**Практическая работа №2**  
**Тема: Машинное обучение на языке R с использованием пакета mlr3.**

**ЛИНЕЙНАЯ РЕГРЕССИЯ В R. Импорт данных из файла формата MS Excel**

**Задание:**

**1) изучить импорт данных из файла формата MS Excel**

**2)изучить дополнительные источники.**

Источники:

# [413939036.pdf (hse.ru)](https://www.hse.ru/mirror/pubs/share/413939036.pdf)

# [Язык R для пользователей Excel #3: Загрузка данных из csv, tsv, excel файлов и Google таблиц в R - YouTube](https://www.youtube.com/watch?v=vVFFtgcBp-w)

# [Регрессия. Регрессионный анализ в Excel - YouTube](https://www.youtube.com/watch?v=UrkVOID7ILo)

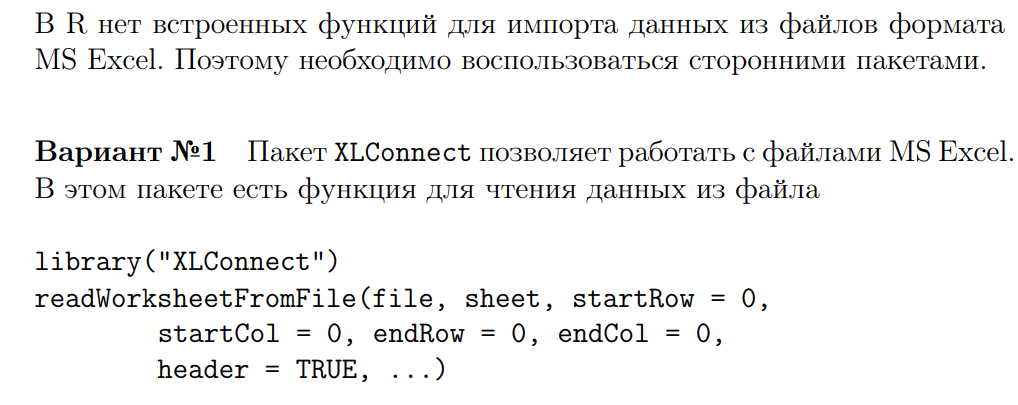
# [Парная регрессия: линейная зависимость - YouTube](https://www.youtube.com/watch?v=3mHoiGm2jVc)

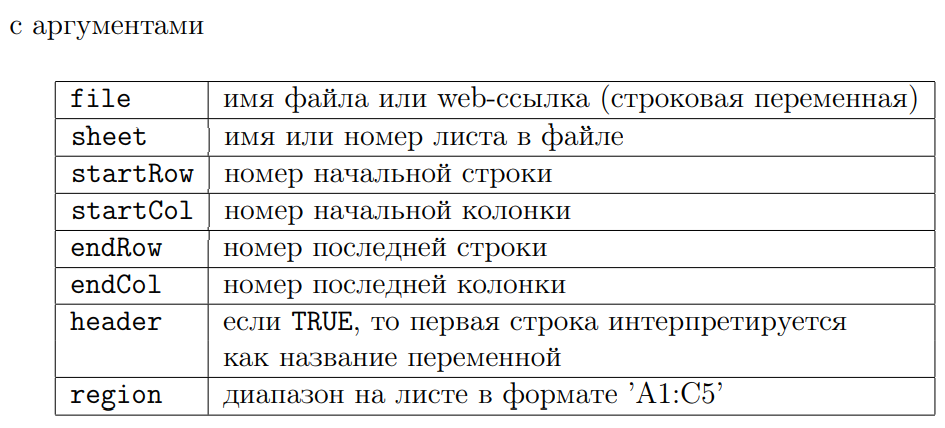
# [Лекция 1. Анализ данных на R в примерах и задачах - YouTube](https://www.youtube.com/watch?v=8mwJ3mEjdIg)

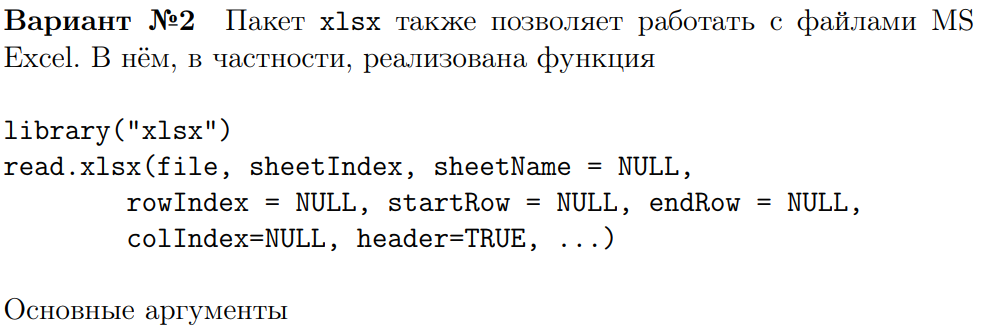
1. [Что такое R — язык программирования и статистическая среда вычислений / What is R language - YouTube](https://www.youtube.com/watch?v=4bd34XDdijY)
2. [R: The R Project for Statistical Computing (r-project.org)](https://www.r-project.org/)
3. [Лекция 1. Анализ данных на R в примерах и задачах - YouTube](https://www.youtube.com/watch?v=8mwJ3mEjdIg)
4. [grishin\_tihov\_R.pdf (unn.ru)](http://www.lib.unn.ru/students/src/grishin_tihov_R.pdf)

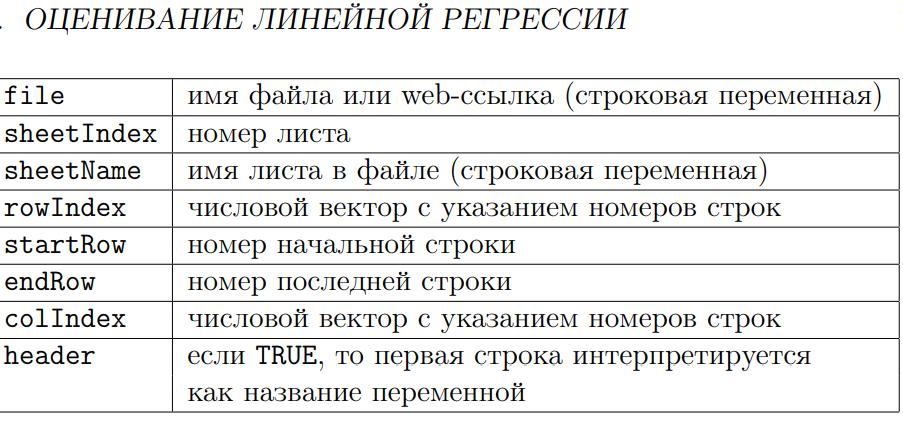
**ЛИНЕЙНАЯ РЕГРЕССИЯ В R. Импорт данных из файла формата MS Excel**

