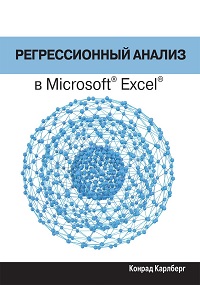
**Конрад Карлберг. Регрессионный анализ в Microsoft Excel**

Регрессионный анализ в Microsoft Excel – наиболее полное руководств по использованию MS Excel для решения задач регрессионного анализа в области бизнес-аналитики. Конрад Карлберг доступно объясняет теоретические вопросы, знание которых поможет вам избежать многих ошибок как при самостоятельном проведении регрессионного анализа, так и при оценке результатов анализа, выполненного другими людьми. Весь материал, от простых корреляций и t-тестов до множественного ковариационного анализа, основан на реальных примерах и сопровождается подробным описанием соответствующих пошаговых процедур.

В книге обсуждаются особенности и противоречия, связанные с функциями Excel для работы с регрессией, рассматриваются последствия использования каждой их опции и каждого аргумента и объясняется, как надежно применять регрессионные методы в самых разных областях, от медицинских исследований до финансового анализа.

Ранее я опубликовал [Левин. Статистика для менеджеров с использованием Microsoft Excel](http://baguzin.ru/wp/?p=5285).

Конрад Карлберг. Регрессионный анализ в Microsoft Excel. – М.: Вильямс, 2017. – 400 с.



Купить книгу в [Ozon](https://www.ozon.ru/context/detail/id/142149278/?partner=baguzin) или [Лабиринте](https://www.labirint.ru/books/612654/?p=13320)

### Глава 1. Оценка изменчивости данных

В распоряжении статистиков имеется множество показателей вариации (изменчивости). Один из них – сумма квадратов отклонений индивидуальных значений от среднего. В Excel для него используется функция КВАДРОТКЛ(). Но чаще используется дисперсия. Дисперсия — это среднее квадратов отклонений. Дисперсия нечувствительна к количеству значений в исследуемом наборе данных (в то время как сумма квадратов отклонений растет с числом измерений).

Программа Excel предлагает две функции, возвращающие дисперсию: ДИСП.Г() и ДИСП.В():

* Используйте функцию ДИСП.Г(), если подлежащие обработке значения образуют генеральную совокупность. Т.е., значения, содержащиеся в диапазоне, являются единственными значениями, которые вас интересуют.
* Используйте функцию ДИСП.В(), если подлежащие обработке значения образуют выборку из совокупности большего объема. Предполагается, что имеются дополнительные значения, дисперсию которых вы также можете оценить.

См. также [СТАНДОТКЛОН.В и СТАНДОТКЛОН.Г: в чем различие?](http://baguzin.ru/wp/?p=15634)

Если такая величина, как среднее значение или коэффициент корреляции, рассчитывается на основе генеральной совокупности, то она называется параметром. Аналогичная величина, рассчитываемая на основе выборки, называется статистикой. Отсчитывая отклонения *от среднего значения* в данном наборе, вы получите сумму квадратов отклонений меньшей величины, чем если бы отсчитывали их от любого другого значения. Аналогичное утверждение справедливо и для дисперсии.

Чем больше объем выборки, тем точнее рассчитанное значение статистики. Но не существует ни одной выборки с объемом меньше объема генеральной совокупности, относительно которой вы могли бы быть уверены в том, что значение статистики совпадает со значением параметра.

Допустим, у вас есть набор из 100 значений роста, среднее которых отличается от среднего по генеральной совокупности, каким бы малым ни было это различие. Рассчитав дисперсию для выборки, вы получите некоторое ее значение, скажем, 4. Это значение меньше любого другого, которое можно получить, рассчитывая отклонение каждого из 100 значений роста относительно любого значения, отличного от среднего по выборке, в там числе и относительно истинного среднего по генеральной совокупности. Поэтому вычисленная дисперсия будет отличаться, причем в меньшую сторону, от дисперсии, которую вы получили бы, если бы каким-то образом узнали и использовали не выборочное среднее, а параметр генеральной совокупности.

Средняя сумма квадратов, определенная для выборки, дает нижнюю оценку дисперсии генеральной совокупности. Вычисленную таким способом дисперсию называют *смещенной* оценкой. Оказывается, чтобы исключить смещение и получить несмещенную оценку, достаточно разделить сумму квадратов отклонений не на *n*, где *n* — размер выборки, а на *n – 1*.

Величина *n – 1* называется количеством (числом) степеней свободы. Существуют разные способы расчета этой величины, хотя все они включают либо вычитание некоторого числа из размера выборки, либо подсчет количества категорий, в которые попадают наблюдения.

Суть различия между функциями ДИСП.Г() и ДИСП.В() состоит в следующем:

* В функции ДИСП.Г() сумма квадратов делится на количество наблюдений и, следовательно, представляет смещенную оценку дисперсии, истинное среднее.
* В функции ДИСП.В() сумма квадратов делится на количество наблюдений минус 1, т.е. на количество степеней свободы, что дает более точную, несмещенную оценку дисперсии генеральной совокупности, из которой была извлечена данная выборка.

Стандартное отклонение (англ. *standard deviation*, SD) – есть квадратный корень из дисперсии:

Возведение отклонений в квадрат переводит шкалу измерений в другую метрику, являющуюся квадратом исходной: метры — в квадратные метры, доллары — в квадратные доллары и т.д. Стандартное отклонение — это корень квадратный из дисперсии, и поэтому оно возвращает нас к исходным единицам измерения. Что удобнее.

Часто приходится рассчитывать стандартное отклонение после того, как данных были подвергнуты некоторым манипуляциям. И хотя в этих случаях результаты несомненно являются стандартными отклонениями, их принято называть *стандартными ошибками*. Существует несколько разновидностей стандартных ошибок, в том числе стандартная ошибка измерения, стандартная ошибка пропорции, стандартная ошибка среднего.

Предположим, вы собрали данные о росте 25 случайно выбранных взрослых мужчин в каждом из 50 штатов. Далее вы вычисляете средний рост взрослых мужчин в каждом штате. Полученные 50 средних значений в свою очередь можно считать наблюдениями. Исходя из этого, вы могли бы рассчитать их стандартное отклонение, которое и является *стандартной ошибкой среднего*. Рис. 1. позволяет сравнить распределение 1250 исходных индивидуальных значений (данные о росте 25 мужчин по каждому из 50 штатов) с распределением средних значений 50 штатов. Формула для оценки стандартной ошибки среднего (т.е. стандартного отклонения средних значений, а не индивидуальных наблюдений):

где – стандартная ошибка среднего; *s* – стандартное отклонение исходных наблюдений; *n* – количество наблюдений в выборке.

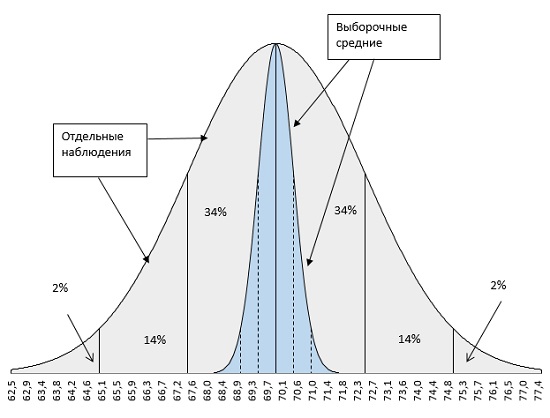


Рис. 1. Вариация средних значений от штата к штату значительно меньше вариации индивидуальных результатов наблюдений

В статистике существует соглашение относительно использования греческих и латинских букв для обозначения статистических величин. Греческими буквами принято обозначать параметры генеральной совокупности, латинскими — выборочные статистики. Следовательно, если речь идет о стандартном отклонении генеральной совокупности, мы записываем его как σ; если же рассматривается стандартное отклонение выборки, то используем обозначение s. Что касается символов для обозначения средних, то они согласуются между собой не столь удачно. Среднее по генеральной совокупности обозначается греческой буквой μ. Однако для представления выборочного среднего традиционно используется символ X̅.

*z-оценка* выражает положение наблюдения в распределении в единицах стандартного отклонения. Например, z = 1,5 означает, что наблюдение отстоит от среднего на 1,5 стандартного отклонения в сторону больших значений. Термин *z-оценка* используют для индивидуальных оценок, т.е. для измерений, приписываемых отдельным элементам выборки. В отношении таких статистик (например, среднее значение по штату) используют термин *z-значение*:

где X̅ – среднее значение выборки, μ – среднее значение генеральной совокупности,  – стандартная ошибка средних набора выборок:

где σ – стандартная ошибка генеральной совокупности (индивидуальных измерений), *n* – размер выборки.

Предположим, вы работаете инструктором в гольф-клубе. Вы имели возможность в течение длительного времени измерять дальность ударов и знаете, что ее среднее значение составляет 205 ярдов, а стандартное отклонение — 36 ярдов. Вам предложили новую клюшку, утверждая, что она увеличит дальность удара на 10 ярдов. Вы просите каждого из последующих 81 посетителей клуба выполнить пробный удар новой клюшкой и записываете его дальность удара. Оказалось, что средняя дальность удара новой клюшкой составляет 215 ярдов. Какова вероятность того, что разница в 10 ярдов (215 – 205) обусловлена исключительно ошибкой выборки? Или по-другому: какова вероятность того, что при более масштабном тестировании новая клюшка не продемонстрирует увеличение дальности удара по сравнению с имеющимся долговременным средним показателем 205 ярдов?

Мы можем проверить это, сформировав z-значение. Стандартная ошибка среднего:

Тогда z-значение:

Нам нужно найти вероятность того, что среднее по выборке будет отстоять от среднего по генеральной совокупности на 2,5σ. Если вероятность будет маленькой, значит отличия обусловлены не случайностью, а качеством новой клюшки. В Excel для определения вероятности z-значения нет готовой функции. Однако можно использовать формулу =1-НОРМ.СТ.РАСП(z-значение;ИСТИНА), где функция НОРМ.СТ.РАСП() возвращает площадь под нормальной кривой слева от z-значения (рис. 2).

