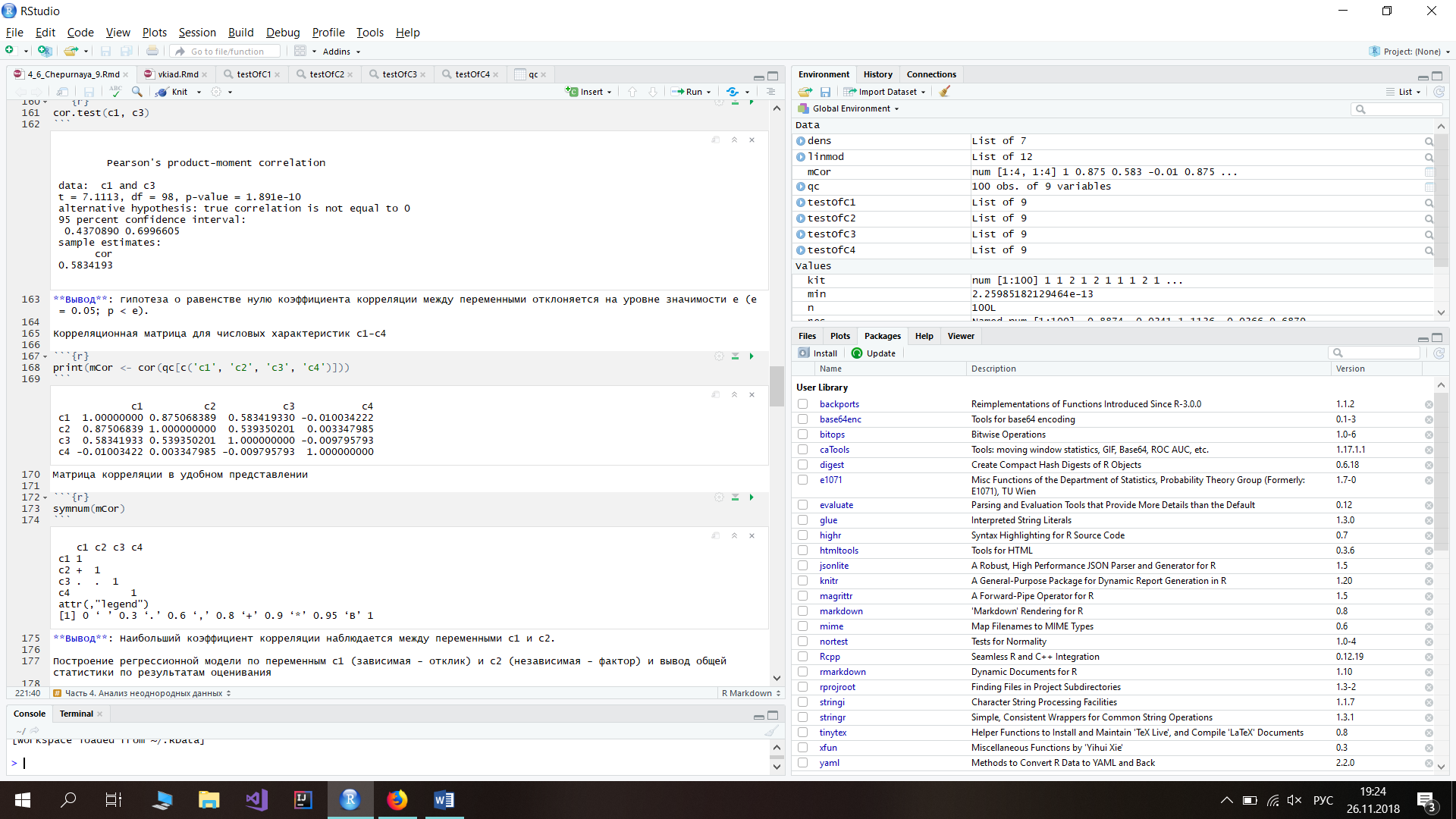
t – t-статистика. Рассчитывается по формуле