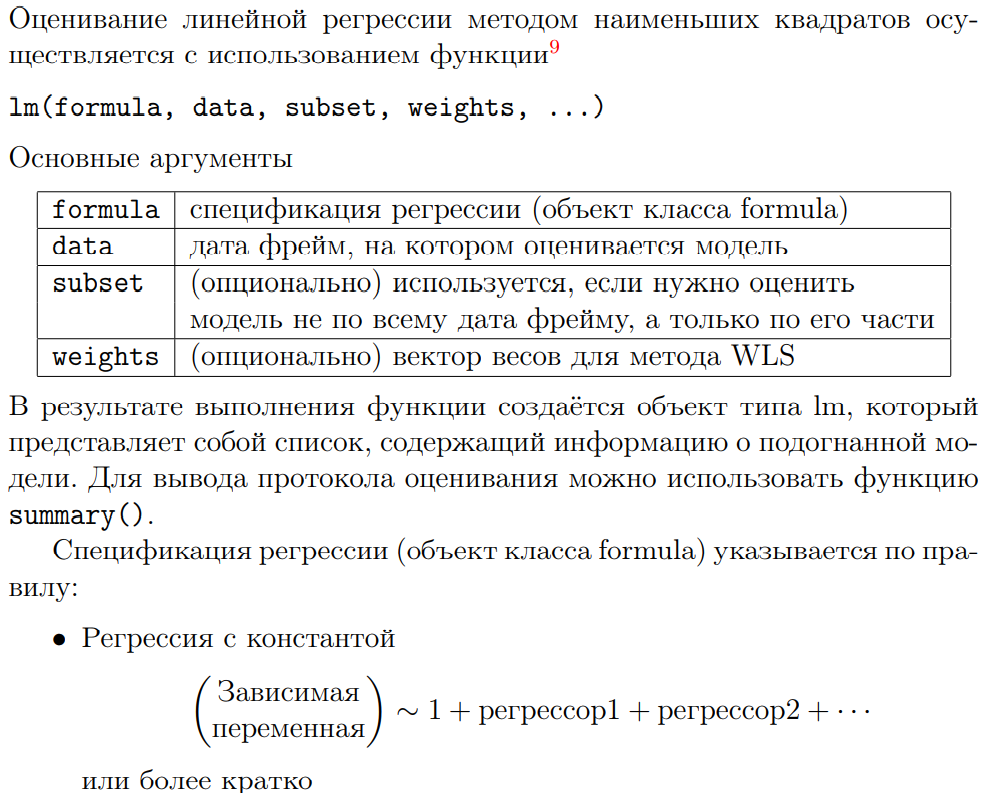
9 lm = Linear Model

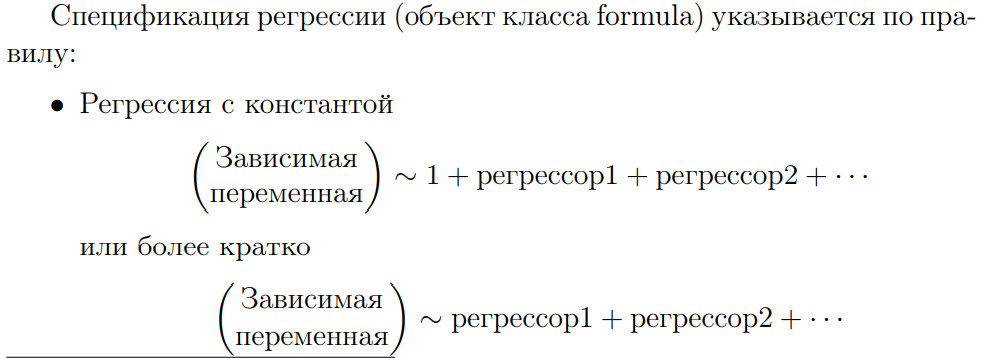


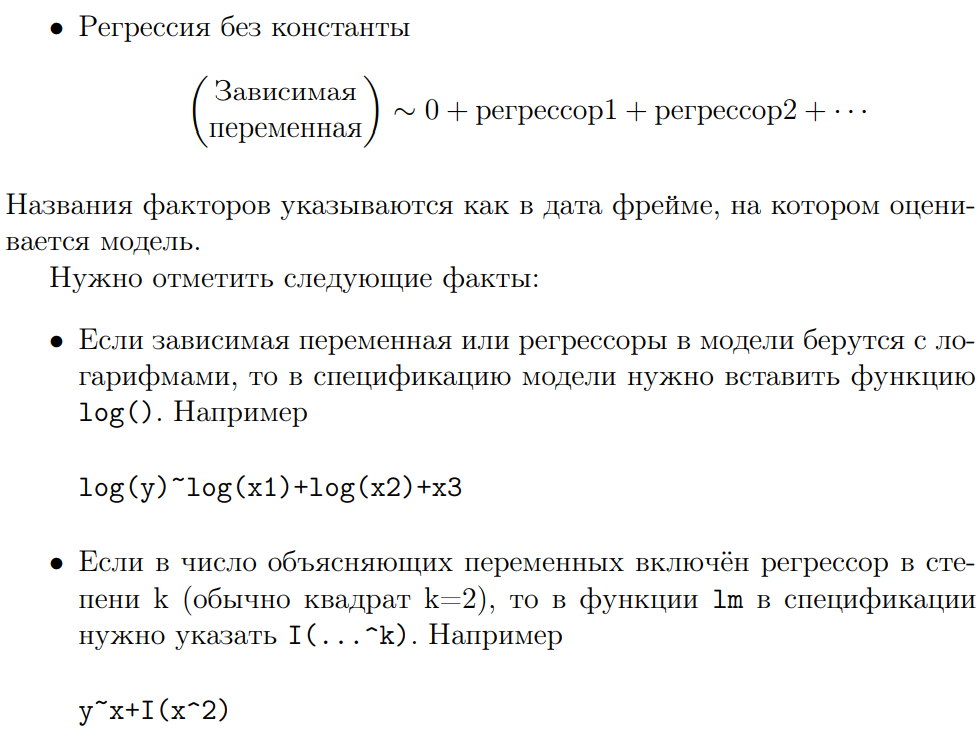




**Оценивание линейной регрессии** 







## Что объясняет регрессия?

Прежде, чем мы приступим к рассмотрению функций MS Excel, позволяющих, решать данные задачи, хотелось бы вам на пальцах объяснить, что, в сущности, предполагает регрессионный анализ. Так вам проще будет сдавать экзамен, а самое главное, интересней изучать предмет.