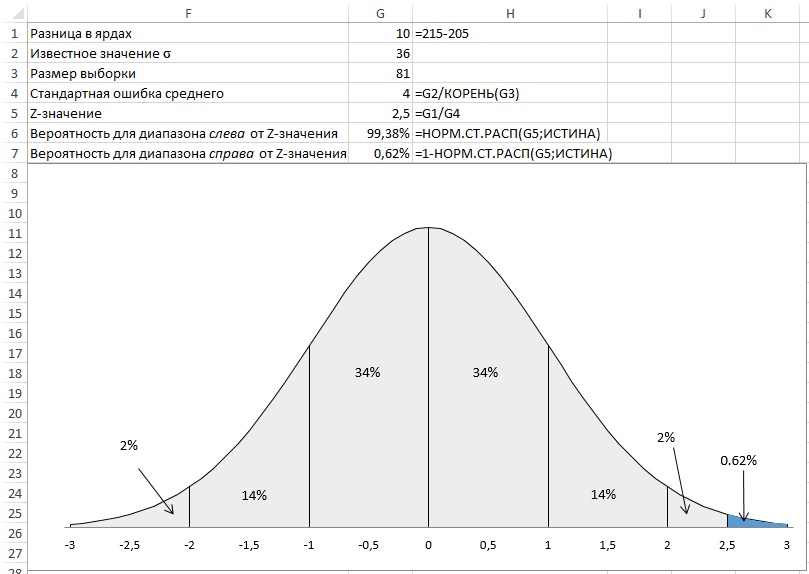


Рис. 2. Функция НОРМ.СТ.РАСП() возвращает площадь под кривой слева от z-значения

Второй аргумент функции НОРМ.СТ.РАСП() может принимать два значения: ИСТИНА – функция возвращает площадь области под кривой слева от точки, заданной первым аргументом; ЛОЖЬ – функция возвращает высоту кривой в точке, заданной первым аргументом.

Если среднее значение (μ) и стандартное отклонение (σ) генеральной совокупности не известны, используется t-значение. Структуры z- и t-значения отличаются тем, что для нахождения t-значения используется стандартное отклонение s, полученное на основе выборочных результатов, а не известное значение параметра генеральной совокупности σ. Нормальная кривая имеет единственную форму, а форма распределения t-значений варьирует в зависимости от количества степеней свободы df (от англ. *degrees of freedom*) выборки, которую оно представляет. Количество степеней свободы выборки равно *n – 1*, где *n* — размер выборки (рис. 3).

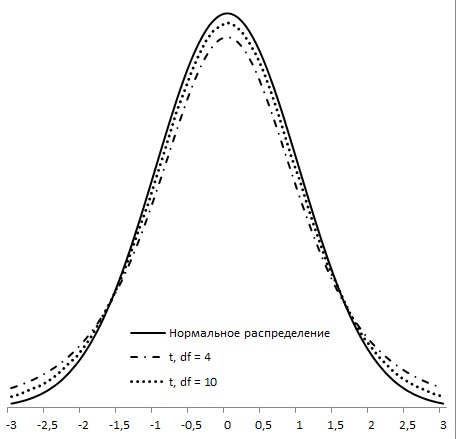


Рис. 3. Форма t-распределений, возникающих в тех случаях, когда параметр σ неизвестен, отличается от формы нормального распределения

В Excel есть две версию функции для t-распределения также называемого распределением Стьюдента: СТЬЮДЕНТ.РАСП() возвращает величину площади под кривой слева от заданного t-значения, а СТЬЮДЕНТ.РАСП.ПХ() – справа.

### Глава 2. Корреляция

Корреляция — это мера зависимости между элементами набора упорядоченных пар. Корреляция характеризуется *коэффициентам корреляции Пирсона* – r. Коэффициент может принимать значения в интервале от –1,0 до +1,0.

где *Sx* и *Sy* – стандартные отклонения переменных *Х* и *Y*, *Sxy* – ковариация:

В этой формуле ковариация делится на стандартные отклонения переменных *Х* и *Y*, тем самым удаляя из ковариации эффекты масштабирования, связанные с единицами измерения. В Excel используется функция КОРРЕЛ(). В названии этой функции отсутствуют уточняющие элементы Г и В, которые используются в названиях таких функций, как СТАНДОТКЛОН(), ДИСП() или КОВАРИАЦИЯ(). Хотя коэффициенте корреляции по выборке предоставляемая смещенную оценку, однако причина смещения иная, нежели в случае дисперсии или стандартного отклонения.

В зависимости от величины генерального коэффициента корреляции (часто обозначаемого греческой буквой *ρ*), коэффициент корреляции *r* дает смещенную оценку, причем эффект смещения усиливается с уменьшением размера выборок. Тем не менее мы не пытаемся корректировать это смещение аналогично тому, как, например, делали это при вычислении стандартного отклонения, когда подставляли в соответствующую формулу не количество наблюдений, а количество степеней свободы. В действительности количество наблюдений, используемое для вычисления ковариации, не оказывает никакого влияния на величину.

Стандартный коэффициент корреляции предназначен для использования с переменными, связанными между собой линейным соотношением. Наличие нелинейности и/или ошибок в данных (выбросы) приводят к неверному расчету коэффициента корреляции. Для диагностики проблем с данными рекомендуется строить точечные диаграммы. Это единственный тип диаграмм в Excel, в котором и горизонтальная, и вертикальная оси трактуются как оси значений. Линейная же диаграмма один из столбцов определяет, как ось категорий, что искажает картину данных (рис. 4).

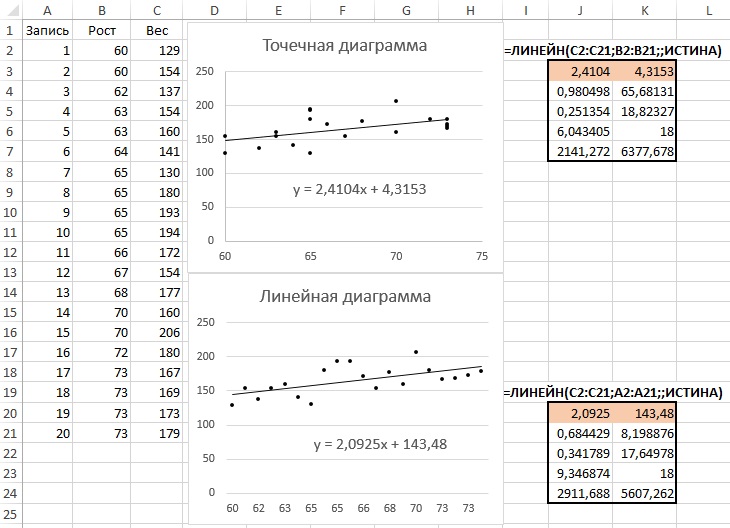


Рис. 4. Линии регрессии кажутся одинаковыми, однако сравните между собой их уравнения

Наблюдения, использованные для построения линейной диаграммы, располагаются вдоль горизонтальной оси эквидистантно. Надписи делений вдоль этой оси — это и есть всего лишь надписи, а не числовые значения.

Несмотря на то что корреляция часто означает наличие причинно-следственной связи, она не может служить доказательством того, что так оно и есть. Статистика не используется для демонстрации того, истинна или ложна теория. Для исключения конкурирующих объяснений результатов наблюдений ставят *плановые эксперименты*. Статистика же привлекается для обобщения информации, собранной в ходе таких экспериментов, и количественной оценки вероятности того, что принимаемое решение может быть неверным при имеющейся доказательной базе.

### Глава 3. Простая регрессия

Если две переменные связаны между собой, так что значение коэффициента корреляции превышает, скажем, 0,5, то в этом случае можно прогнозировать (с некоторой точностью) неизвестное значение одной переменной по известному значению другой. Для получения прогнозных значений цены, исходя из данных, приведенных на рис. 5, можно использовать любой из нескольких возможных способов, но почти наверняка вы не будете использовать тот, который представлен на рис. 5. И все же вам стоит с ним ознакомиться, поскольку ни один другой способ не позволяет так же отчетливо продемонстрировать связь между корреляцией и прогнозированием, как этот. На рис. 5 в диапазоне В2:С12 представлена случайная выборка из десяти домов и приведены данные о площади каждого дома (в квадратных футах) и его продажной цене.

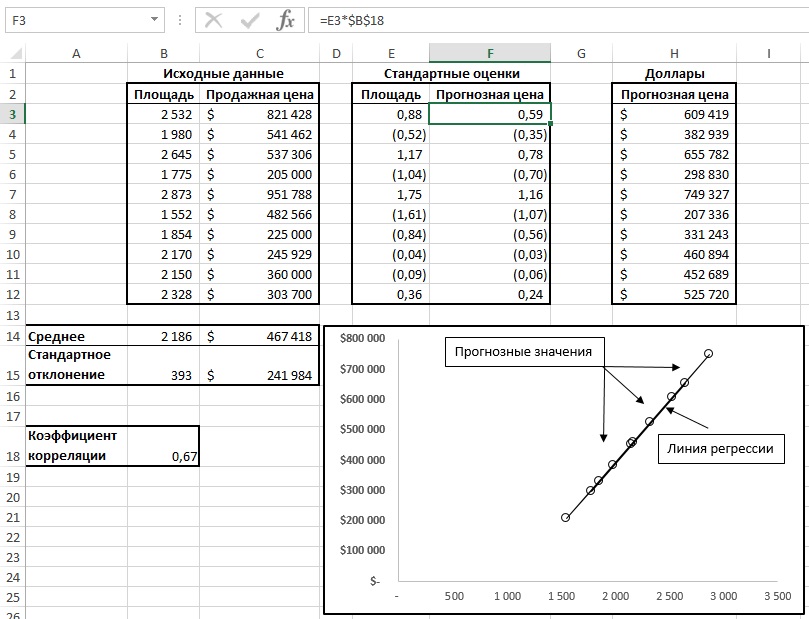


Рис. 5. Прогнозные значения продажной цены образуют прямую линию

Найдите средние значения, стандартные отклонения и коэффициент корреляции (диапазон А14:С18). Рассчитайте z-оценки площади (Е2:Е12). Например, ячейка ЕЗ содержит формулу: =(В3-$В$14)/$В$15. Вычислите z-оценки прогнозной цены (F2:F12). Например, ячейка F3 содержит формулу: =ЕЗ\*$В$18. Переведите z-оценки в цены в долларах (Н2:Н12). В ячейке НЗ формула: =F3\*$C$15+$C$14.

