# 3. Линейная регрессия (linear regression)

## Постановка задачи линейной регрессии. Вероятностная интерпретация.

### Постановка задачи

Для любых вычислений удобнее работать не с дискретными данными, а с общим законом, определяющим их. В случае с отдельными точками в пространстве координат стараются воссоздать функциональную зависимость. Если точки получены в эксперименте, они неизбежно содержат ошибку измерений, шум, поэтому разумнее требовать, чтобы функция передавала общую тенденцию, а не точно проходила через все точки. Для этих целей применяется регрессия.

В случае с линейной регрессией стараются найти уравнение вида:  
   
для которого отклонение от действительных данных .

В матричном представлении .

Цель регрессии — найти коэффициенты этой линейной комбинации , и тем самым определить регрессионную функцию (которую также называют моделью).

### Вероятностная интерпретация

Пусть каждая точка обладает скрытой характеристикой . Тогда точки распределяются с плотностью . В большинстве случаев ожидается, что распределена нормально относительно некоторого мат.ожидания .

Для поиска зависимости используется метод максимальной правдоподобности  
.  
Упрощая формулу, , что эквивалентно методу наименьших квадратов.

По формуле условной вероятности .  
При получаем .

## Метод наименьших квадратов. Алгебраическое и оптимизационное решения.

Метод наименьших квадратов – математический метод, применяемый для решения различных задач, основанный на минимизации суммы квадратов отклонений некоторых функций от искомых переменных.  
.

В общем случае решение сводится к решению системы уравнений с производными:  
. Откуда . В матричной форме:  
.

## Ковариация, корреляция.

Ковариация – мера линейной зависимости двух случайных величин.  
.  
Эта характеристика может быть сколь угодно большой по модулю и быть как положительной, так и отрицательной. Чем ближе значение к нулю, тем меньше зависимость между величинами оказывается линейной.

Корреляция – статистическая взаимосвязь нескольких случайных величин. Мерой корреляции является коэффициент корреляции. Он соответствует нормированной ковариации: . После нормирования характеристика может принимать значения от -1 до 1.

Для характеристики линейной корреляционной связи между одной случайной величиной и некоторым множеством случайных величин используется множественный коэффициент корреляции. Он вычисляется с помощью корреляционной матрицы:  
, где – определитель матрицы, – алгебраическое дополнение элемента .  
В общем случае сравнивается величина и её линейная аппроксимация .

## Коэффициент деретминации (критерий ).

Коэффициент детерминации – доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости. Можно рассматривать как универсальную меру зависимости одной случайной величины от множества других.  
.

В случае выборочной оценки значений дисперсий используется выборочный коэффициент детерминации:  
.

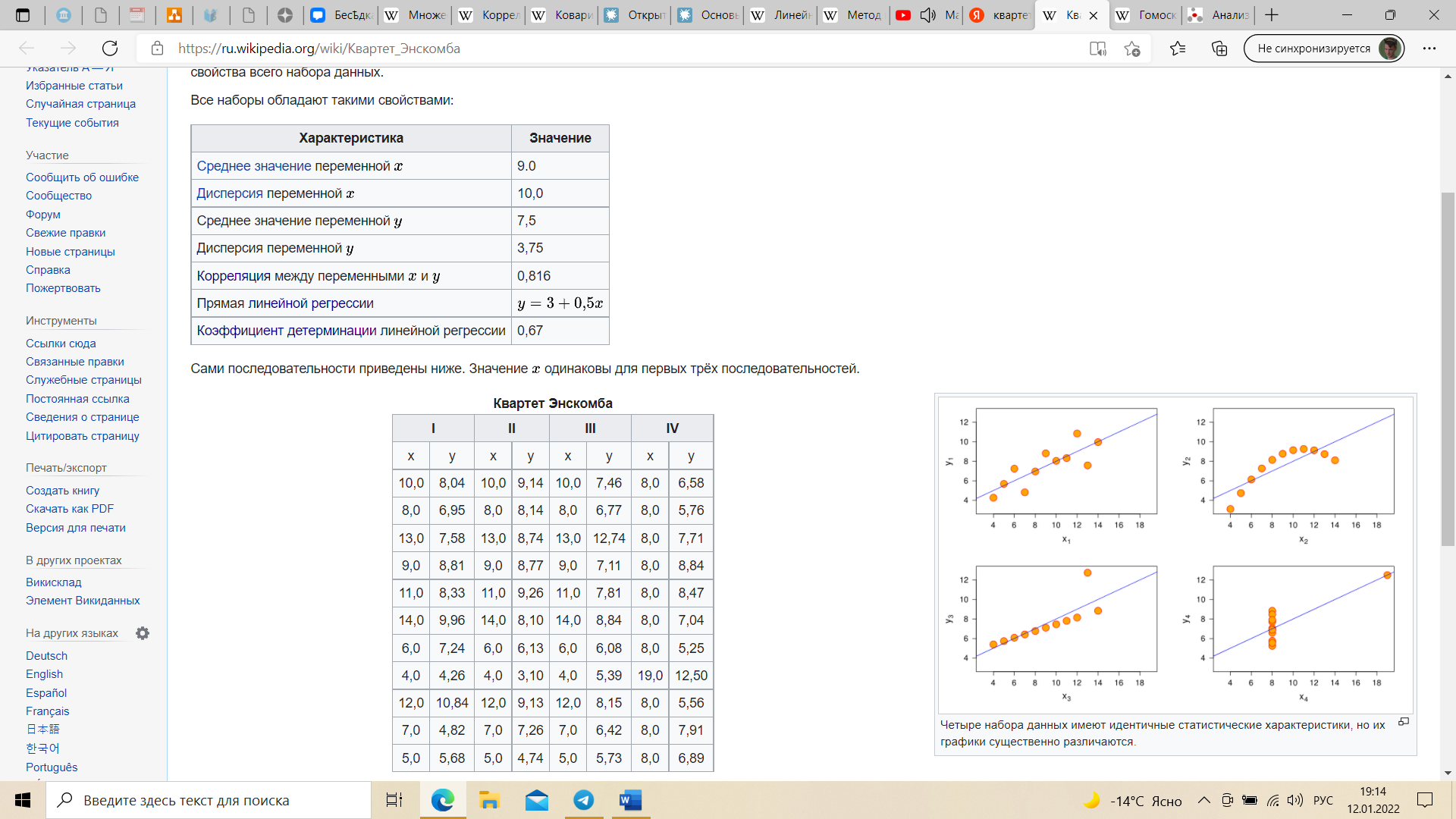
Коэффициент детерминации может принимать значения от 0 до 1. В случае линейной зависимости эта характеристика является квадратом множественного коэффициента корреляции.