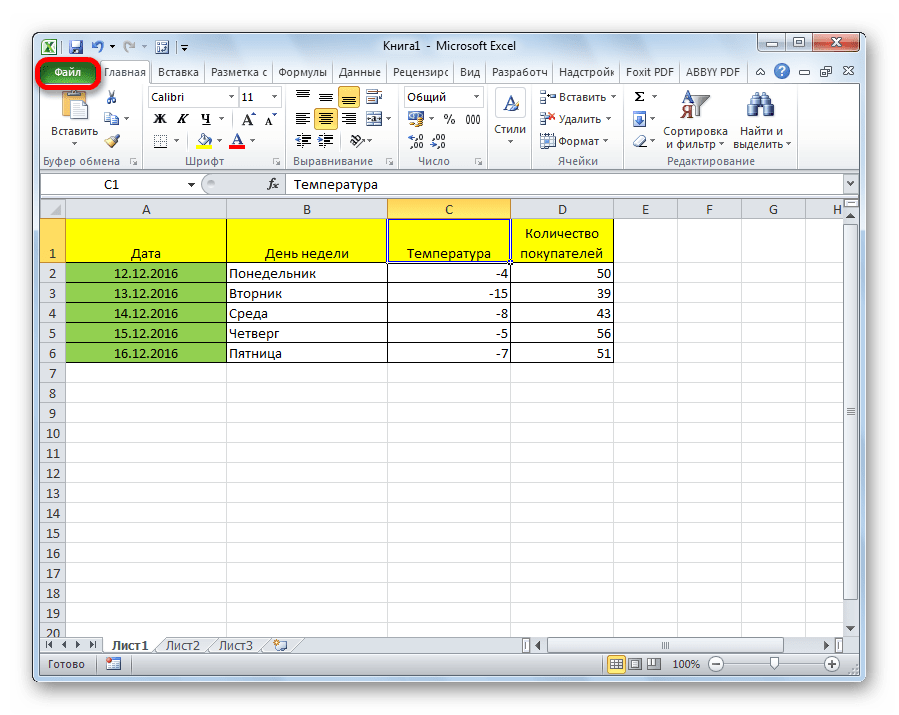
# Регрессионный анализ в excel пример

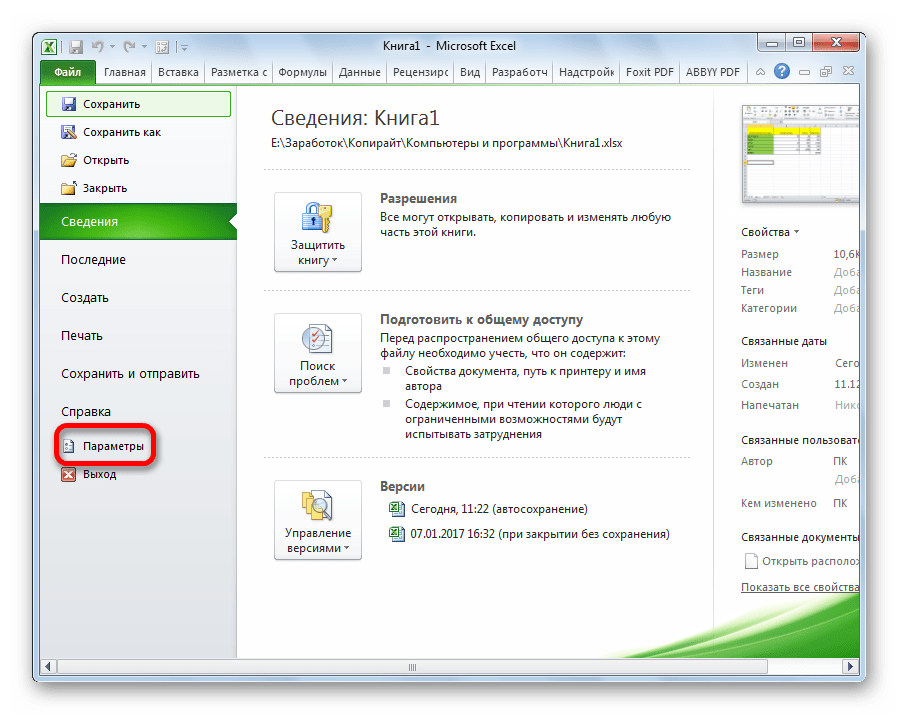
егрессионный анализ является одним из самых востребованных методов статистического исследования. С его помощью можно установить степень влияния независимых величин на зависимую переменную. В функционале Microsoft Excel имеются инструменты, предназначенные для проведения подобного вида анализа. Давайте разберем, что они собой представляют и как ими пользоваться.

## Подключение пакета анализа

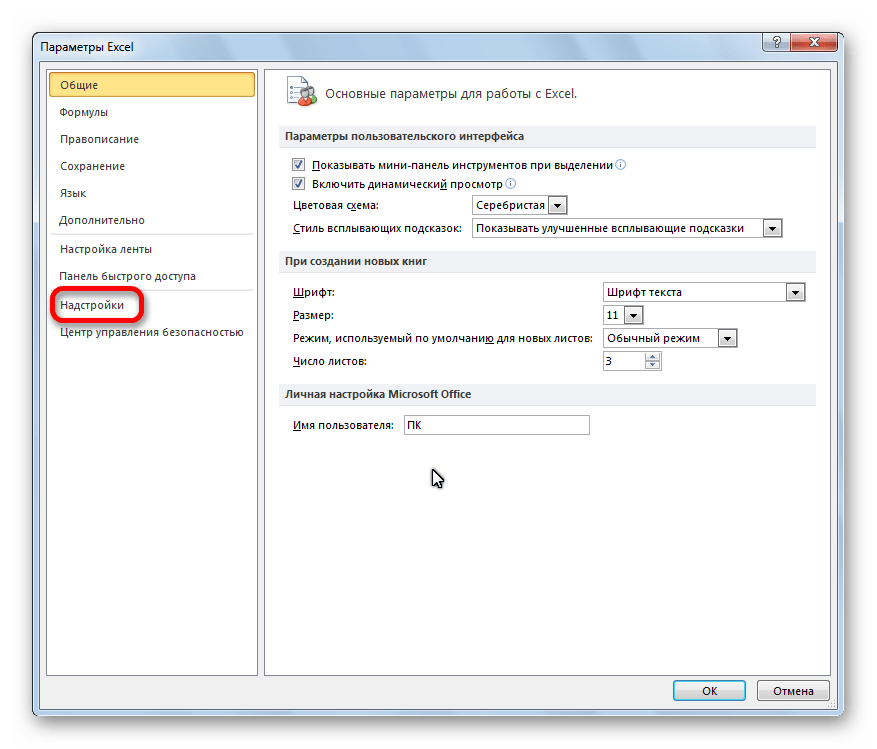
Но, для того, чтобы использовать функцию, позволяющую провести регрессионный анализ, прежде всего, нужно активировать Пакет анализа. Только тогда необходимые для этой процедуры инструменты появятся на ленте Эксель.

Перемещаемся во вкладку «Файл».

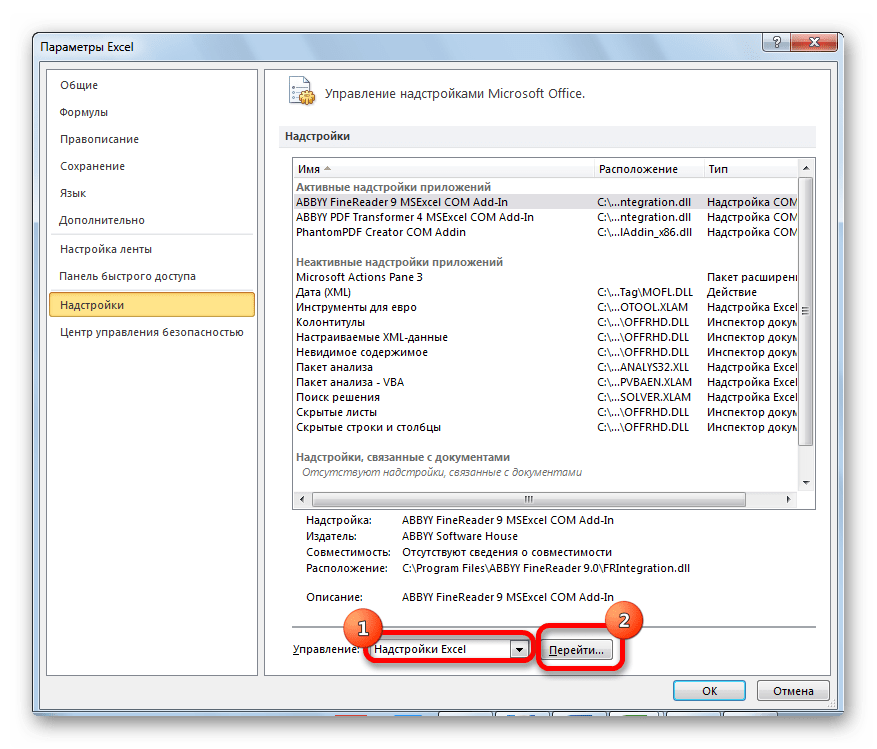


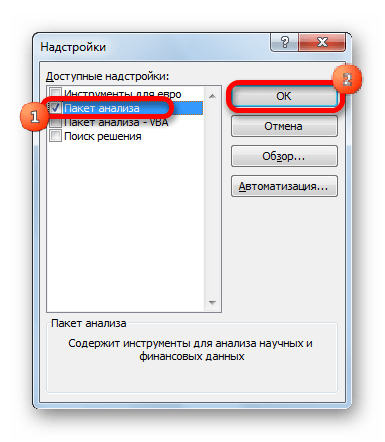


Открывается окно параметров Excel. Переходим в подраздел «Надстройки».