Обратите внимание: прогнозное значение всегда стремится сместиться в сторону среднего, равного 0. Чем ближе к нулю коэффициент корреляции, тем ближе к нулю прогнозная z-оценка. В нашем примере коэффициент корреляции между площадью и продажной ценой равен 0,67, и прогнозная цена равна 1,0\*0,67, т.е. 0,67. Этому соответствует превышение значения над средним значением, равное двум третям стандартного отклонения. Если бы коэффициент корреляции был равен 0,5, то прогнозная цена составила бы 1,0\*0,5, т.е. 0,5. Этому соответствует превышение значения над средним значением, равное лишь половине стандартного отклонения. Всякий раз, когда значение коэффициента корреляции отличается от идеального, т.е. больше -1,0 и меньше 1,0, оценка прогнозируемой переменной должна быть ближе к своему среднему значению, чем оценка предикторной (независимой) переменной к своему. Это явление называется регрессией к среднему, или просто регрессией.

В Excel есть несколько функций для определения коэффициентов уравнения линии регрессии (в Excel она называется линией тренда) *у = kx + b*. Для определения *k* служит функция

*=НАКЛОН(известные\_значения\_у; известные\_значения\_х)*

Здесь *у* – прогнозируемая переменная, а *х* – независимая переменная. Вы должны строго следовать этому порядку переменных. Наклон линии регрессии, коэффициент корреляции, стандартные отклонения переменных и ковариация тесно связаны между собой (рис. 6). Функция ОТРЕЗОК() возвращает значение, отсекаемое линией регрессии на вертикальной оси:

=ОТРЕЗОК(известные\_значения\_у; известные\_значения\_х)

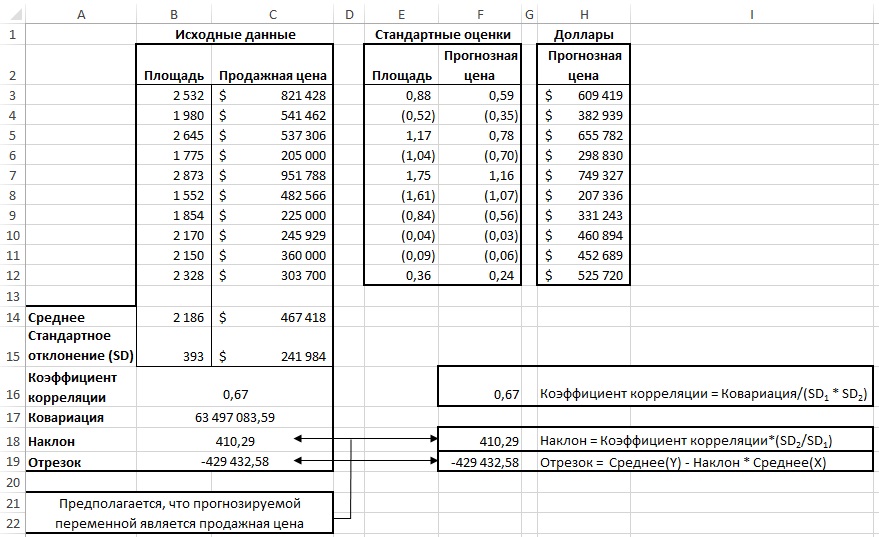


Рис. 6. Соотношение между стандартными отклонениями преобразует ковариацию в коэффициент корреляции и наклон линии регрессии

Обратите внимание, что количество значений х и у, предоставляемых функциям НАКЛОН() и ОТРЕЗОК() в качестве аргументов, должно быть одинаковым.

В регрессионном анализе используется еще один важный показатель – R2 (R-квадрат), или коэффициент детерминации. Он определяет, какой вклад в общую изменчивость данных вносит выявленная с помощью регрессии зависимость между *х* и *у*. В Excel для него есть функция КВПИРСОН(), которая принимает точно те же аргументы, что и функция КОРРЕЛ().

О двух переменных с ненулевым коэффициентом корреляции между ними говорят, что они объясняют дисперсию или имеют объясненную дисперсию. Обычно объясненная дисперсия выражается в процентах. Так *R2* = 0,81 означает, что 81% дисперсии (разброса) двух переменных является объясненной. Остальные 19% обусловлены случайными флуктуациями.

В Excel имеется функция ТЕНДЕНЦИЯ, которая упрощает вычисления. Функция ТЕНДЕНЦИЯ():

* принимает предоставляемые вами известные значения *х* и известные значения *у*;
* вычисляет наклон линии регрессии и константу (отрезок);
* возвращает прогнозные значения *у*, определяемые на основании применения уравнения регрессии к известным значениям *х* (рис. 7).

Функция ТЕНДЕНЦИЯ() является функцией массива (если вы ранее не сталкивались с такими функциями, рекомендую [Excel. Введение в формулы массива](http://baguzin.ru/wp/?p=3248)).

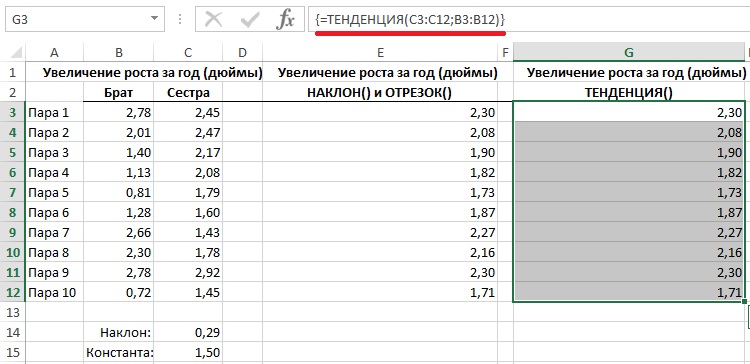


Рис. 7. Использование функции ТЕНДЕНЦИЯ() позволяет ускорить и упростить вычисления по сравнению с использованием пары функций НАКЛОН() и ОТРЕЗОК()

Чтобы ввести функцию ТЕНДЕНЦИЯ() в виде формулы массива в ячейки G3:G12, выделите диапазон G3:G12, введите формулу ТЕНДЕНЦИЯ (СЗ:С12;ВЗ:В12), нажмите и удерживайте клавиши <Ctrl+Shift> и только после этого нажмите клавишу <Enter>. Обратите внимание, что формула заключена в фигурные скобки: { и }. Так Excel сообщает вам о том, что данная формула воспринята именно как формула массива. Не вводите сами скобки: если вы попытаетесь ввести их самостоятельно в составе формулы, Excel воспримет ваш ввод как обычную текстовую строку.

У функции ТЕНДЕНЦИЯ() есть еще два аргумента: *новые\_значения\_х* и *конст*. Первый позволяет построить прогноз на будущее, а второй может заставить линию регрессии пройти через начало координат (значение ИСТИНА говорит Excel использовать расчетную константу, значение ЛОЖЬ – константу = 0). Excel позволяет нарисовать регрессионную прямую на графике так, чтобы она проходила через начало координат. Начните с построения точечной диаграммы, после чего щелкните правой кнопкой мыши на одном из маркеров ряда данных. Выберите в открывшемся контекстном меню пункт *Добавить линию тренда*; выберите вариант *Линейная*; при необходимости прокрутите панель вниз, установите флажок *Настроить пересечение*; убедитесь, что в связанном с ним текстовом поле задано значение 0,0.

Если у вас есть три переменных, и вы хотите определить корреляцию между двумя из них, исключив влияние третьей, можно использовать *частную корреляцию*. Предположим, вас интересует связь между процентной долей жителей города, закончивших колледж, и количеством книг в городских библиотеках. Вы собрали данные по 50 городам, но… Проблема в том, что оба этих параметра могут зависеть от благосостояния жителей того или иного города. Разумеется, очень трудно подобрать другие 50 городов, характеризующихся в точности одинаковым уровнем благосостояния жителей.

Применяя статистические методы для исключения влияния, оказываемого фактором благосостояния как на финансовую поддержку библиотек, так и на доступность обучения в колледже, вы могли бы получить более точную количественную оценку степени зависимости между интересующими вас переменными, а именно: количеством книг и количеством выпускников. Такая условная корреляция между двумя переменными, когда значения других переменных фиксированы, и называется частной корреляцией. Один из способов ее расчета заключается в использовании уравнения:

Где *rCB.W* — коэффициент корреляции между переменными Колледж (College) и Книги (Books) при исключенном влиянии (фиксированном значении) переменной Благосостояние (Wealth); *rCB* — коэффициент корреляции между переменными Колледж и Книги; *rCW* — коэффициент корреляции между переменными Колледж и Благосостояние; *rBW* — коэффициент корреляции между переменными Книги и Благосостояние.

С другой стороны, частную корреляцию можно рассчитать на основе анализа остатков, т.е. разностей между прогнозными значениями и связанными с ними результатами фактических наблюдений (оба метода представлены на рис. 8).