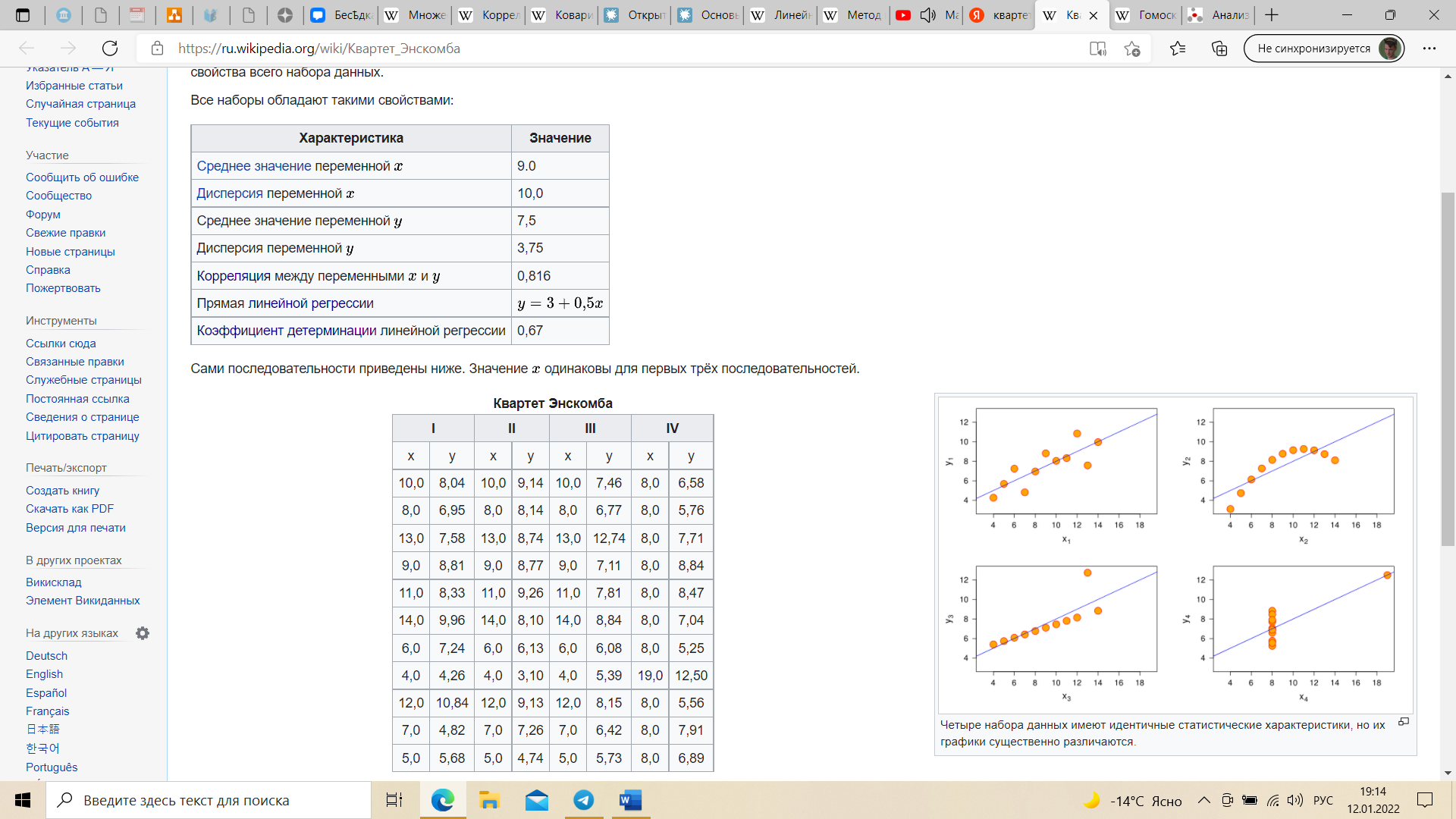
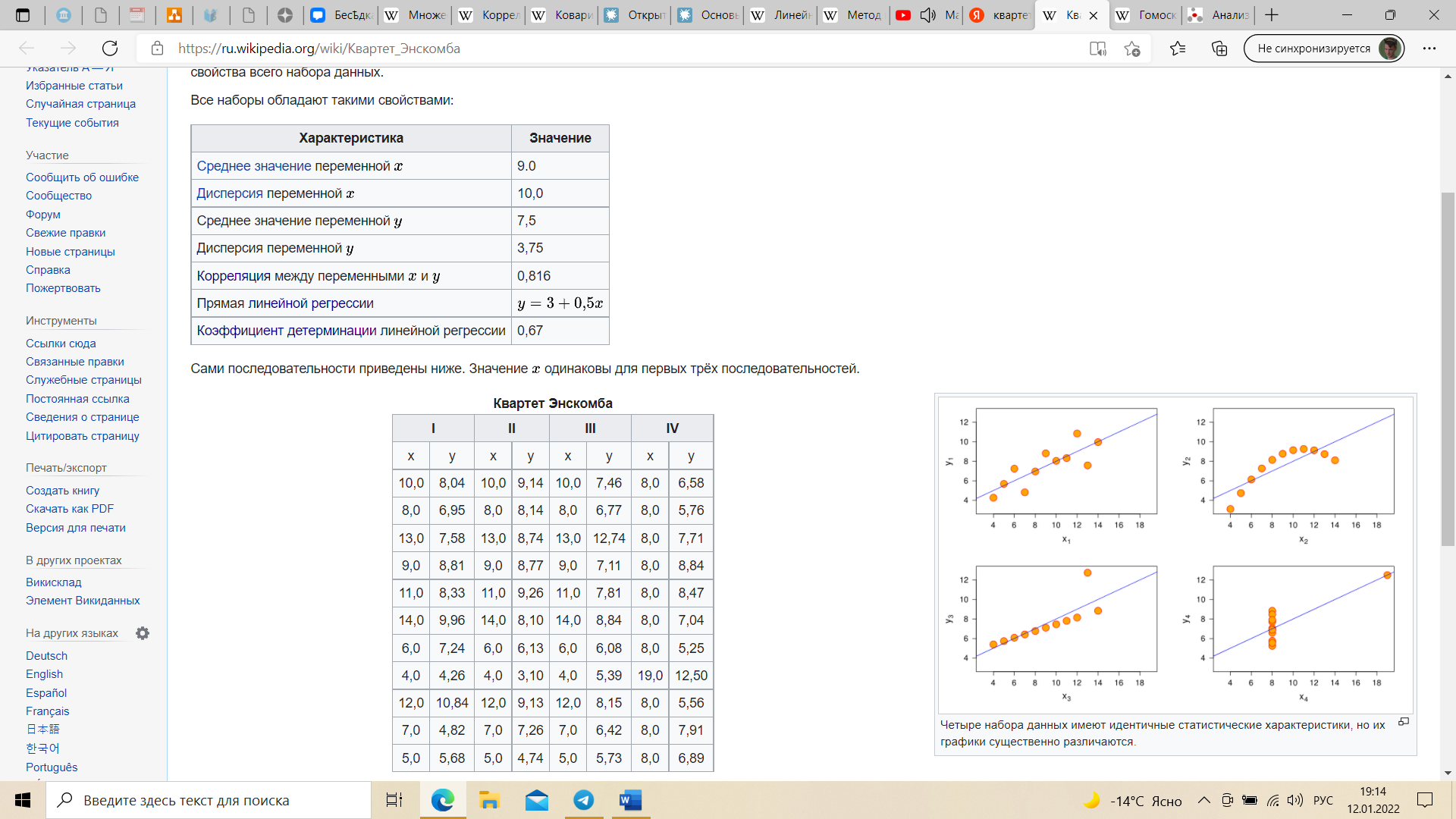
## Анализ остатков. Гомоскедастичность. Квартет Анскомба.