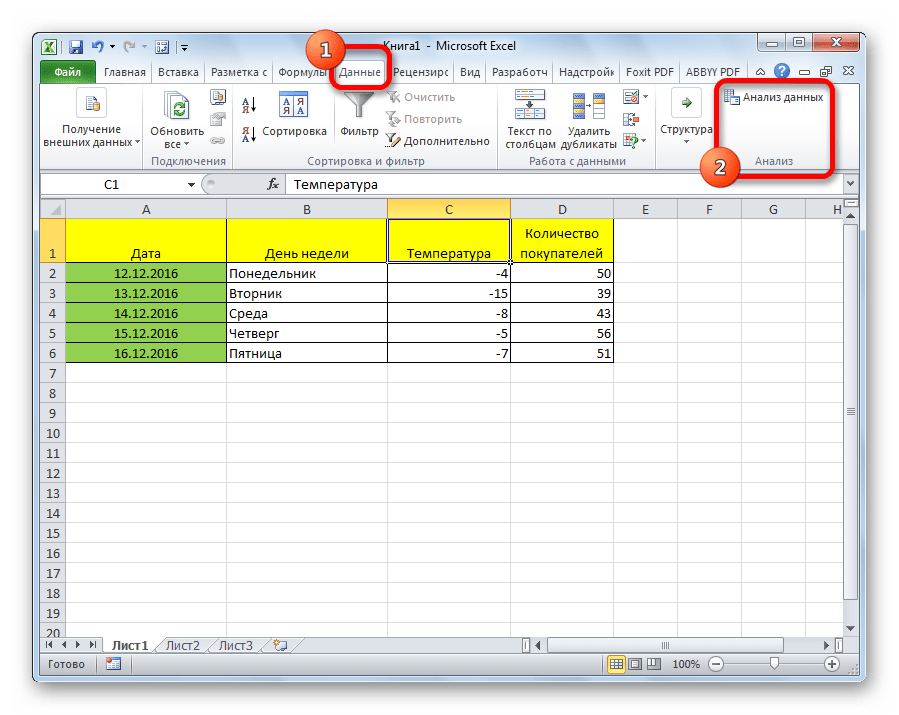


В самой нижней части открывшегося окна переставляем переключатель в блоке «Управление» в позицию «Надстройки Excel», если он находится в другом положении. Жмем на кнопку «Перейти».





Теперь, когда мы перейдем во вкладку «Данные», на ленте в блоке инструментов «Анализ» мы увидим новую кнопку – «Анализ данных».



## Виды регрессионного анализа

Существует несколько видов регрессий:

* параболическая;
* степенная;
* логарифмическая;
* экспоненциальная;
* показательная;
* гиперболическая;
* линейная регрессия.

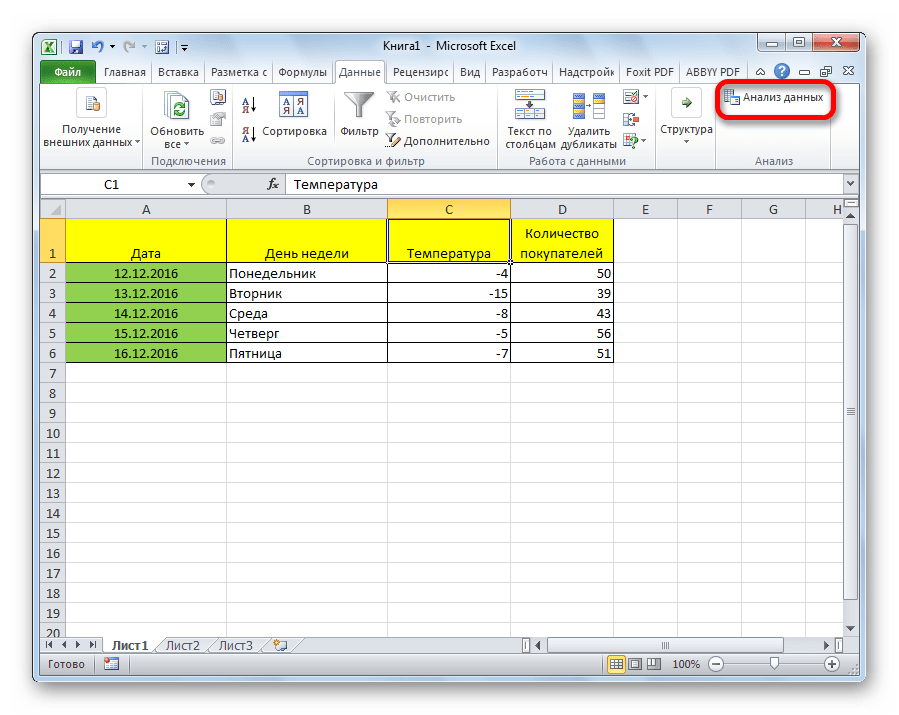
О выполнении последнего вида регрессионного анализа в Экселе мы подробнее поговорим далее.

## Линейная регрессия в программе Excel

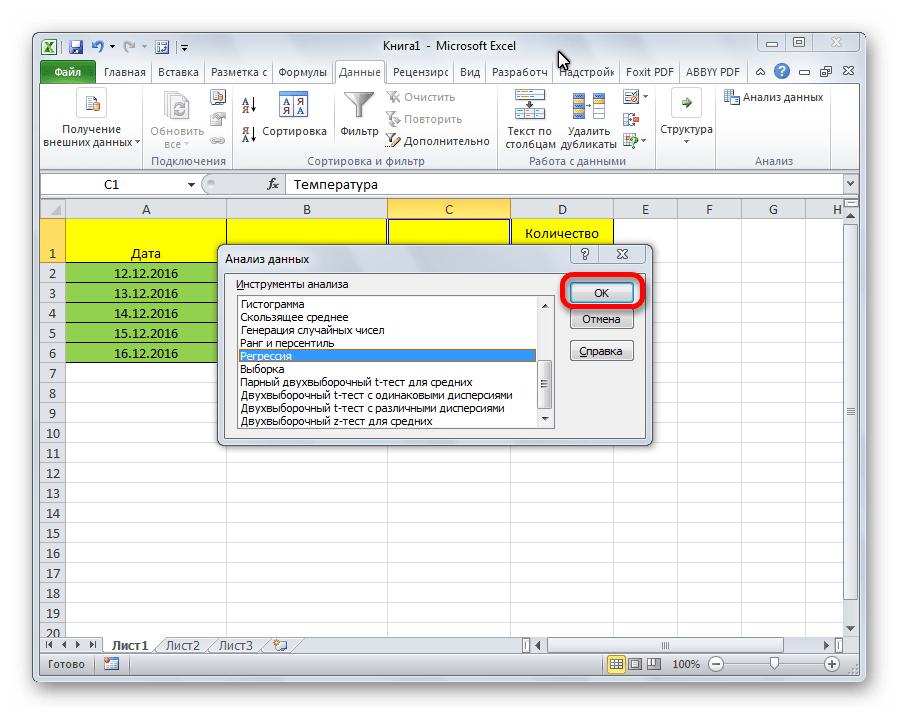
Внизу, в качестве примера, представлена таблица, в которой указана среднесуточная температура воздуха на улице, и количество покупателей магазина за соответствующий рабочий день. Давайте выясним при помощи регрессионного анализа, как именно погодные условия в виде температуры воздуха могут повлиять на посещаемость торгового заведения.