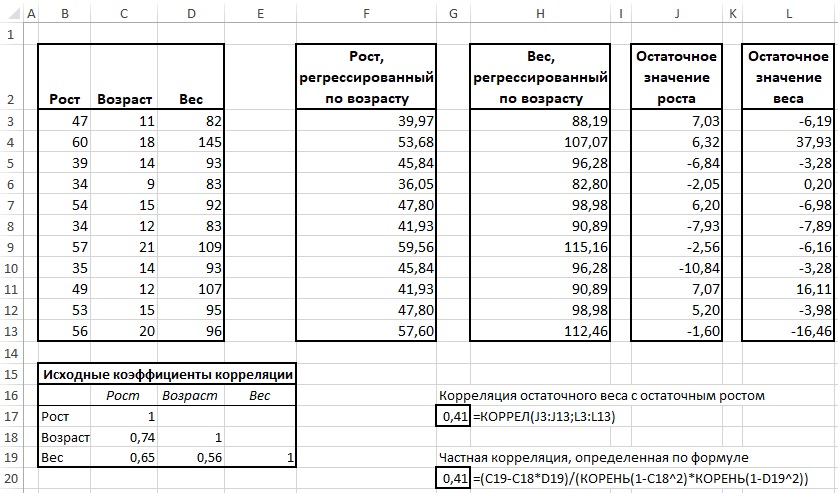


Рис. 8. Частная корреляция, как корреляция остатков

Для упрощения подсчета матрицы коэффициентов корреляции (В16:Е19) используйте пакет анализа Excel (меню *Данные* –> *Анализ* –> *Анализ данных*). По умолчанию этот пакет в Excel не активен. Для его установки пройдите по меню *Файл* –> *Параметры* –> *Надстройки*. Внизу открывшегося окна *Параметры Excel* найдите поле *Управление*, выберите *Надстройки Excel*, кликните *Перейти*. Поставьте галочку напротив надстройки *Пакет анализа*. Кликните А*нализ данных*, выберите опцию *Корреляция*. В качестве входного интервала укажите $B$2:$D$13, поставьте галочку *Метки в первой строке*, в качестве выходного интервала укажите $B$16:$E$19.

Еще одна возможность – определить получастную корреляцию. Например, вы исследуете влияние роста и возраста на вес. Таким образом, у вас две предикторные переменные – рост и возраст, и одна прогнозируемая переменная – вес. Вы хотите исключить влияние одной предикторной переменной на другую, но не на прогнозную переменную:

где Н – Рост (Height), W– Вес (Weight), А – Возраст (Age); в индексе получастного коэффициента корреляции используются круглые скобки, с помощью которых указывается, влияние какой переменной устраняется и из какой именно переменной. В данном случае обозначение W(Н.А) указывает на то, что влияние переменной Возраст удаляется из переменной Рост, но не из переменной Вес.

Может создаться впечатление, что обсуждаемый вопрос не имеет существенного значения. Ведь важнее всего то, насколько точно работает общее уравнение регрессии, тогда как проблема относительных вкладов отдельных переменных в суммарную объясненную дисперсию представляется второстепенной. Однако это далеко не так. Как только вы начинаете задумываться над тем, стоит ли вообще использовать какую-то переменную в уравнении множественной регрессии, проблема становится важной. Она может влиять на оценку правильности выбора модели для анализа.

### Глава 4. Функция ЛИНЕЙН()

Функция ЛИНЕЙН() возвращает 10 статистик регрессионного анализа. Функция ЛИНЕЙН() является функцией массива. Для ее ввода выделите диапазон, содержащий пять строк и два столбца, напечатайте формулу, и нажмите <Ctrl+Shift+Enter> (рис. 9):

=ЛИНЕЙН(B2:B21;A2:A21;ИСТИНА;ИСТИНА)

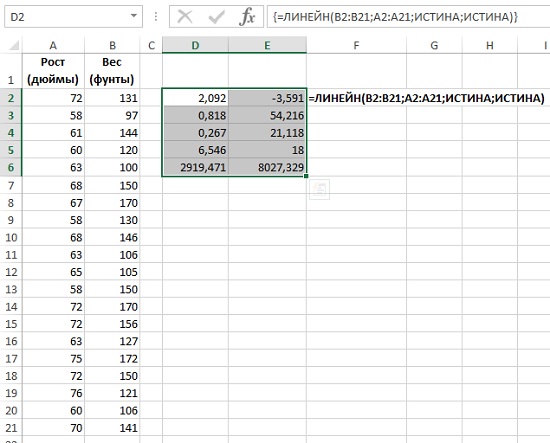


Рис. 9. Функция ЛИНЕЙН(): а) выделите диапазон D2:E6, б) введите формулу, как показано в строке формул, в) нажмите <Ctrl+Shift+Enter>

Функция ЛИНЕЙН() возвращает:

* коэффициент регрессии (или наклон, ячейка D2);
* отрезок (или константа, ячейка Е3);
* стандартные ошибки коэффициента регрессии и константы (диапазон D3:E3);
* коэффициент детерминации R2 для регрессии (ячейка D4);
* стандартная ошибка оценки (ячейка Е4);
* F-критерий для полной регрессии (ячейка D5);
* количество степеней свободы для остаточной суммы квадратов (ячейка Е5);
* регрессионная сумма квадратов (ячейка D6);
* остаточная сумма квадратов (ячейка Е6).

Рассмотрим каждую из этих статистик и их взаимодействие.

**Стандартная ошибка** в нашем случае – это стандартное отклонение, вычисляемое для ошибок выборки. Т.е., это ситуация, когда генеральная совокупность имеет одну статистику, а выборка – другую. Разделив коэффициент регрессии на стандартную ошибку, вы получите значение 2,092/0,818 = 2,559. Иными словами, коэффициент регрессии, равный 2,092, отстоит от нуля на две с половиной стандартные ошибки.

Если коэффициент регрессии равен нулю, то наилучшей оценкой прогнозируемой переменной является ее среднее значение. Две с половиной стандартные ошибки — это довольно большая величина, и вы с уверенностью можете полагать, что коэффициент регрессии для генеральной совокупности имеет ненулевое значение.

Можно определить вероятность получения выборочного коэффициента регрессии 2,092, если его фактическое значение в генеральной совокупности равно 0,0 с помощью функции

=СТЬЮДЕНТ.РАСП.ПХ(t-критерий = 2,559; количество степеней свободы =18)

В общем количество степеней свободы = n – k – 1, где n — количество наблюдений, а k — количество предикторных переменных.

Эта формула возвращает значение 0,00987 или, округленно, 1%. Оно сообщает нам следующее: если коэффициент регрессии для генеральной совокупности равен 0%, то вероятность получения выборки из 20 человек, для которой расчетное значение коэффициента регрессии равно 2,092, составляет скромный 1%.

F-критерий (ячейка D5 на рис. 9) выполняет те же функции по отношению к полной регрессии, что и t-критерий по отношению к коэффициенту простой парной регрессии. F-критерий используется для проверки того, действительно ли коэффициент детерминации R2 для регрессии имеет достаточно большую величину, позволяющую отбросить гипотезу о том, что в генеральной совокупности он имеет значение 0,0, которое указывает на отсутствие дисперсии, объясняемой предикторной и прогнозируемой переменной. При наличии только одной предикторной переменной F-критерий в точности равен квадрату t-критерия.

До сих пор мы рассматривали интервальные переменные. Если же у вас переменные, которые могут принимать несколько значений, представляющих собой простые имена, например, Мужчина и Женщина или Пресмыкающееся, Земноводное и Рыба, представьте их в виде числового кода. Такие переменные называются номинальными.

**Статистика R2** дает количественную оценку доли объясненной дисперсии.

**Стандартная ошибка оценки.** На рис. 4.9 представлены прогнозные значения переменной Вес, полученные на основании ее связи с переменной Рост. В диапазоне Е2:Е21 содержатся значения остатков для переменной Вес. Точнее эти остатки называть ошибками — отсюда и следует термин стандартная ошибка оценки.

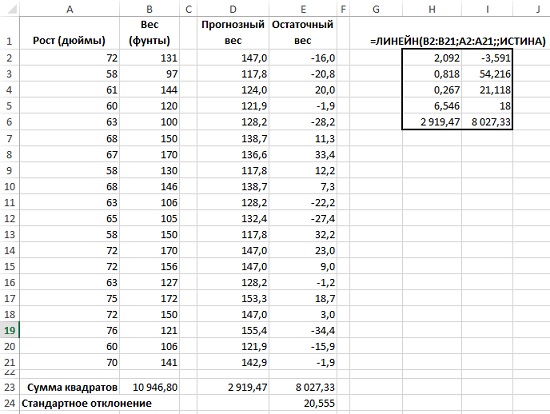


Рис. 10. Как R2, так и стандартная ошибка оценки выражают точность прогнозов, получаемых с помощью регрессии

Чем меньше стандартная ошибка оценки, тем точнее уравнение регрессии и тем более близкого совпадения любого прогноза, полученного с помощью уравнения, с фактическим наблюдением вы ожидаете. Стандартная ошибка оценки предоставляет способ количественной оценки этих ожиданий. Вес 95% людей, обладающих неким ростом, будет находиться в диапазоне:

(рост \* 2,092 – 3,591) ± 2,092\*21,118

**F-статистика** – это отношение межгрупповой дисперсии к внутригрупповой дисперсии. Это название было введено статистиком Джорджем Снедекором в честь сэра [Рональда Фишера](http://baguzin.ru/wp/?p=15657), разработавшего в начале XX столетия дисперсионный анализ (ANOVA, Analysis of Variance).

Коэффициент детерминации R2 выражает долю общей суммы квадратов, связанную с регрессией. Величина (1 – R2) выражает долю общей суммы квадратов, связанную с остатками — ошибками прогнозирования. F-критерий можно получить с использованием функции ЛИНЕЙН (ячейка F5 на рис. 11), с использованием сумм квадратов (диапазон G10:J11), с использованием долей дисперсии (диапазон G14:J15). Формулы можно изучить в прилагаемом файле Excel.