Анализ остатков заключается в проверке следующих гипотез:  
, соответствующее проверке .  
, что является проверкой на гомоскедастичность.  
, соответствующее проверке , где – количество наблюдений в интервале, – ожидаемое число попаданий в интервал.  
Все независимы, что подтверждается коэффициентом корреляции Дарбина-Уотсона  
.

Гомоскедастичность – свойство, означающее постоянство условной дисперсии вектора или последовательности случайных величин. Оно необходимо для эффективности работы метода наименьших квадратов.   
При проверке гомоскедастичности сравниваются дисперсии различных выборок из общего множества величин. Но в частном случае её можно проверить через равенство элементов диагонали ковариационной матрицы.

Квартет Энскомба – четыре набора числовых данных, у которых простые статистические свойства идентичны, но их графики существенно отличаются. Каждый набор состоит из 11 пар чисел. Квартет был составлен в 1973 году английским математиком Ф. Дж. Энскомбом для иллюстрации важности применения графиков для статистического анализа и влияния выбросов значений на свойства всего набора данных.





## Решение для неквадратных и плохо обусловленных матриц.

Число обусловленности – характеристика функции, показывающая, насколько изменится её значение при небольшом изменении аргумента. Оно отражает, насколько функция чувствительна к ошибкам в данных. При линейной регрессии обусловленность используется в качестве диагностики мультиколлинеарности.  
 при .  
Как правило, при можно потерять до знаков точности.

При анализе плохо обусловленных матриц коэффициенты уравнения становятся слишком большими, и становится трудно судить о значимости признака для получения результата. Кроме того, при малейшем изменении значений признаков, например, при шумах или добавлении новых признаков, вектор коэффициентов сильно изменяется.

Для борьбы с плохой обусловленностью используется уменьшение числа признаков через отбор или сведение одних признаков к другим. Кроме того, применяется регуляризация нормы вектора коэффициентов.

## Регуляризация LASSO, Ridge, Elastic.

### LASSO

Метод регрессии лассо (LASSO, Least Absolute Shrinkage and Selection Operator) — это вариация линейной регрессии, специально адаптированная для данных, которые демонстрируют сильную мультиколлинеарность (то есть сильную корреляцию признаков друг с другом). LASSO использует сжатие коэффициентов (shrinkage), то есть процесс, в котором значения данных приближаются к центральной точке (например, среднему значению). Регрессия лассо использует регуляризацию L1, то есть взвешивает ошибки по их абсолютному значению.

Вместо того чтобы корректировать сложность модели, компенсируя сложность данных, подобно методам регрессии с высокой дисперсией нейронных сетей и дерева решений, лассо пытается уменьшить сложность данных так, чтобы их можно было обрабатывать простыми методами регрессии, искривляя пространство, на котором они лежат. В этом процессе лассо автоматически помогает устранить или исказить сильно коррелированные и избыточные функции в методе с низкой дисперсией.

.

### Ridge

Гребневая регрессия или ридж-регрессия очень похожа на регрессию LASSO в том, что она применяет сжатие. Оба алгоритма хорошо подходят для наборов данных с большим количеством признаков, которые не являются независимыми друг от друга (коллинеарность). Однако самое большое различие между ними в том, что гребневая регрессия использует регуляризацию L2, то есть ни один из коэффициентов не *становится* нулевым, как это происходит в регрессии LASSO. Вместо этого коэффициенты всё больше приближаются к нулю, но не имеют большого стимула достичь его из-за природы регуляризации L2.

Гребневая регрессия лучше подходит в ситуации, когда мы хотим сделать приоритетными большое количество переменных, каждая из которых имеет небольшой эффект. Если в модели требуется учитывать несколько переменных, каждая из которых имеет средний или большой эффект, лучшим выбором будет лассо.

.

### Elastic

ElasticNet стремится объединить лучшее из гребневой регрессии и регрессии лассо, комбинируя регуляризацию L1 и L2.

.

# Обобщённые аддитивные модели (generalized additive models).

При линейной регрессии мы ищем . Если взять отдельный член этого уравнения, то он представляет собой линейную функцию . Однако он может быть и более сложной функцией. Тогда .

Если входящие функции имеют сложную структуру, применяются различные методы сглаживания и понижения ранга.

# Partial Least Squares