df(степень свободы) = n-2 (n у нас равно 100)

p-value – оно отвечает нам на наш вопрос. Так как оно <0.05, то наше предположение отклоняется, оно неверно. Т.е. коэффициент между с1 и с3 не равен нулю или около нуля. Сл-но между ними есть какая то связь. И в самом низу, где 0.583… - это наш коэффициент корреляции. Из предыдущего описания коэффициента, мы можем сказать, что связь между с1 и с3 –прямая и чуть сильнее, чем просто слабая.

***Задание 5***



Корреляционная матрица – это матрица из коэффициентов корреляции. На диагоналях стоят единички (это то, о чем я говорила в задании 2). На пересечениях переменных – их коэффициенты корреляции.

Матрица корреляции в удобном виде

Числа заменены условными обозначениями.

Обычный пробел – число между 0 и 0.3 (числа по модулю берутся)

Точка (.) – число между 0.3 и 0.6

Запятая (,) – число между 0.6 и 0.8

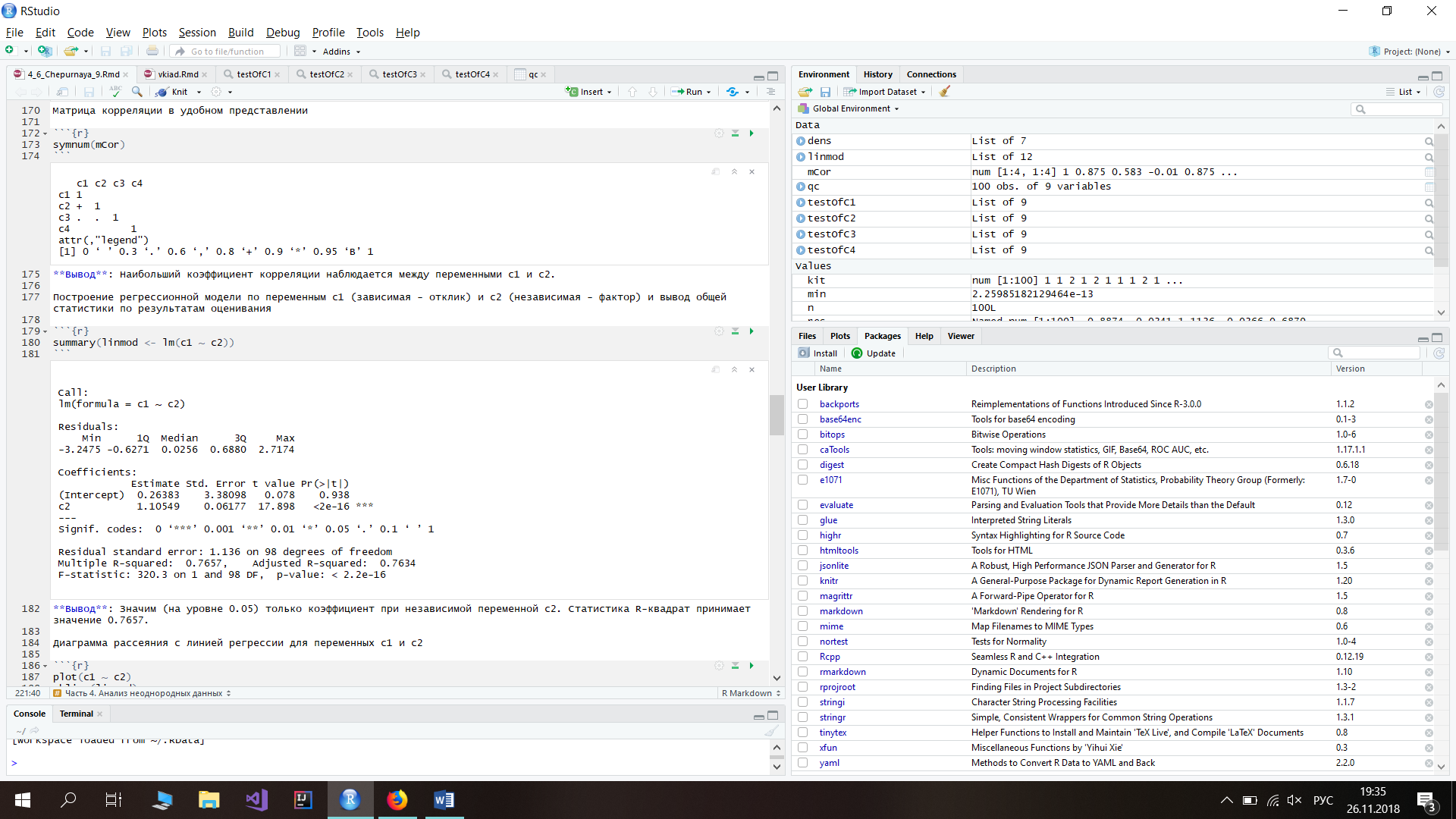
Плюс (+) – число между 0.8 и 0.9

Звезда (\*) – число между 0.9 и 0.95

В – число между 0.95 и 1

Ищем наибольший коэффициент. Видим плюсик, следовательно, между с1 и с2 наибольший коэффициент корреляции.

***Задание 6***



Смотрим столбец Pr (>|t|)

1ая строка отвечает за с1. 0.938>0.05

2ая строка отвечает за с2. 2e-16<0.05

Отсюда следует, что значимы коэффициенты только при с2 (чуть что уточни это у Полины, она разбиралась)

R2  показывает нам, насколько наша модель близка к идеальной модели (нормальному распределению)

Adjusted R-squared — Проблема с R2 в том, что он по любому растет с числом факторов, поэтому высокое значение данного коэффициента может быть обманчивым, когда в модели присутствует множество факторов. Для того, чтобы изъять из *коэффициента корреляции* данное свойство был придуман скорректированный коэффициент детерминации.

F-statistic — Используется для оценки значимости модели регрессии в целом, является соотношением объяснимой дисперсии, к необъяснимой. Если модель линейной регрессии построена удачно, то она объясняет значительную часть дисперсии, оставляя в знаменателе малую часть. Чем больше значение параметра — тем лучше.