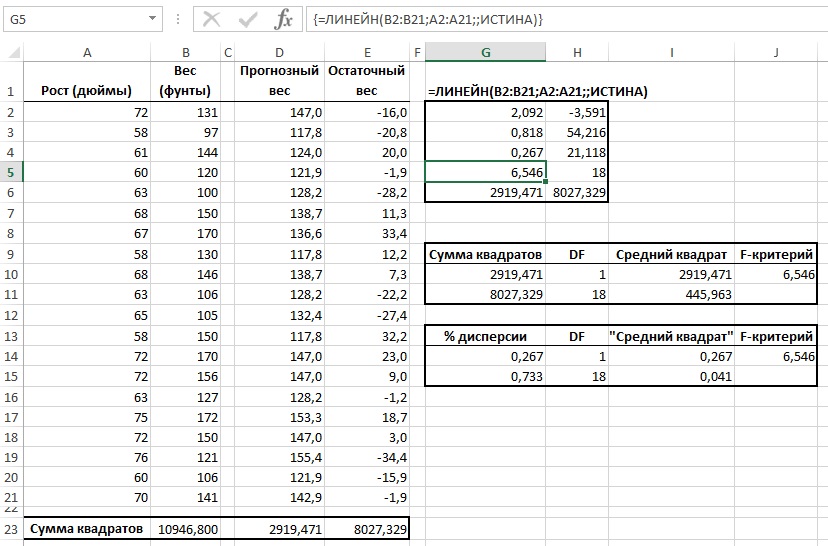


Рис. 11. Расчет F-критерия

При использовании номинальных переменных используется фиктивное кодирование (рис. 12). Для кодирования значений удобно использовать значения 0 и 1. Вероятность F рассчитывается с помощью функции:

=F.РАСП.ПХ(К2;I2;I3)

Здесь функция F.РАСП.ПХ() возвращает вероятность получения F-критерия, подчиняющегося центральному F-распределению (рис. 13) для двух наборов данных с количествами степеней свободы, приведенными в ячейках I2 и I3, значение которого совпадает со значением, приведенным в ячейке К2.

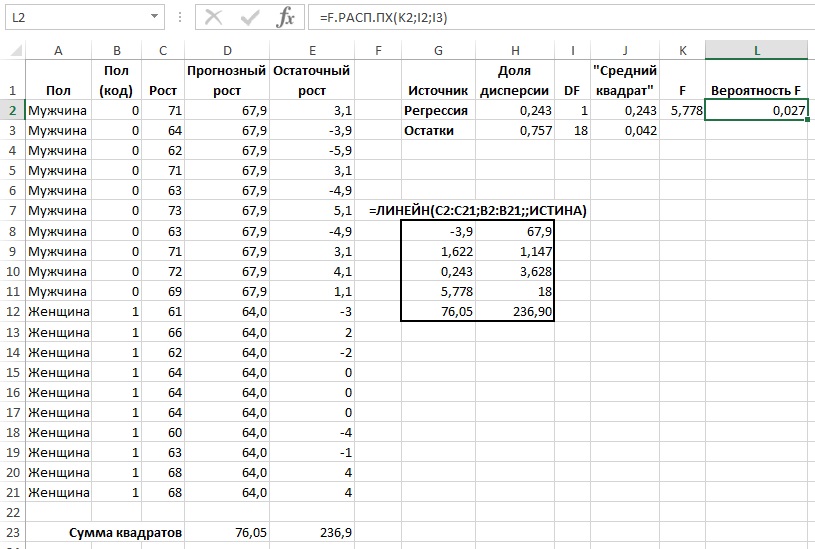


Рис. 12. Регрессионный анализ с использованием фиктивных переменных

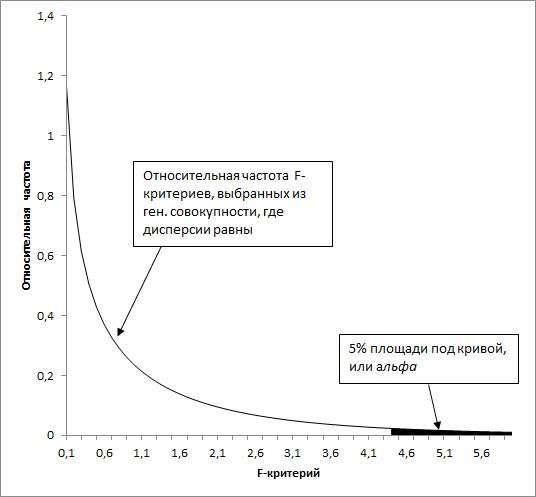


Рис. 13. Центральное F-распределение при λ = 0

### Глава 5. Множественная регрессия

Переходя от простой парной регрессии с одной предикторной переменной к множественной регрессии, вы добавляете одну или несколько предикторных переменных. Сохраняйте значения предикторных переменных в смежных столбцах, например, в столбцах А и В в случае двух предикторов или А, В и С в случае трех предикторов. Прежде чем вводить формулу, включающую функцию ЛИНЕЙН(), выберите пять строк и столько столбцов, сколько имеется предикторных переменных, плюс еще один для константы. В случае регрессии с двумя предикторными переменными можно использовать следующую структуру:

=ЛИНЕЙН(А2: А41; В2: С41;;ИСТИНА)

Точно так же в случае трех переменных:

=ЛИНЕЙН(А2:А61;В2:D61;;ИСТИНА)

Предположим, вы хотите изучить возможное влияние возраста и диеты на содержание ЛПНП — липопротеинов низкой плотности, которые считаются ответственными за образование атеросклеротических бляшек, служащих причиной атеротромбоза (рис. 14).

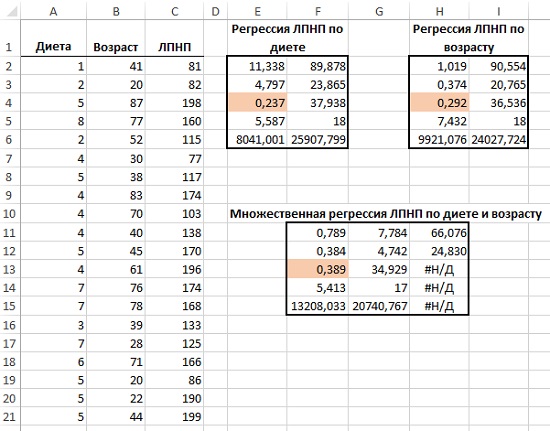


Рис. 14. Множественная регрессия

R2 множественной регрессии (отражаемый в ячейке F13), больше, чем R2 любой простой регрессии (Е4, Н4). В множественной регрессии одновременно используются несколько предикторных переменных. При этом R2 почти всегда увеличивается.

Для любого простого линейного уравнения регрессии с одной предикторной переменной между прогнозными значениями и значениями предикторной переменной всегда будет наблюдаться идеальная корреляция, поскольку в таком уравнении значения предиктора умножаются на одну константу и к каждому произведению прибавляется другая константа. Этот эффект не сохраняется во множественной регрессии.

Отображение результатов, возвращаемых функцией ЛИНЕЙН() для множественной регрессии (рис. 15). Коэффициенты регрессии выводятся в составе результатов, возвращаемых функцией ЛИНЕЙН() *в порядке обратном расположению переменных* (G–H–I соответствует С–В–А).

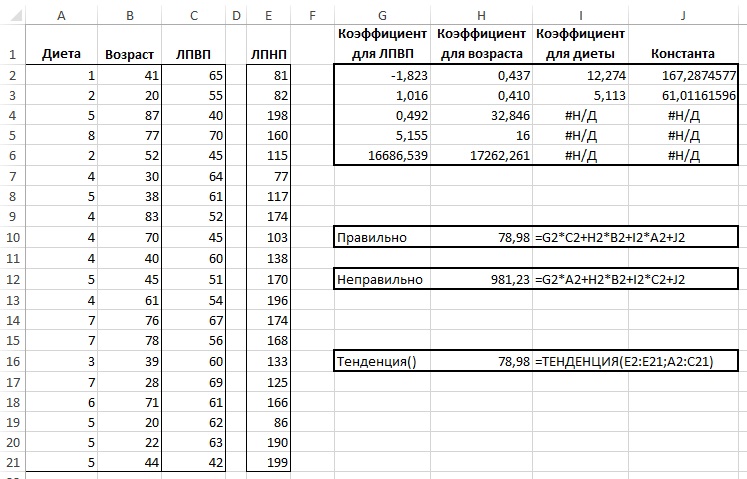


Рис. 15. Коэффициенты и их стандартные ошибки отображаются в обратном порядке их следования на рабочем листе

Принципы и процедуры, используемые в регрессионном анализе с одной предикторной переменной, легко адаптируются для учета нескольких предикторных переменных. Оказывается, что многое в этой адаптации зависит от устранения влияния предикторных переменных друг на друга. Последнее связано с частной и получастной корреляциями (рис. 16).



Рис. 16. Множественная регрессия может быть выражена через парную регрессию остатков (формулы см. в Excel-файле)

В Excel, имеются функции, предоставляющие информацию о t- и F-распределениях. Функции, имена которых включают часть РАСП, такие как СТЬЮДЕНТ.РАСП() и F.РАСП(), принимают t- или F-критерий в качестве аргумента и возвращают вероятность наблюдения указанного значения. Функции, имена которых включают часть ОБР, такие как СТЬЮДЕНТ.ОБР() и F.ОБР(), принимают значение вероятности в качестве аргумента и возвращают значение критерия, соответствующее указанной вероятности.

Поскольку мы ищем критические значения t-распределения, которые отсекают края его хвостовых областей, мы передаем 5% в качестве аргумента одной из функций СТЬЮДЕНТ.ОБР(), которая возвращает значение, соответствующее этой вероятности (рис. 17, 18).

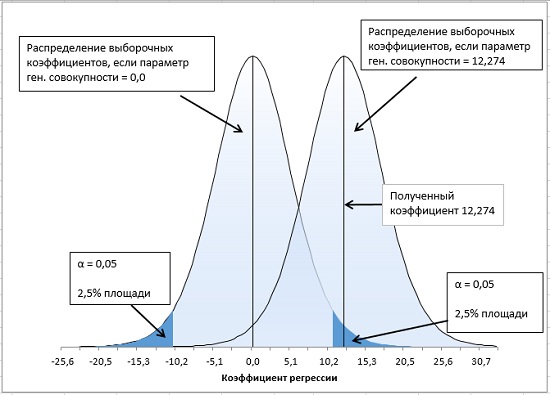


Рис. 17. Двусторонний t-тест

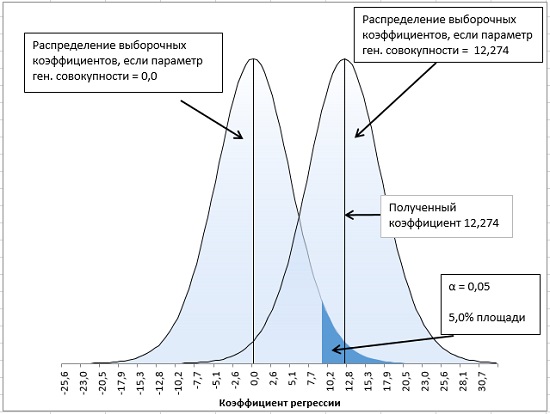


Рис. 18. Односторонний t-тест

Устанавливая правило принятия решений в случае однохвостовой альфа-области, вы увеличиваете статистическую мощность теста. Если, приступая к эксперименту, вы уверены в том, что у вас есть все основания ожидать получения положительного (или отрицательного) коэффициента регрессии, то вам следует выполнить однохвостовой тест. В этом случае вероятность того, что вы принимаете правильное решение, отвергая гипотезу о нулевом коэффициенте регрессии в генеральной совокупности, будет выше.