Общее уравнение регрессии линейного вида выглядит следующим образом: У = а0 + а1х1 +…+акхк . В этой формуле Y означает переменную, влияние факторов на которую мы пытаемся изучить. В нашем случае, это количество покупателей. Значение x – это различные факторы, влияющие на переменную. Параметры a являются коэффициентами регрессии. То есть, именно они определяют значимость того или иного фактора. Индекс k обозначает общее количество этих самых факторов.

Кликаем по кнопке «Анализ данных». Она размещена во вкладке «Главная» в блоке инструментов «Анализ».



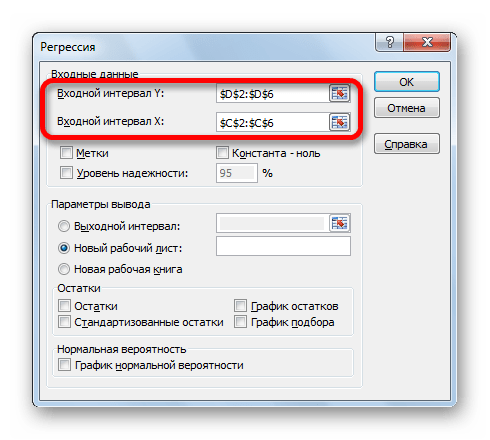
Открывается небольшое окошко. В нём выбираем пункт «Регрессия». Жмем на кнопку «OK».



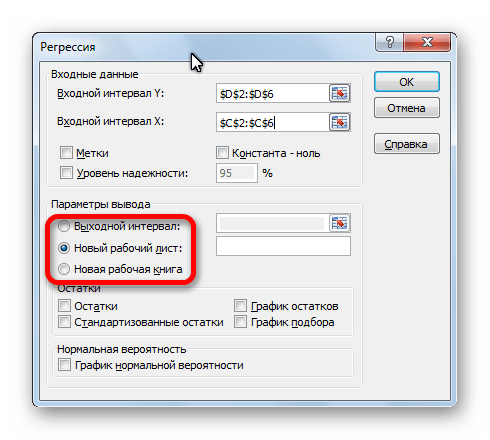
Открывается окно настроек регрессии. В нём обязательными для заполнения полями являются «Входной интервал Y» и «Входной интервал X». Все остальные настройки можно оставить по умолчанию.

В поле «Входной интервал Y» указываем адрес диапазона ячеек, где расположены переменные данные, влияние факторов на которые мы пытаемся установить. В нашем случае это будут ячейки столбца «Количество покупателей». Адрес можно вписать вручную с клавиатуры, а можно, просто выделить требуемый столбец. Последний вариант намного проще и удобнее.

В поле «Входной интервал X» вводим адрес диапазона ячеек, где находятся данные того фактора, влияние которого на переменную мы хотим установить. Как говорилось выше, нам нужно установить влияние температуры на количество покупателей магазина, а поэтому вводим адрес ячеек в столбце «Температура». Это можно сделать теми же способами, что и в поле «Количество покупателей».



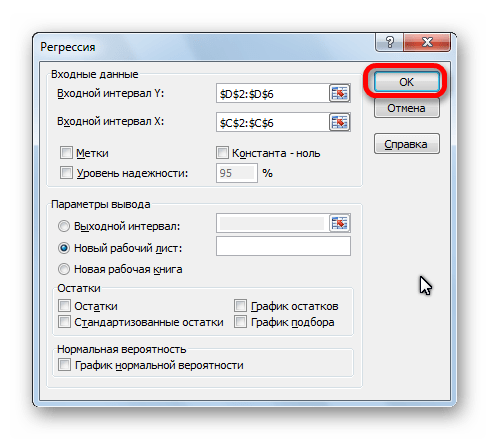
С помощью других настроек можно установить метки, уровень надёжности, константу-ноль, отобразить график нормальной вероятности, и выполнить другие действия. Но, в большинстве случаев, эти настройки изменять не нужно. Единственное на что следует обратить внимание, так это на параметры вывода. По умолчанию вывод результатов анализа осуществляется на другом листе, но переставив переключатель, вы можете установить вывод в указанном диапазоне на том же листе, где расположена таблица с исходными данными, или в отдельной книге, то есть в новом файле.



[Узнать больше](https://an.yandex.ru/count/WwCejI_zO5a3pHm0L3Cd9uKZEuHpdGK0MGGnETdxOG00000umAjSG0n80c2C66W4SDA6YPA9pyw4nG600TlihWI80UFUdmwG0RY5kCFMW8200fW1k8MumrQm0QYnawq8k07edV_Q8y010jW1vFkneW7W0S3png41w06A0lW1qexUlW680WAW0gZ3_HYv0gCphuZiejDUy0ArWV6N0S2H9RW4_m7u1A0EY0MW3f05i1IW1SOKg0N65B05nXIu1R4Tm0MXBCW5wZ_G1Q81u0Lsk0Q-0ia6SqIMIV4Zc6Uf1wltz2Q6uPC4i0U0W90qk0U01Pm60j08ceg0WSAGWOO8u0ZKuRuCW0eAoGhzn3TmbkdgFuWB1AeB47abBlYhzm00DWbN9ntNw0km5F0B1eWCqexUlW7e39y6c0ssgZFW3OA2WRaENg40mGvwEAllmytYiFZ91BWF580Gn_YE8OWGp-_qa0QG493nxBNVcl20Jv70WC88q12-cwzV-13q__y1a141c161yHsX4LAu7eNhKSY0d8EN7E0HajsY1EWHnVAJjUAbkU0EwH8EG3DVhOiTP_0IyHC1c1C6g1FXjVsFXCYCvnRW4x4TW1IWW628582VgRtejelhfm6W5B4Tg1Im5CWK1z3teOy6u1G1w1I40j0LqFUXZmRO5S6AzkoZZxpyO_2O5gxtX8y6eAWMw0Um5kq5oHRG5jAEthu1s1Q15uWN0PWNkE25BgWN2RWN2C0NjZRG5z260zWNmkSww1S2e1W7i1YKcU-F1hWO0j0O8VWOyEMtifFGZQ7v0O0PsQ-ub_AYtyQy0OaP02070000002W6SO2k1d06RZAdSgu_Rg3WW7I6H9vOM9pNtDbSdPbSYzoDZ4vBJ7W6GBe6S85y1c0mWE16l__JxfTh0RGY1h0X3sO6jJ3Kx0QvEU9YxZqhU1kzHe10000-1gsgZEO6x38BB0RMBWR0-aR0000e9oNYq7m6vwFaHFu6yF9z0Y270rITKCwMNP2RpLeRbSta1oWd06O7E-40UaSW1t_VvaTeHrgygpo7PsvFwWT0PWUlDOYg1u1s1wwZH-W7_AxmI6m7mN87_Bzc57I7mKrCpKqDHy0E8IaFcDAm1B1c6QV7z9MgCWOIhhAuBji3kE5_52SVA7m_keEH-DsDi9bb-xj8p4eV0av_5w5wVtCLyjaS6boLts7CG7CkzDTqV8oqB3SzhmNp1anS2G-l2W53lSWnm3H~1" \t "_blank)