***Задание 7***

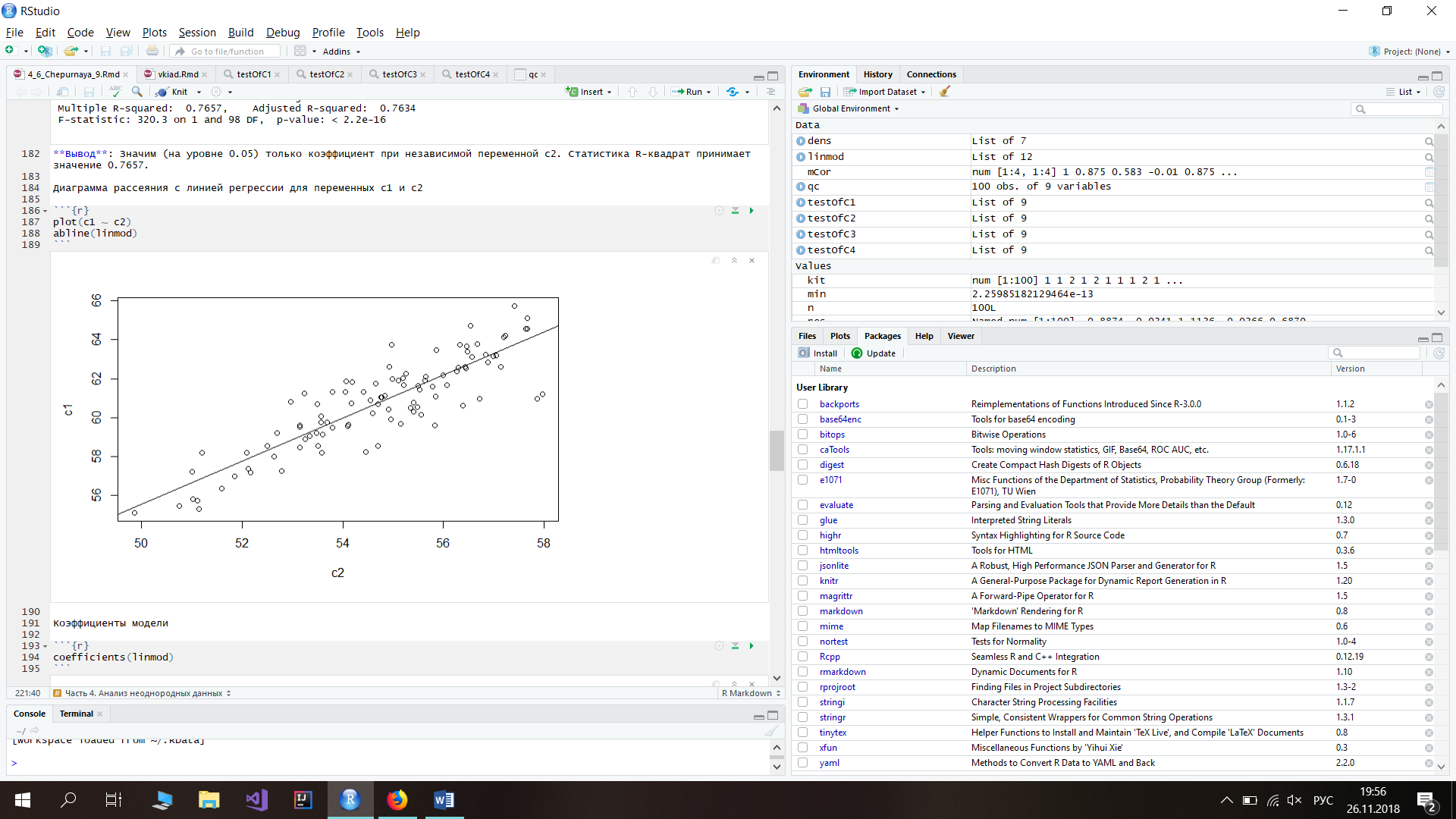


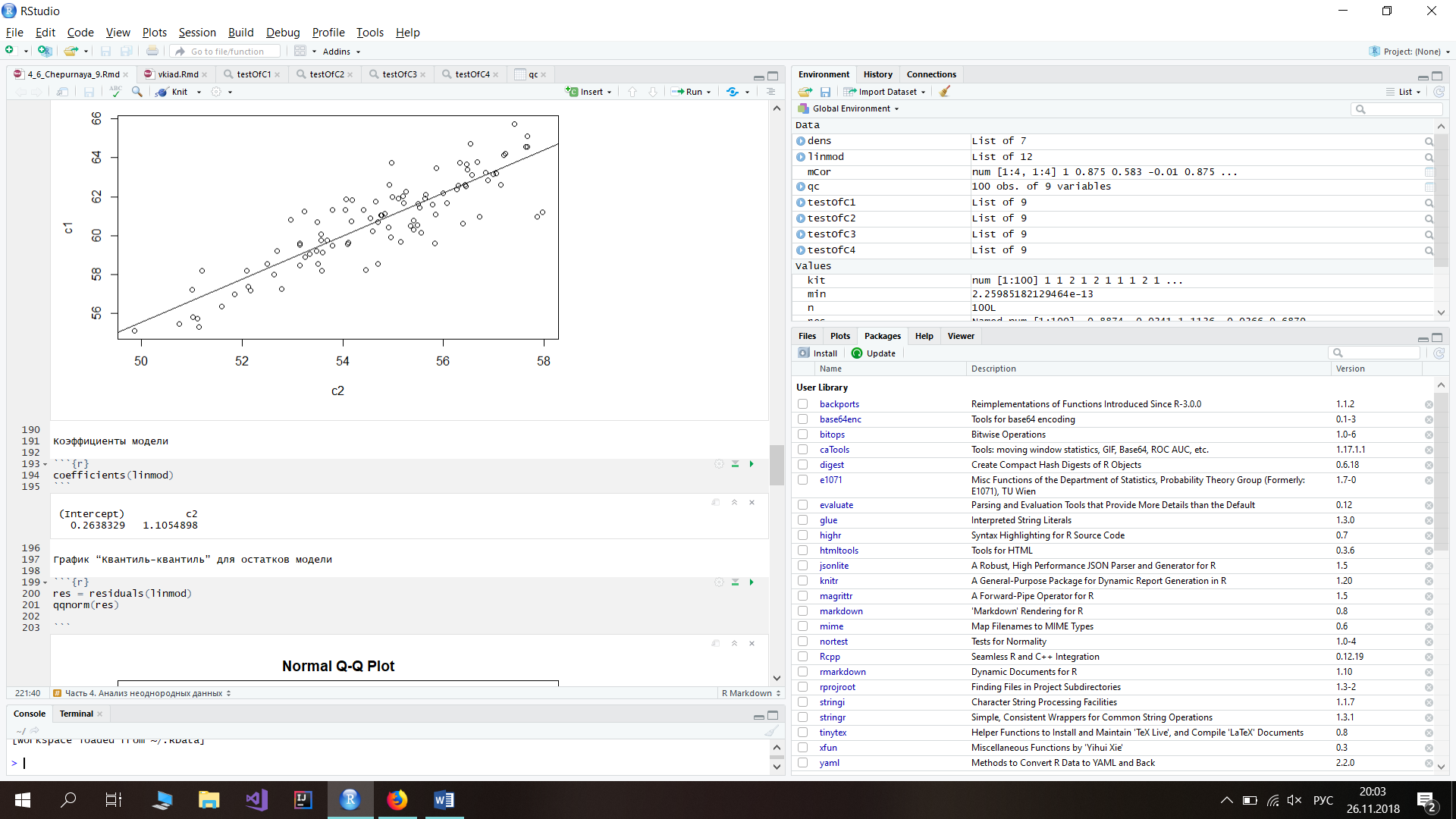
Диаграмма рассеяния— математическая диаграмма, изображающая значения двух переменных в виде точек на декартовой плоскости.

На диаграмме рассеяния каждому наблюдению (или элементарной единице набора данных) соответствует точка, координаты которой (в декартовой системе координат) равны значениям двух каких-то параметров этого наблюдения. Если предполагается, что один из параметров зависит от другого, то обычно значения независимого параметра откладывается по горизонтальной оси, а значения зависимого — по вертикальной. Диаграммы рассеяния используются для демонстрации наличия или отсутствия корреляции между двумя переменными.

Про линию регрессии:

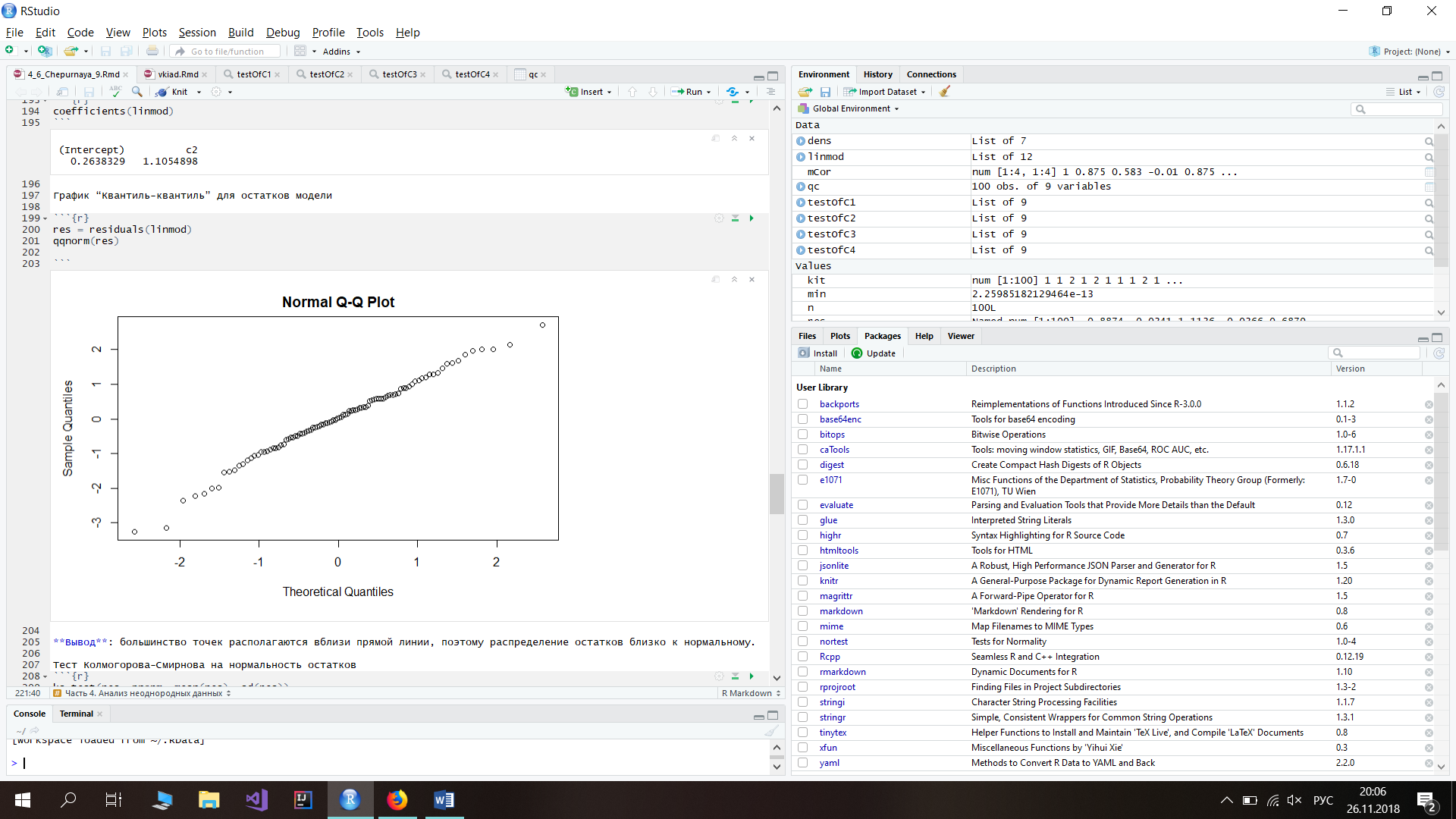
<http://statistica.ru/theory/osnovy-lineynoy-regressii/>

Она задается коэффициентами



В нашем случае, уравнением будет: у = 1.105…\*х+0.263…

***Задание 8***

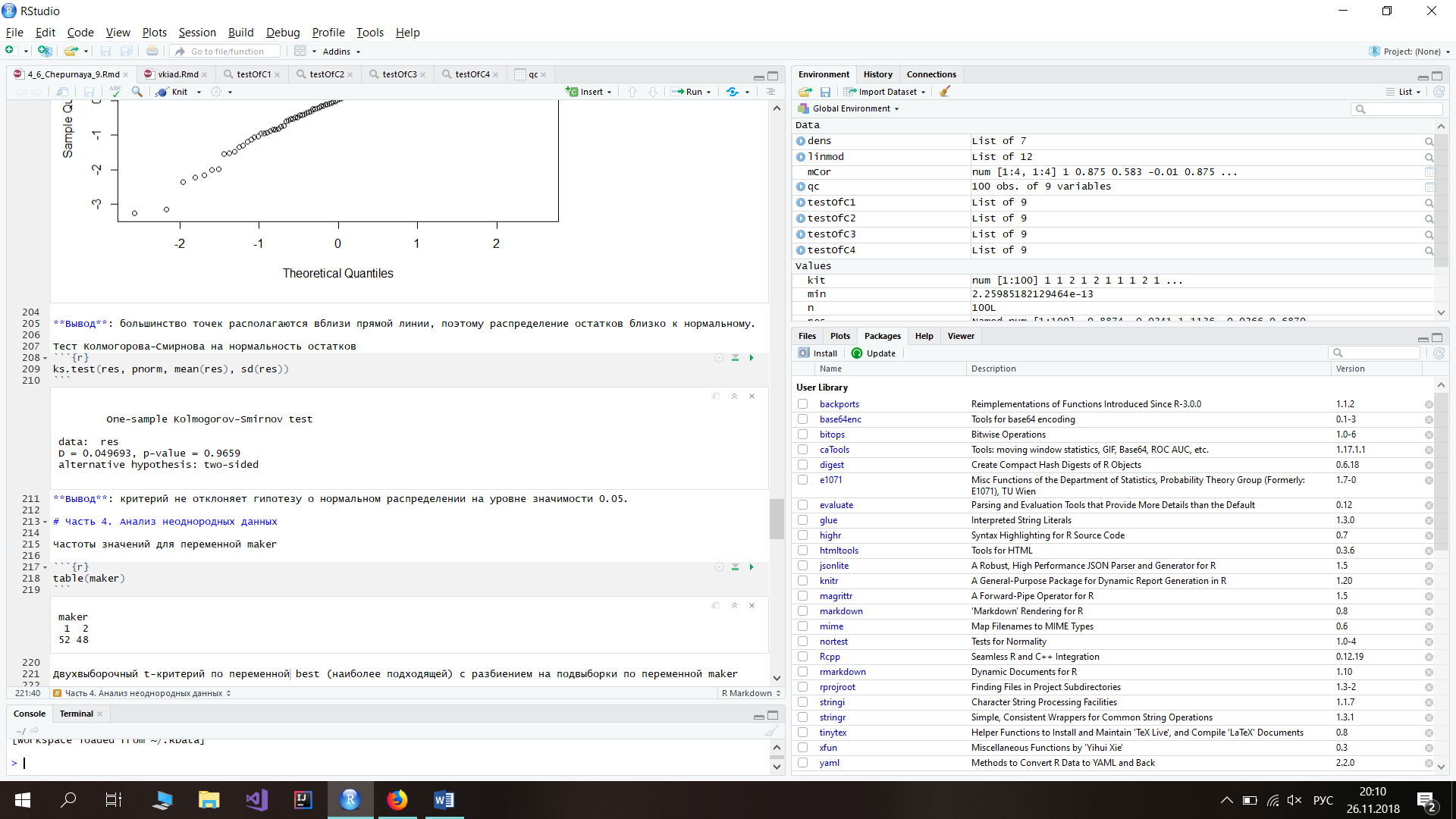


Остатки модели (регрессии) - это разности между наблюдаемыми значениями и предсказанными значениями (идеальными значениями/значениями при нормальном распределении).

Если наблюдаемые значения попадают на прямую линию, то теоретическое распределение хорошо подходит к наблюдаемым данным.

У нас оно так и есть почти, поэтому пишем соответствующий вывод.

***Задание 9***



Этим тестом мы проверяем, что наши остатки действительно нормальные. Получаем р>0.05, следовательно, наше предположение, что наши остатки норм, не отклоняется (т.е. верно). И у нас действительно норм остатки (это мы увидели еще в предыдущем задании).