Статистики предпочитают использовать термин *направленный тест* вместо термина *однохвостовой тест* и термин *ненаправленный тест* вместо термина *двуххвостовой тест*. Термины направленный и ненаправленный предпочтительнее, поскольку делают акцент на типе гипотезы, а не на природе хвостов распределения.

**Подход к оценке влияния предикторов, основанный на сравнении моделей.** На рис. 19 представлены результаты регрессионного анализа, в котором тестируется вклад переменной Диета в уравнение регрессии.

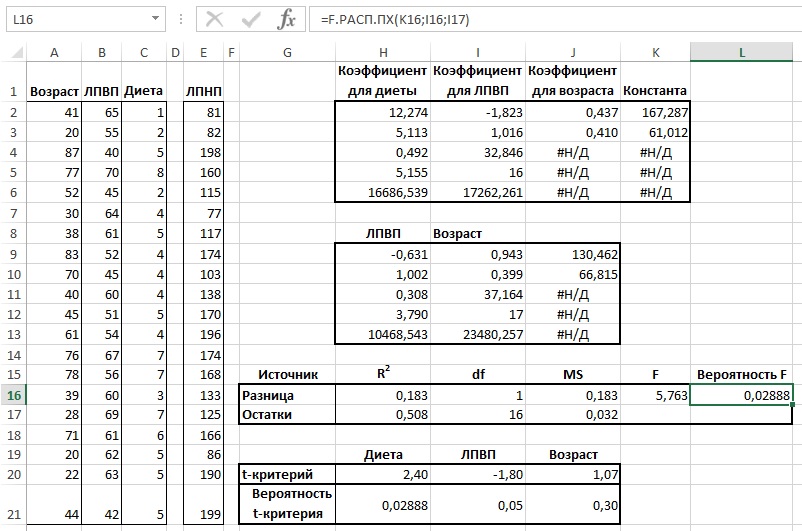


Рис. 19. Сравнение двух моделей путем проверки различий в их результатах

Результаты функции ЛИНЕЙН() (диапазон Н2:К6) имеют отношение к тому, что я называю полной моделью, в которой выполняется регрессия переменной ЛПНП по переменным Диета, Возраст и ЛПВП. В диапазоне Н9:J1З представлены расчеты без учета предикторной переменной Диета. Я называю это ограниченной моделью. В полной модели 49,2% дисперсии зависимой переменной ЛПНП объясняется предикторными переменными. В ограниченной модели лишь 30,8% ЛПНП объясняется переменными Возраст и ЛПВП. Потеря R2, обусловленная исключением переменной Диета из модели, составляет 0,183. В диапазоне G15:L17 сделаны расчеты, которые показывают, что лишь с вероятностью 0,0288 влияние переменной Диета является случайным. В остальных 97,1% Диета оказывает влияние на ЛПНП.

### Глава 6. Допущения и предостережения в отношении регрессионного анализа

Термин «допущение» не определен достаточно строго, а способ его использования предполагает, что если допущение не соблюдается, то результаты всего анализа являются по меньшей мере сомнительными или, возможно, не имеющими силы. На самом деле это не так, хотя, безусловно, существуют случаи, когда нарушение допущения в корне меняет картину. Основные допущения: а) остатки переменной Y нормально распределены в любой точке X вдоль линии регрессии; б) значения Y находятся в линейной зависимости от значений X; в) дисперсия остатков примерно одинакова в каждой точке Х; г) между остатками отсутствует зависимость.

Если допущения не играют существенной роли, статистики говорят о робастности анализа по отношению к нарушению допущения. В частности, когда вы используете регрессию для тестирования различий между групповыми средними, допущение о том, что значения Y — а значит, и остатки — нормально распределены, не играет существенной роли: тесты робастны по отношению к нарушению допущения о нормальности. При этом важно анализировать данные с помощью диаграмм. Например, включенных в надстройку *Анализ данных* инструмент *Регрессия*.

Если данные не соответствуют допущениям линейной регрессии, в вашем распоряжении имеются другие подходы, отличные от линейного. Один из них – логистическая регрессия (рис. 20). Вблизи верхнего и нижнего предельных значений предикторной переменной линейная регрессия приводит к нереалистичным прогнозам.

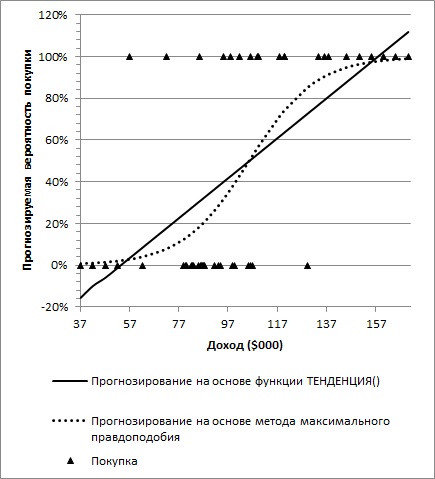


Рис. 20. Логистическая регрессия

На рис. 6.8 отображены результаты двух методов анализа данных, направленного на исследование связи между ежегодным доходом и вероятностью покупки дома. Очевидно, вероятность совершения покупки будет увеличиваться с увеличением дохода. Диаграммы упрощают выявление различий между результатами, прогнозирующими вероятность покупки дома посредством линейной регрессии, и результатами, которые вы могли бы получить, используя другой подход.

В основе линейного (а в данном контексте — и криволинейного) регрессионного анализа лежат корреляции. Фундамент логистической регрессии образуют шансы и отношения шансов. Несмотря на использование слова регрессия в названиях обоих методов, между этими подходами имеются значительные различия. В частности, логистическая регрессия не обязательно дает прямую регрессионную линию, как в случае линейной регрессии. Это свойство логистической регрессии означает, что ее линия максимального правдоподобия, которой на рис. 20 соответствует S-образная точечная кривая, меняет свой наклон при перемещении вдоль континуума значений X (в данном случае — дохода). Поэтому линия логистической регрессии никогда не доходит до отрицательных значений вероятности покупки в области нижней границы шкалы дохода и до значений вероятности свыше 100% в области верхней границы шкалы (см. также [Использование модели Раша в пересчете баллов ЕГЭ](http://baguzin.ru/wp/?p=3370)).

На языке статистиков отбрасывание нулевой гипотезы, когда в действительности она является истинной, называется ошибкой I рода.

В надстройке *Анализ данных* предлагается удобный инструмент для генерации случайных чисел, предоставляющий пользователю возможность задать желаемую форму распределения (например, Нормальное, Биномиальное или Пуассона), а также среднее значение и стандартное отклонение.

**Различия между функциями семейства СТЬЮДЕНТ.РАСП().** Начиная с версии Excel 2010 доступны три разные формы функции, возвращающей долю распределения слева и/или справа от заданного значения t-критерия. Функция СТЬЮДЕНТ.РАСП() возвращает долю площади под кривой распределения слева от указанного вами значения t-критерия. Предположим, у вас имеется 36 наблюдений, и поэтому количество степеней свободы для анализа равно 34, а значение t-критерия = 1,69. В этом случае формула

=СТЬЮДЕНТ.РАСП(+1,69;34;ИСТИНА)

возвращает значение 0,05, или 5% (рис. 21). Третий аргумент функции СТЬЮДЕНТ.РАСП() может иметь значение ИСТИНА или ЛОЖЬ. Если он задан равным ИСТИНА, функция возвращает кумулятивную площадь под кривой слева от заданного t-критерия, выраженную в виде доли. Если же он равен ЛОЖЬ, функция возвращает относительную высоту кривой в точке, соответствующей t-критерию. Другие версии функции СТЬЮДЕНТ.РАСП() — СТЬЮДЕНТ.РАСП.ПХ() и СТЬЮДЕНТ.РАСП.2Х() — принимают в качестве аргументов только значение t-критерия и количество степеней свободы и не требуют задания третьего аргумента.

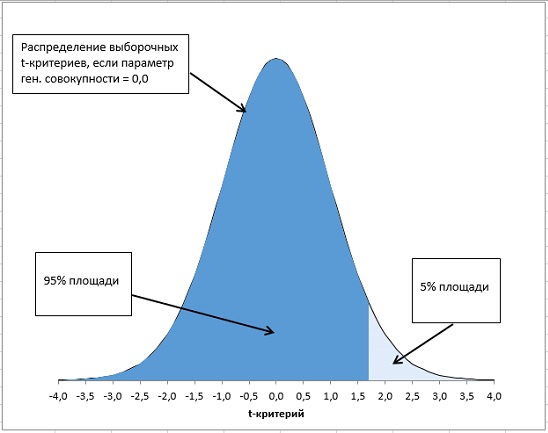


Рис. 21. Более темная затененная область в левом хвосте распределения соответствует доле площади под кривой слева от большого положительного значения t-критерия

Чтобы определить площадь справа от t-критерия используйте одну из формул:

=1 - СТЫОДЕНТ.РАСП (1, 69;34;ИСТИНА)

=СТЬЮДЕНТ.РАСП.ПХ(1,69;34)

Вся площадь под кривой должна составлять 100%, поэтому вычитание из 1 доли площади слева от значения t-критерия, которую возвращает функция, дает долю площади, располагающейся справа от значения t-критерия. Возможно, вам покажется более предпочтительным вариант непосредственного получения интересующей вас доли площади с помощью функции СТЬЮДЕНТ.РАСП.ПХ(), где ПХ означает правый хвост распределения (рис. 22).

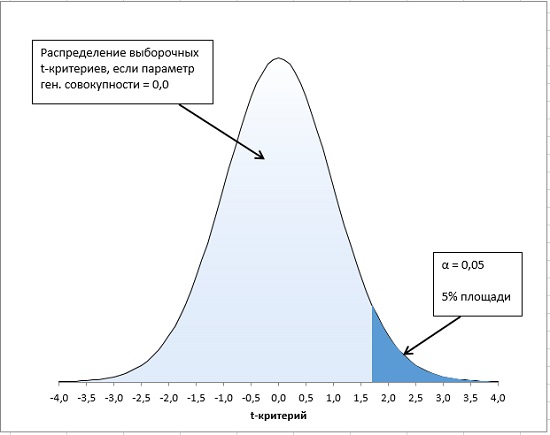


Рис. 22. 5%-ная альфа область для направленного теста

Использование функций СТЬЮДЕНТ.РАСП() или СТЬЮДЕНТ.РАСП.ПХ () подразумевает, что вы выбрали направленную рабочую гипотезу. Направленная рабочая гипотеза в сочетании с установкой значения альфа на уровне 5% означает, что вы помещаете все 5% в правый хвост распределениями. Вы должны будете отвергнуть нулевую гипотезу лишь в том случае, если вероятность полученного вами значения t-критерия составит 5% и менее. Направленные гипотезы обычно приводят к более чувствительным статистическим тестам (эту большую чувствительность также называют большей статистической мощностью).

При ненаправленном тесте значение альфа остается на том же уровне 5%, но распределение будет иным. Поскольку вы должны допускать два исхода вероятность ложноположительного результата должна быть распределена между двумя хвостами распределения. Общепринято распределять эту вероятность поровну (рис. 23).

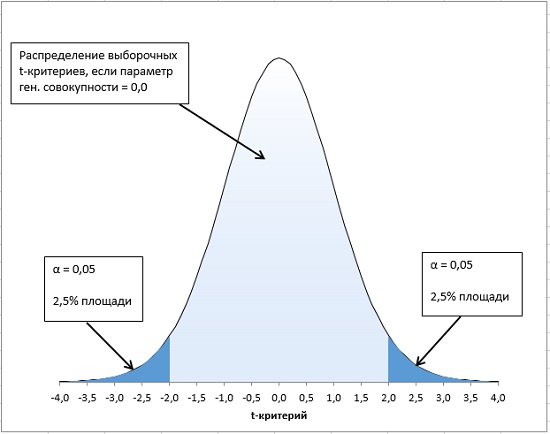


Рис. 23. Ненаправленный тест

Используя то же самое полученное значение t-критерия и то же количество степеней свободы, что и в предыдущем примере, воспользуйтесь формулой

=СТЬЮДЕНТ.РАСП.2Х(1,69;34)

Без каких-либо особых на то причин функция СТЬЮДЕНТ.РАСП.2Х() возвращает код ошибки #ЧИСЛО!, если в качестве первого аргумента ей предоставляется отрицательное значение t-критерия.

Если выборки содержат разное число данных, воспользуйтесь двухвыборочным t-тестом с различными дисперсиями, включенным в пакет *Анализ данных*.

### Глава 7. Использование регрессии для тестирования различий между групповыми средними

Переменные, которые ранее фигурировали под названием прогнозируемых переменных, в этой главе будут называться результативными переменными, а вместо термина предикторные переменные будет использоваться термин факторные переменные.

Простейшим из подходов к кодированию номинальной переменной является **фиктивное кодирование** (рис. 24).

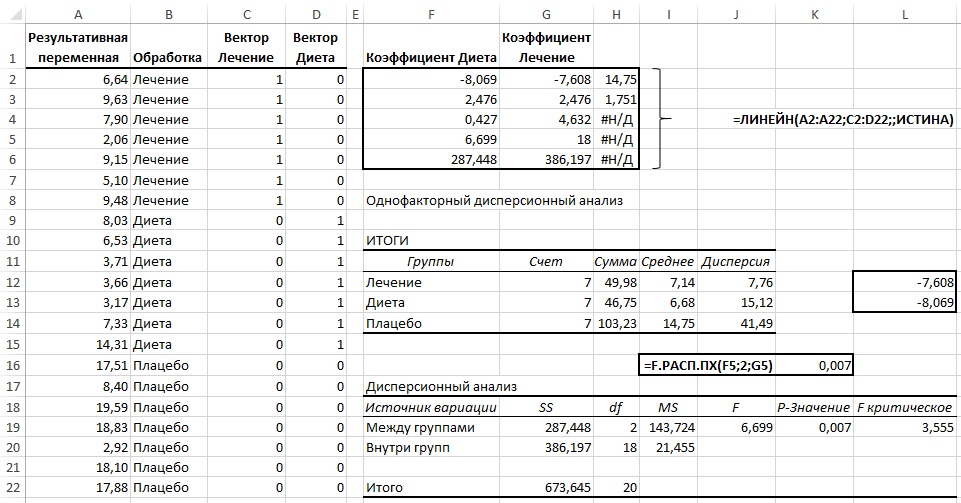


Рис. 24. Регрессионный анализ на основе фиктивного кодирования

При использовании фиктивного кодирования любого рода следует придерживаться правил:

* Количество столбцов, резервируемых для новых данных, должно быть равным количеству уровней фактора минус 1.
* Каждый вектор представляет один уровень фактора.
* Субъекты одного из уровней, которым часто является контрольная группа, получают код 0 во всех векторах.

Формула в ячейках F2:H6 =ЛИНЕЙН(A2:A22;C2:D22;;ИСТИНА) возвращает регрессионные статистики. Для сравнения на рис. 24 отображены результаты традиционного дисперсионного анализа, возвращаемого инструментом *Однофакторный дисперсионный анализ* надстройки *Анализ данных*.

**Кодирование эффектов.** В другом типе кодирования, получившем название *кодирование эффектов,* среднее каждой группы сравнивается со средним групповых средних. Этот аспект кодирования эффектов обусловлен использованием значения -1 вместо 0 в качестве кода для группы, которая получает один и тот же код во всех кодовых векторах (рис. 25).

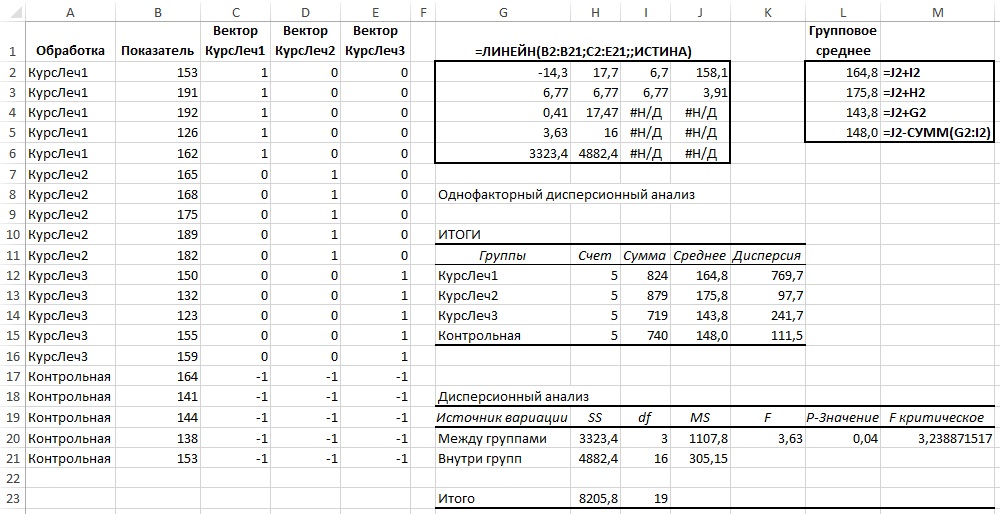


Рис. 25. Кодирование эффектов

Когда используется фиктивное кодирование, значение константы, возвращаемое функцией ЛИНЕЙН(), совпадает со средним группы, которой во всех векторах назначены нулевые коды (обычно это контрольная группа). В случае кодирования эффектов константа равна общему среднему (ячейка J2).

Общая линейная модель — полезный способ концептуализации компонентов значения результирующей переменной:

Yij = μ + αj + εij

Использование в этой формуле греческих букв вместо латинских подчеркивает тот факт, что она относится к генеральной совокупности, из которой извлекаются выборки, но ее можно переписать в виде, указывающем на то, что она относится к выборкам, извлекаемым изданной генеральной совокупности:

Yij = Y̅ + aj + eij

Идея состоит в том, что каждое наблюдение Yij можно рассматривать как сумму следующих трех компонентов: общее среднее, μ; эффект обработки j, аj; величина eij, которая представляет отклонение индивидуального количественного показателя Yij от комбинированного значения общего среднего и эффекта j-й обработки (рис. 26). Целью уравнения регрессии является минимизация суммы квадратов остатков.

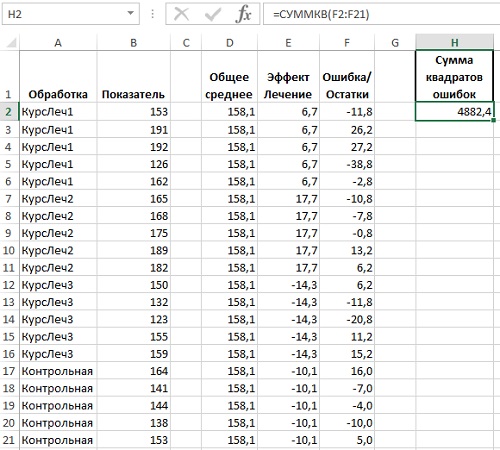


Рис. 26. Наблюдения, разложенные на компоненты общей линейной модели

**Факторный анализ.** Если исследуется связь между результативной переменной и одновременно двумя или более факторами, то в этом случае говорят об использовании факторного анализа. Добавление одного или нескольких факторов в однофакторный дисперсионный анализ может увеличивать статистическую мощность. В однофакторном дисперсионном анализе вариация результативной переменной, которая не может быть приписана фактору, включается в остаточный средний квадрат. Но вполне может быть так, что эта вариация с вязана с другим фактором. Тогда эта вариация может быть удалена из среднеквадратической ошибки, уменьшение которой приводит к увеличению значений F-критерия, а значит, к увеличению статистической мощности теста. Надстройка *Анализ данных* включает инструмент, обеспечивающий обработку двух факторов одновременно (рис. 27).

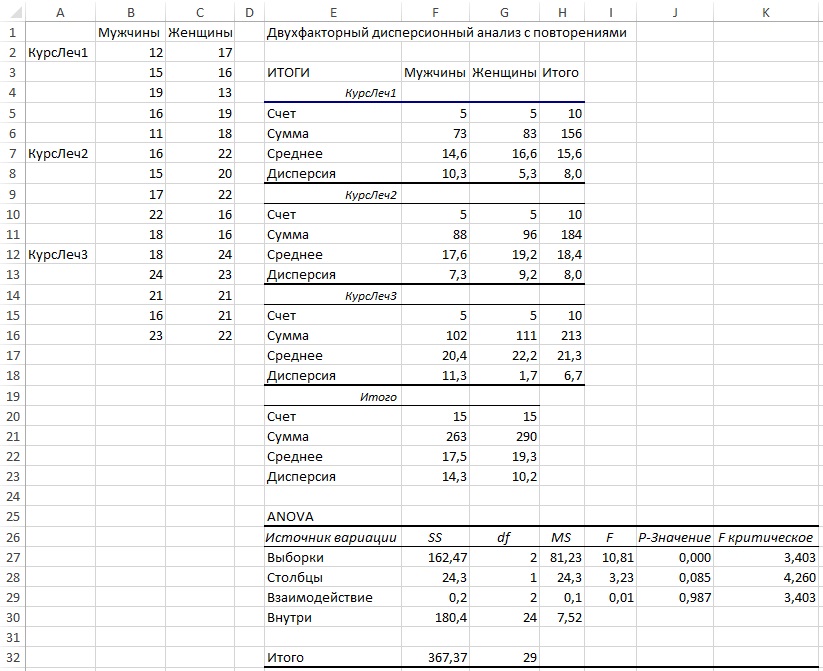


Рис. 27. Инструмент Двухфакторный дисперсионный анализ с повторениями Пакета анализа

Использованный на этом рисунке инструмент дисперсионного анализа, полезен тем, что он возвращает среднее и дисперсию результативной переменной, а также значение счетчика для каждой группы, включенной в план. В таблице *Дисперсионный анализ* отображаются два параметра, отсутствующие в выходной информации однофакторной версии инструмента дисперсионного анализа. Обратите внимание на источники вариации *Выборка* и *Столбцы* в строках 27 и 28. Источник вариации *Столбцы* относится к полу. Источник вариации *Выборка* относится к любой переменной, значения которой занимают различные строки. На рис. 27 значения для группы КурсЛеч1 находятся в строках 2-6, группы КурсЛеч2 — в строках 7-11, а группы КурсЛечЗ — в строках 12-16.

Главный момент заключается в том, что оба фактора, Пол (подпись Столбцы в ячейке Е28) и Лечение (подпись Выборка в ячейке Е27), включены в таблицу Дисперсионный анализ как источники вариации. Средние для мужчин отличаются от средних для женщин, и это создает источник вариации. Средние для трех видов лечения также различаются — вот вам еще один источник вариации. Существует также третий источник — Взаимодействие, который относится к объединенному эффекту переменных Пол и Лечение.

### Глава 8. Ковариационный анализ

Ковариационный анализ, или ANCOVA (Analysis of Covariation) уменьшает смещения и увеличивает статистическую мощность. Напомню, что одним из способов оценки надежности регрессионного уравнения являются F-тесты:

F = MSRegression/MSResidual

где MS (Mean Square) — средний квадрат, а индексы Regression и Residual указывают на регрессионную и остаточную компоненты соответственно. Расчет MSResidual выполняется по формуле:

MSResidual = SSResidual / dfResidual

где SS (Sum of Squares) — сумма квадратов, a df – количество степеней свободы. Когда вы добавляете ковариацию в уравнение регрессии, некоторая доля общей суммы квадратов включается не в SSResiduaI, а в SSRegression. Это приводит к уменьшению SSResidual, а значит, и MSResidual. Чем меньше MSResidual, тем больше F-критерий и тем вероятнее, что вы отвергнете нулевую гипотезу об отсутствии различий между средними. В результате вы перераспределяете изменчивость результативной переменной. В ANOVA, когда ковариация не учитывается, изменчивость переходит в ошибку. Но в ANCOVA часть изменчивости, ранее относившаяся к ошибке, назначается ковариате и становится частью SSRegression.

Рассмотрим пример, в котором один и тот же набор данных анализируется сначала с помощью ANOVA, а затем с помощью ANCOVA (рис. 28).

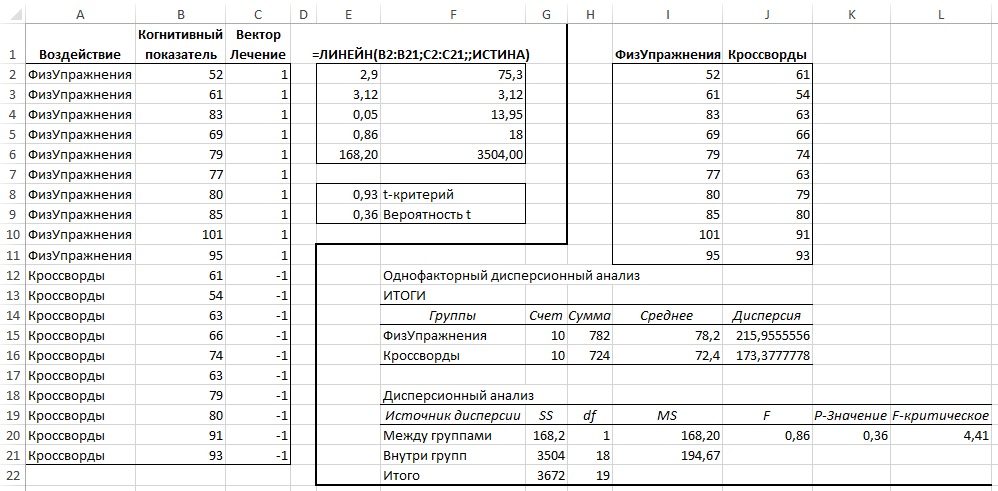


Рис. 28. Анализ ANOVA указывает на то, что результаты, полученные с помощью уравнения регрессии, ненадежны

В исследовании сравниваются относительные эффекты физических упражнений, развивающих мышечную силу, и когнитивных упражнений (разгадывание кроссвордов), активизирующих мозговую деятельность. Субъекты были случайным образом распределены по двум группам, чтобы в начале эксперимента обе группы находились в одинаковых условиях. По прошествии трех месяцев были измерены когнитивные характеристики субъектов. Результаты этих измерений приведены в столбце В.

В диапазоне А2:С21 размещены исходные данные, передаваемые функции ЛИНЕЙН() для выполнения анализа с использованием кодирования эффектов. Результаты работы функции ЛИНЕЙН() приведены в диапазоне E2:F6, где в ячейке Е2 отображается коэффициент регрессии, связанный с вектором воздействия. В ячейке Е8 содержится t-критерий = 0,93, а в ячейке Е9 тестируется надежность этого t-критерия. Содержащееся в ячейке Е9 значение говорит о том, что вероятность встретить различие между групповыми средними, наблюдаемое в данном эксперименте, составляет 36%, если в генеральной совокупности групповые средние равны. Лишь немногие признают этот результат статистически значимым.

На рис. 29 показано, что произойдет при добавлении ковариаты в анализ. В данном случае я добавил в набор данных возраст каждого субъекта. Коэффициент детерминации R2 для уравнения регрессии, в котором используется ковариата, равен 0,80 (ячейка F4). Значение R2 в диапазоне F15:G19, в котором я воспроизвел результаты ANOVA, полученные без использования ковариаты, равно всего лишь 0,05 (ячейка F17). Следовательно, уравнение регрессии, включающее ковариату, предсказывает значения переменной Когнитивный показатель намного точнее, чем с использованием только вектора Воздействие. Для ANCOVA вероятность случайного получения значения F-критерия, отображаемого в ячейке F5, равна менее чем 0,01%.

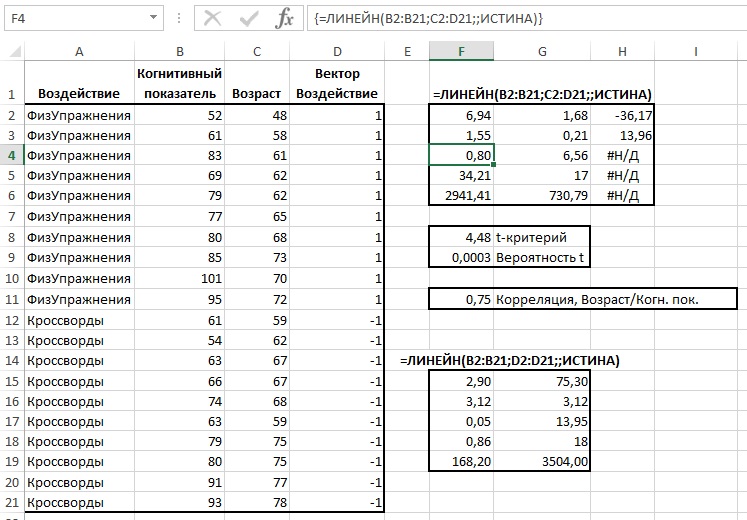


Рис. 29. ANCOVA возвращает совершенно иную картину