**РЕКЛАМА**

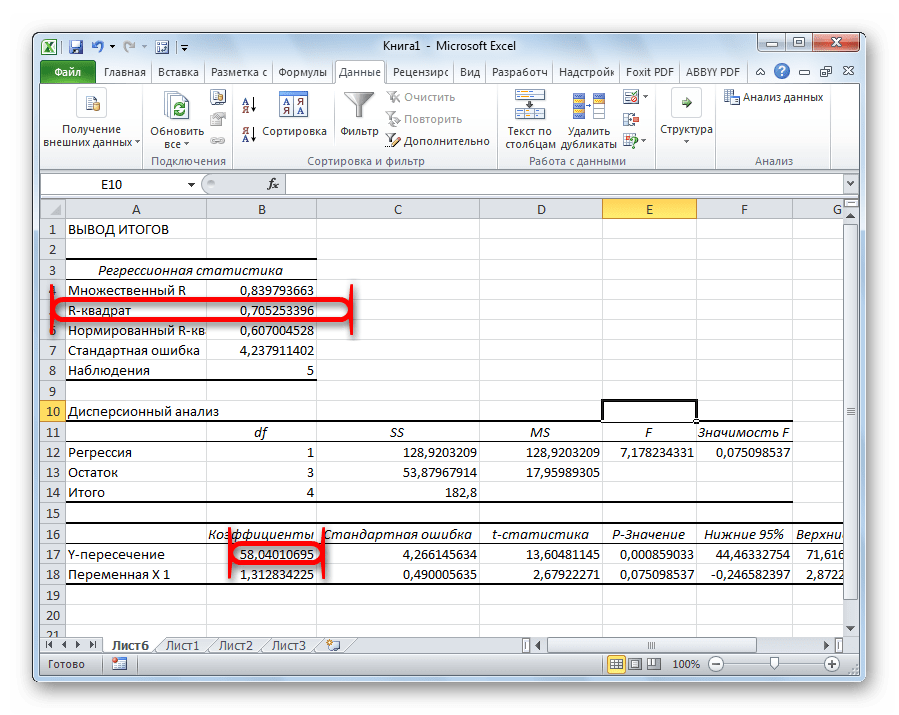
После того, как все настройки установлены, жмем на кнопку «OK».



## Разбор результатов анализа

Результаты регрессионного анализа выводятся в виде таблицы в том месте, которое указано в настройках.

**[Читайте также:  Драйвера для ссд самсунг 860 evo](https://planshet-info.ru/kompjutery/drajvera-dlja-ssd-samsung-860-evo" \t "_blank)**



Одним из основных показателей является R-квадрат. В нем указывается качество модели. В нашем случае данный коэффициент равен 0,705 или около 70,5%. Это приемлемый уровень качества. Зависимость менее 0,5 является плохой.

Ещё один важный показатель расположен в ячейке на пересечении строки «Y-пересечение» и столбца «Коэффициенты». Тут указывается какое значение будет у Y, а в нашем случае, это количество покупателей, при всех остальных факторах равных нулю. В этой таблице данное значение равно 58,04.

Значение на пересечении граф «Переменная X1» и «Коэффициенты» показывает уровень зависимости Y от X. В нашем случае — это уровень зависимости количества клиентов магазина от температуры. Коэффициент 1,31 считается довольно высоким показателем влияния.

Как видим, с помощью программы Microsoft Excel довольно просто составить таблицу регрессионного анализа. Но, работать с полученными на выходе данными, и понимать их суть, сможет только подготовленный человек.

Отблагодарите автора, поделитесь статьей в социальных сетях.

Регрессионный и корреляционный анализ – статистические методы исследования. Это наиболее распространенные способы показать зависимость какого-либо параметра от одной или нескольких независимых переменных.

Ниже на конкретных практических примерах рассмотрим эти два очень популярные в среде экономистов анализа. А также приведем пример получения результатов при их объединении.

## Регрессионный анализ в Excel

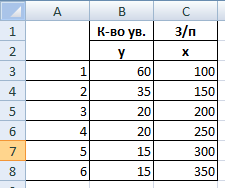
Показывает влияние одних значений (самостоятельных, независимых) на зависимую переменную. К примеру, как зависит количество экономически активного населения от числа предприятий, величины заработной платы и др. параметров. Или: как влияют иностранные инвестиции, цены на энергоресурсы и др. на уровень ВВП.

Результат анализа позволяет выделять приоритеты. И основываясь на главных факторах, прогнозировать, планировать развитие приоритетных направлений, принимать управленческие решения.

* линейной (у = а + bx);
* параболической (y = a + bx + cx 2 );
* экспоненциальной (y = a \* exp(bx));
* степенной (y = a\*x^b);
* гиперболической (y = b/x + a);
* логарифмической (y = b \* 1n(x) + a);
* показательной (y = a \* b^x).

Рассмотрим на примере построение регрессионной модели в Excel и интерпретацию результатов. Возьмем линейный тип регрессии.

Задача. На 6 предприятиях была проанализирована среднемесячная заработная плата и количество уволившихся сотрудников. Необходимо определить зависимость числа уволившихся сотрудников от средней зарплаты.



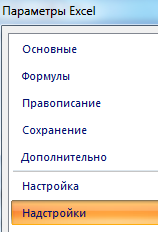
Модель линейной регрессии имеет следующий вид:

Где а – коэффициенты регрессии, х – влияющие переменные, к – число факторов.

В нашем примере в качестве У выступает показатель уволившихся работников. Влияющий фактор – заработная плата (х).

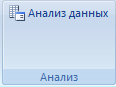
В Excel существуют встроенные функции, с помощью которых можно рассчитать параметры модели линейной регрессии. Но быстрее это сделает надстройка «Пакет анализа».

Активируем мощный аналитический инструмент:

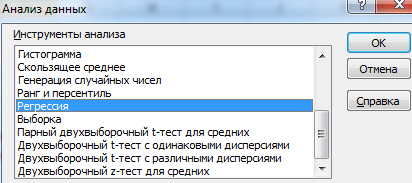
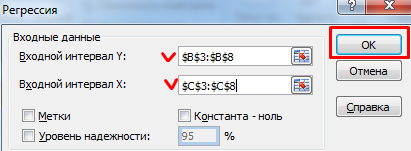
1. Нажимаем кнопку «Офис» и переходим на вкладку «Параметры Excel». «Надстройки».
2. Внизу, под выпадающим списком, в поле «Управление» будет надпись «Надстройки Excel» (если ее нет, нажмите на флажок справа и выберите). И кнопка «Перейти». Жмем.https://exceltable.com/otchety/images/otchety4-3.png
3. Открывается список доступных надстроек. Выбираем «Пакет анализа» и нажимаем ОК.

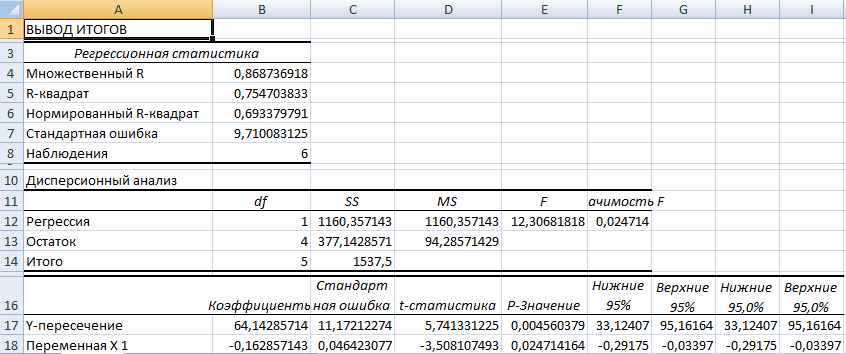


После активации надстройка будет доступна на вкладке «Данные».



Теперь займемся непосредственно регрессионным анализом.

1. Открываем меню инструмента «Анализ данных». Выбираем «Регрессия».
2. Откроется меню для выбора входных значений и параметров вывода (где отобразить результат). В полях для исходных данных указываем диапазон описываемого параметра (У) и влияющего на него фактора (Х). Остальное можно и не заполнять.
3. После нажатия ОК, программа отобразит расчеты на новом листе (можно выбрать интервал для отображения на текущем листе или назначить вывод в новую книгу).



В первую очередь обращаем внимание на R-квадрат и коэффициенты.

R-квадрат – коэффициент детерминации. В нашем примере – 0,755, или 75,5%. Это означает, что расчетные параметры модели на 75,5% объясняют зависимость между изучаемыми параметрами. Чем выше коэффициент детерминации, тем качественнее модель. Хорошо – выше 0,8. Плохо – меньше 0,5 (такой анализ вряд ли можно считать резонным). В нашем примере – «неплохо».

Коэффициент 64,1428 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.

**[Читайте также:  Black desert free to play](https://planshet-info.ru/kompjutery/black-desert-free-to-play" \t "_blank)**

Коэффициент -0,16285 показывает весомость переменной Х на Y. То есть среднемесячная заработная плата в пределах данной модели влияет на количество уволившихся с весом -0,16285 (это небольшая степень влияния). Знак «-» указывает на отрицательное влияние: чем больше зарплата, тем меньше уволившихся. Что справедливо.

## Корреляционный анализ в Excel

Корреляционный анализ помогает установить, есть ли между показателями в одной или двух выборках связь. Например, между временем работы станка и стоимостью ремонта, ценой техники и продолжительностью эксплуатации, ростом и весом детей и т.д.

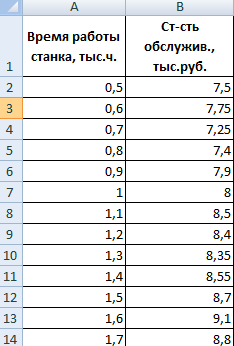
Если связь имеется, то влечет ли увеличение одного параметра повышение (положительная корреляция) либо уменьшение (отрицательная) другого. Корреляционный анализ помогает аналитику определиться, можно ли по величине одного показателя предсказать возможное значение другого.

Коэффициент корреляции обозначается r. Варьируется в пределах от +1 до -1. Классификация корреляционных связей для разных сфер будет отличаться. При значении коэффициента 0 линейной зависимости между выборками не существует.

Рассмотрим, как с помощью средств Excel найти коэффициент корреляции.

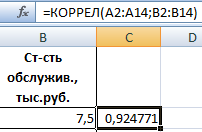
Для нахождения парных коэффициентов применяется функция КОРРЕЛ.

Задача: Определить, есть ли взаимосвязь между временем работы токарного станка и стоимостью его обслуживания.



Ставим курсор в любую ячейку и нажимаем кнопку fx.

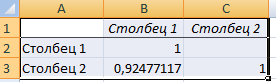
1. В категории «Статистические» выбираем функцию КОРРЕЛ.
2. Аргумент «Массив 1» – первый диапазон значений – время работы станка: А2:А14.
3. Аргумент «Массив 2» – второй диапазон значений – стоимость ремонта: В2:В14. Жмем ОК.



Чтобы определить тип связи, нужно посмотреть абсолютное число коэффициента (для каждой сферы деятельности есть своя шкала).

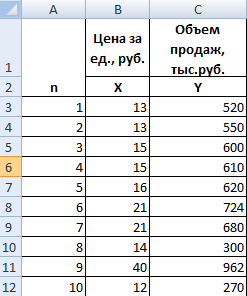
Для корреляционного анализа нескольких параметров (более 2) удобнее применять «Анализ данных» (надстройка «Пакет анализа»). В списке нужно выбрать корреляцию и обозначить массив. Все.

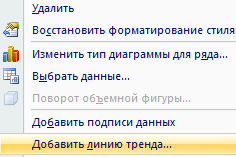
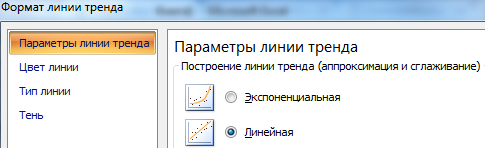
Полученные коэффициенты отобразятся в корреляционной матрице. Наподобие такой:



## Корреляционно-регрессионный анализ

На практике эти две методики часто применяются вместе.



1. Строим корреляционное поле: «Вставка» – «Диаграмма» – «Точечная диаграмма» (дает сравнивать пары). Диапазон значений – все числовые данные таблицы.
2. Щелкаем левой кнопкой мыши по любой точке на диаграмме. Потом правой. В открывшемся меню выбираем «Добавить линию тренда».
3. Назначаем параметры для линии. Тип – «Линейная». Внизу – «Показать уравнение на диаграмме».
4. Жмем «Закрыть».



Теперь стали видны и данные регрессионного анализа.

Регрессионный анализ — это статистический метод исследования, позволяющий показать зависимость того или иного параметра от одной либо нескольких независимых переменных. В докомпьютерную эру его применение было достаточно затруднительно, особенно если речь шла о больших объемах данных. Сегодня, узнав как построить регрессию в Excel, можно решать сложные статистические задачи буквально за пару минут. Ниже представлены конкретные примеры из области экономики.

## Виды регрессии

Само это понятие было введено в математику Фрэнсисом Гальтоном в 1886 году. Регрессия бывает:

* линейной;
* параболической;
* степенной;
* экспоненциальной;
* гиперболической;
* показательной;
* логарифмической.

## Пример 1

Рассмотрим задачу определения зависимости количества уволившихся членов коллектива от средней зарплаты на 6 промышленных предприятиях.

Задача. На шести предприятиях проанализировали среднемесячную заработную плату и количество сотрудников, которые уволились по собственному желанию. В табличной форме имеем:

Для задачи определения зависимости количества уволившихся работников от средней зарплаты на 6 предприятиях модель регрессии имеет вид уравнения Y = а + а1x1 +…+аkxk, где хi — влияющие переменные, ai — коэффициенты регрессии, a k — число факторов.

Для данной задачи Y — это показатель уволившихся сотрудников, а влияющий фактор — зарплата, которую обозначаем X.

РЕКЛАМА

## Использование возможностей табличного процессора «Эксель»

Анализу регрессии в Excel должно предшествовать применение к имеющимся табличным данным встроенных функций. Однако для этих целей лучше воспользоваться очень полезной надстройкой «Пакет анализа». Для его активации нужно:

* с вкладки «Файл» перейти в раздел «Параметры»;
* в открывшемся окне выбрать строку «Надстройки»;
* щелкнуть по кнопке «Перейти», расположенной внизу, справа от строки «Управление»;
* поставить галочку рядом с названием «Пакет анализа» и подтвердить свои действия, нажав «Ок».

**[Читайте также:  Технические характеристики поколений эвм](https://planshet-info.ru/kompjutery/tehnicheskie-harakteristiki-pokolenij-jevm" \t "_blank)**

Если все сделано правильно, в правой части вкладки «Данные», расположенном над рабочим листом «Эксель», появится нужная кнопка.

## Линейная регрессия в Excel

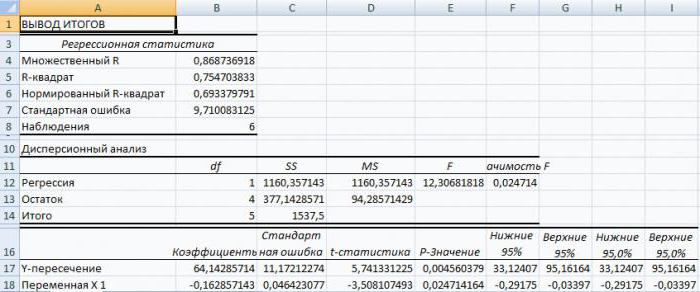
Теперь, когда под рукой есть все необходимые виртуальные инструменты для осуществления эконометрических расчетов, можем приступить к решению нашей задачи. Для этого:

* щелкаем по кнопке «Анализ данных»;
* в открывшемся окне нажимаем на кнопку «Регрессия»;
* в появившуюся вкладку вводим диапазон значений для Y (количество уволившихся работников) и для X (их зарплаты);
* подтверждаем свои действия нажатием кнопки «Ok».

В результате программа автоматически заполнит новый лист табличного процессора данными анализа регрессии. Обратите внимание! В Excel есть возможность самостоятельно задать место, которое вы предпочитаете для этой цели. Например, это может быть тот же лист, где находятся значения Y и X, или даже новая книга, специально предназначенная для хранения подобных данных.

## Анализ результатов регрессии для R-квадрата

В Excel данные полученные в ходе обработки данных рассматриваемого примера имеют вид:



Прежде всего, следует обратить внимание на значение R-квадрата. Он представляет собой коэффициент детерминации. В данном примере R-квадрат = 0,755 (75,5%), т. е. расчетные параметры модели объясняют зависимость между рассматриваемыми параметрами на 75,5 %. Чем выше значение коэффициента детерминации, тем выбранная модель считается более применимой для конкретной задачи. Считается, что она корректно описывает реальную ситуацию при значении R-квадрата выше 0,8. Если R-квадрата 2 (RI) представляет собой числовую характеристику доли общего разброса и показывает, разброс какой части экспериментальных данных, т.е. значений зависимой переменной соответствует уравнению линейной регрессии. В рассматриваемой задаче эта величина равна 84,8%, т. е. статистические данные с высокой степенью точности описываются полученным УР.

F-статистика, называемая также критерием Фишера, используется для оценки значимости линейной зависимости, опровергая или подтверждая гипотезу о ее существовании.

Значение t-статистики (критерий Стьюдента) помогает оценивать значимость коэффициента при неизвестной либо свободного члена линейной зависимости. Если значение t-критерия > tкр, то гипотеза о незначимости свободного члена линейного уравнения отвергается.

В рассматриваемой задаче для свободного члена посредством инструментов «Эксель» было получено, что t=169,20903, а p=2,89Е-12, т. е. имеем нулевую вероятность того, что будет отвергнута верная гипотеза о незначимости свободного члена. Для коэффициента при неизвестной t=5,79405, а p=0,001158. Иными словами вероятность того, что будет отвергнута верная гипотеза о незначимости коэффициента при неизвестной, равна 0,12%.

Таким образом, можно утверждать, что полученное уравнение линейной регрессии адекватно.

## Задача о целесообразности покупки пакета акций

Множественная регрессия в Excel выполняется с использованием все того же инструмента «Анализ данных». Рассмотрим конкретную прикладную задачу.

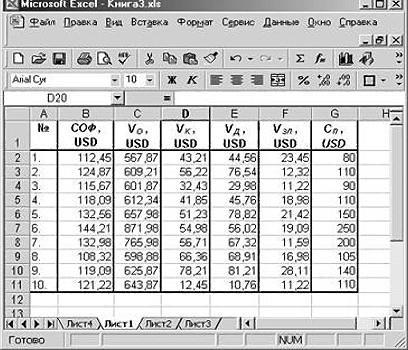
Руководство компания «NNN» должно принять решение о целесообразности покупки 20 % пакета акций АО «MMM». Стоимость пакета (СП) составляет 70 млн американских долларов. Специалистами «NNN» собраны данные об аналогичных сделках. Было принято решение оценивать стоимость пакета акций по таким параметрам, выраженным в миллионах американских долларов, как:

* кредиторская задолженность (VK);
* объем годового оборота (VO);
* дебиторская задолженность (VD);
* стоимость основных фондов (СОФ).

Кроме того, используется параметр задолженность предприятия по зарплате (V3 П) в тысячах американских долларов.

## Решение средствами табличного процессора Excel

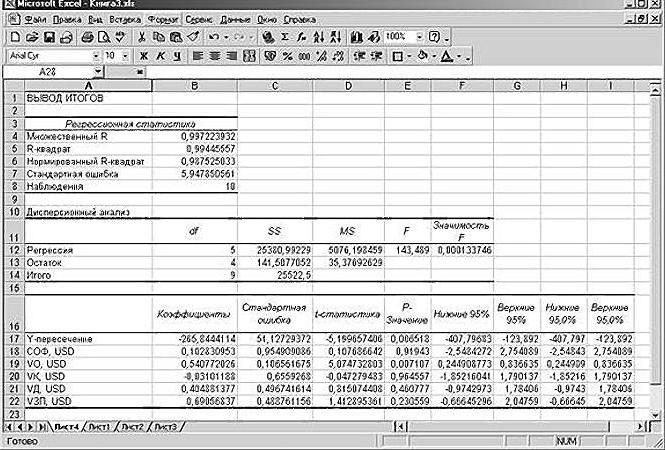
Прежде всего, необходимо составить таблицу исходных данных. Она имеет следующий вид:



* вызывают окно «Анализ данных»;
* выбирают раздел «Регрессия»;
* в окошко «Входной интервал Y» вводят диапазон значений зависимых переменных из столбца G;
* щелкают по иконке с красной стрелкой справа от окна «Входной интервал X» и выделяют на листе диапазон всех значений из столбцов B,C, D, F.

Отмечают пункт «Новый рабочий лист» и нажимают «Ok».

Получают анализ регрессии для данной задачи.



## Изучение результатов и выводы

«Собираем» из округленных данных, представленных выше на листе табличного процессора Excel, уравнение регрессии:

СП = 0,103\*СОФ + 0,541\*VO – 0,031\*VK +0,405\*VD +0,691\*VZP – 265,844.

В более привычном математическом виде его можно записать, как:

y = 0,103\*x1 + 0,541\*x2 – 0,031\*x3 +0,405\*x4 +0,691\*x5 – 265,844

Данные для АО «MMM» представлены в таблице: