**Лекция №9 Регрессионный анализ**

**Цель лекции:**

* Изучить линейную регрессию на примере парной линейной регрессии
* Познакомиться с различными методами построения модели
* Построить модель линейной регрессии
* Научиться оценивать полученную математическую модель с помощью различных метрик
* Познакомиться с логистической регрессией.

**Материал прошлого урока:**

Мы уже обсудили с вами, что такие методы как корреляционный анализ и регрессионный анализ помогают проследить нам зависимость одной переменной от другой. Но если корреляционный анализ помогает нам оценить тесноту связи, то регрессионный анализ может ответить на вопрос, как именно изменяется зависимая переменная при изменении независимой переменной

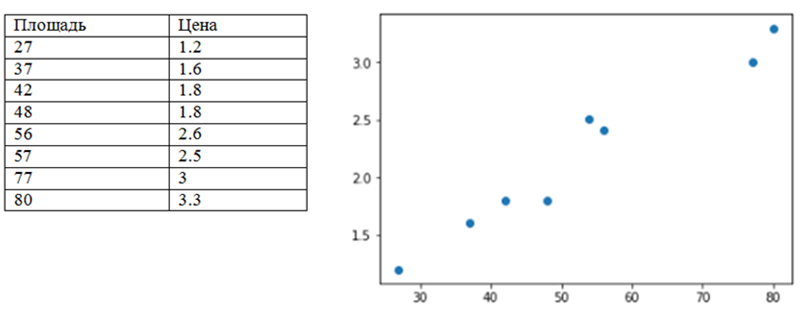
**План урока:**

1. Регрессионный анализ
2. Парная линейная регрессия
3. Метод наименьших квадратов
4. Условия применимости линейной регрессии
5. Трансформация данных
6. Поиск коэффициентов для парной линейной регрессии
7. Функция потерь
8. Матричный метод расчета коэффициентов линейной регрессии
9. Расчет коэффициентов методом градиентного спуска
10. Функция в Python для построения линейной регрессии
11. Коэффициент детерминации
12. Оценка значимости математической модели
13. Оценка значимости отдельных коэффициентов
14. Логистическая регрессия

**Регрессионный анализ**

Мы уже обсудили с вами, что такие методы как корреляционный анализ и регрессионный анализ помогают проследить нам зависимость одной переменной от другой. Но если корреляционный анализ помогает нам оценить тесноту связи, то регрессионный анализ может ответить на вопрос, как именно изменяется зависимая переменная при изменении независимой переменной . Регрессионный анализ помогает нам подобрать математическую модель определенного типа, с помощью которой мы опишем изменение зависимой переменной при изменении независимой переменной .

Давайте вернемся к тем двум переменным, площадь и цена, с которыми мы работали на прошлом уроке. У нас есть 8 парных значений, мы можем нанести их на график синими точками. Каждая точка имеет 2 координаты (площадь, цена). Площадь здесь будет независимой переменной (признак, причина), а цена квартиры – зависимой переменной (результирующая, следствие).



Но мы могли бы иметь несколько признаков, например, удаленность от метро. Тогда бы каждая точка на графике задавалась бы 3 координатами. И такие данные уже будут называться многомерными. Занимается изучением этих данных многомерный статистический анализ.

Т.е. **многомерный статистический анализ – это раздел статистики, который посвящен исследованиям экспериментов с многомерными наблюдениями.** Вернемся к этому понятию чуть позже. Обратите внимание, что регрессии могут быть, как линейные, так и нелинейные. И начнем с самого простого и очень часто используемого метода парной линейной регрессии.

Но давайте еще раз проговорим, для чего нужна математическая модель? С помощью нее мы сможем предсказывать результирующую переменную. Мы подберем модель, а затем, имея только значение признака, сможем предсказать результирующую. В нашем случае с помощью площади сможем предсказать стоимость квартиры.

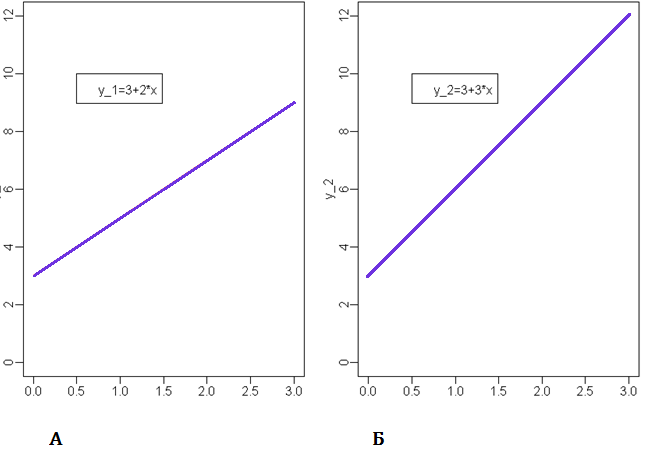
**Парная линейная регрессия**

Если признак один, то такая линейная регрессия называется парной. Она описывает связь признака с результирующей . В основе математической модели, описывающей эту зависимость, лежит предположение, что зависимость выражена следующим образом:

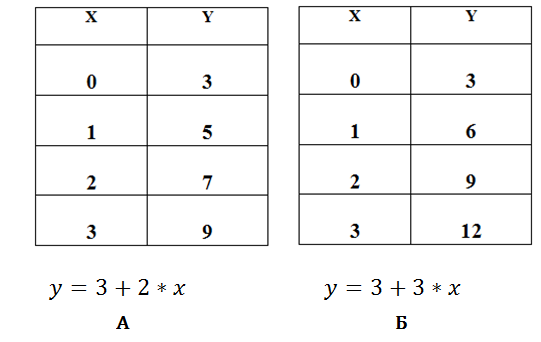
Это уравнение прямой. Чтобы сделать такое предположение, нам нужно для начала взглянуть на график и оценить, прослеживается ли между переменными линейная зависимость. Эти данные знакомы нам уже с прошлого урока, поэтому мы точно знаем, что линейная зависимость есть, к тому же, посмотрев на расположение синих точек на графике, можно убедиться, что она напоминает прямую.

Давайте для начала рассмотрим коэффициенты и в данном уравнении. Коэффициент называется в статистике интерсептом (перехватом).

Посмотрим на два рисунка ниже. На них представлены 2 математические модели в виде уравнения и графика. Если мы мысленно продолжим фиолетовую прямую, то она пересечет ось , когда его значение будет равно 3. Т.е. зависимая переменная равна 3, когда независимая переменная равна 0. Значение 3 и будет являться интерсептом. Для обеих моделей он одинаков.



Также мы видим, что на рисунке А зависимая переменная изменяется не так быстро, как на рисунке Б. Давайте построим простые таблицы, как мы делали в школе, для обоих графиков. Независимую переменную будем изменять на 1.

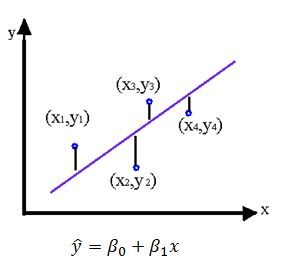


В таблице для математической модели А мы видим, что изменяется на 1, а при этом изменяется на 2 единицы. А в таблице, построенной для математической модели Б, видно, что при изменении на единицу при этом меняется на 3 единицы. Т.е. коэффициент перед признаком показывает, на сколько единиц изменяется зависимая переменная при изменении на 1. Вот мы и видим, что модель Б описывает данные, где изменяется быстрее, чем в данных, описанных с помощью модели А.

**Метод наименьших квадратов**

Чтобы построить линейную регрессию для начала нам надо накопить данные. Эти данные мы будем использовать, чтобы подобрать коэффициенты. У нас есть все те же данные, площадь и цена, мы убедились на прошлом уроке, что они имеют линейную зависимость, теперь запишем модель, которая будет описывать эту зависимость.

В статистике вместо коэффициентов и принято использовать коэффициенты . Имея данные, мы подбираем коэффициенты таким образом, что прямая (графическая математическая модель) проходит максимально близко ко всем точкам. Т.е. коэффициенты подбираются так, что сумма квадратов расстояний от синих точек до прямой минимальна. Иными словами, в основе линейной регрессии лежит метода наименьших квадратов. После того, как мы подберем коэффициенты, мы можем подставить признак () в это уравнение и посчитать оценочные значения ( с крышкой).



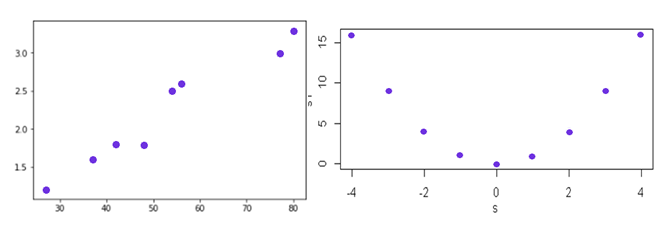
**Условия применимости линейной регрессии**

Помимо предварительной оценки на наличие линейной зависимости, есть еще три условия применимости. Перечислим все условия применимости ниже:

1. Наличие линейной зависимости между зависимой и независимой переменной.
2. Независимость остатков
3. Для любого значения значение зависимой переменной распределено нормально.
4. Гомоскедастичность.

Теперь каждое условие рассмотрим более подробно.

1. Наличие линейной зависимости можно увидеть с помощью графика. На первом графике условие соблюдается, на втором, где видим параболу, нет.

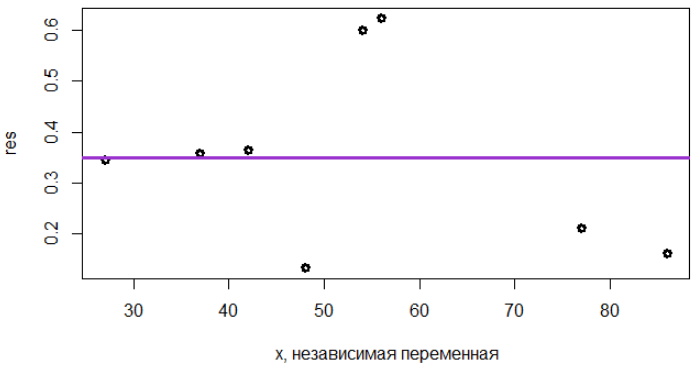


Решить эту проблему можем с помощью нелинейной трансформации и/или . Здесь мы можем применить

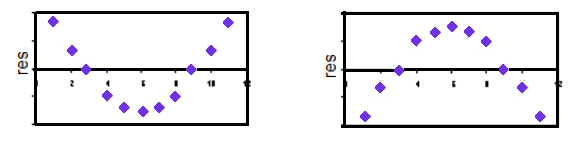
* логарифм,
* квадратный корень
* умножение на обратное число

1. Независимость остатков

Проверить можем с помощью графика, на котором с течением времени мы не должны видеть каких-то паттернов в поведении остатков. Например, с течением времени остатки не должны расти. Они должны быть разбросаны случайным образом вокруг горизонтальной линии, как, например, на рисунке ниже.



А если мы видим расположение остатков, например, напоминающих параболу, то мы приходим к выводу, что для таких данных лучше подойдет нелинейная модель.

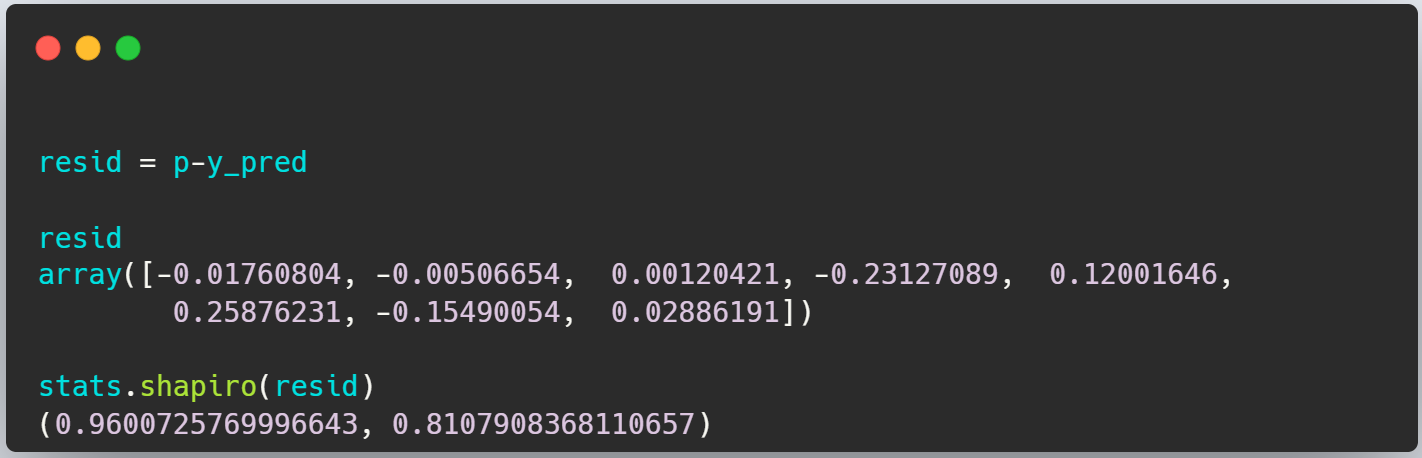


Эта проблема тоже решаема с помощью трансформации данных.

1. Нормальное распределение остатков. Проверить данное утверждение можно после построение математической модели линейной регрессии.

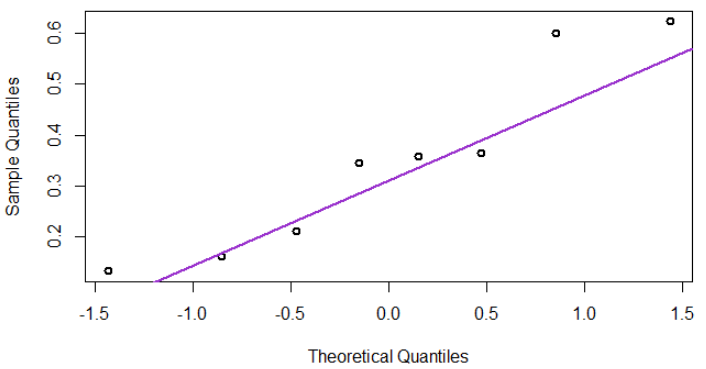
Методы проверки:

* Тест Шапиро-Уилка
* QQ –график



Мы видим, что статистика 0.96, а p-value = 0.81. Если сравнить с заранее установленным уровнем статистической значимости α, возьмем, как обычно, 0.05, то видим что p-value > α. Если p-value > α, то делаем вывод всегда в пользу нулевой гипотезы. А нулевая гипотеза гласит о том, что различий нет. Т.е. мы не нашли отличия от нормального распределения.

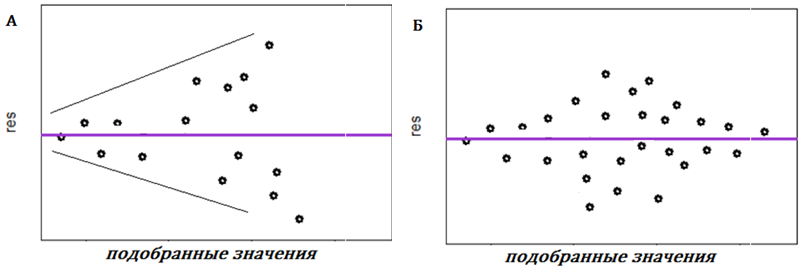
А теперь убедимся в правильности вывода с помощью графика QQ-plot.



И здесь практически все точки лежат на прямой.

1. **Гомоскедастичность. Под этим свойством понимается постоянство дисперсии при всех значениях .**

Например, если расположение остатков напоминает конус, мы говорим, о гетороскедастичности. Ниже на рисунке изображено два примера гетероскедастичности. На рисунке А с увеличением растет дисперсия. На рисунке Б дисперсия максимальна при средних значениях .



**Трансформация данных**

Чтобы обеспечить условия применимости линейной регрессии иногда нам необходимо сделать трансформацию данных. Но нельзя забывать, что получив модель, нам нужно сделать обратную трансформацию. Основные методы приведены в таблице ниже.

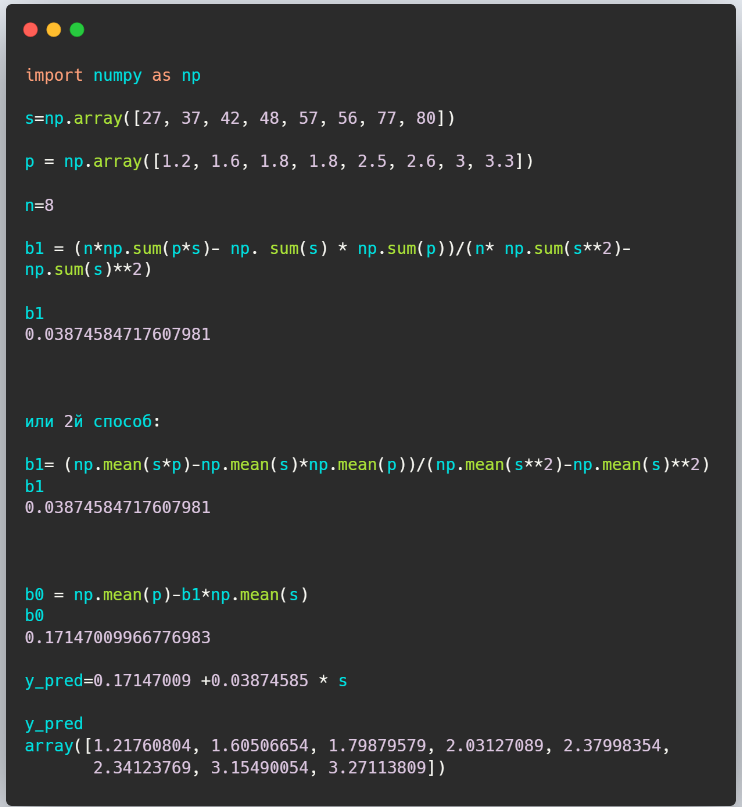
| **Трансформация** | **Уравнение регрессии** | **Оценочное значение** |
| --- | --- | --- |
|  | = |  |
|  | = |  |
| s | s= |  |
|  | = ) |  |
|  | ) |  |

**Поиск коэффициентов для парной линейной регрессии (формулы)**

Если у нас есть модель вида , то расчет коэффициентов производится по следующим формулам:

Сделаем эти вычисление ниже в Python. Один из расчетов ниже для коэффициента будет сделан по формуле, указанной выше. Другая формула (2 способ) также встречается в литературе и приводит к точно такому же результату. После того, как мы рассчитаем коэффициенты, получим модель вида:

Если мы обратно подставим в эту модель площадь и посчитаем соответствующую цену за квартиру, то это уже будут оценочные значения *y\_pred.*

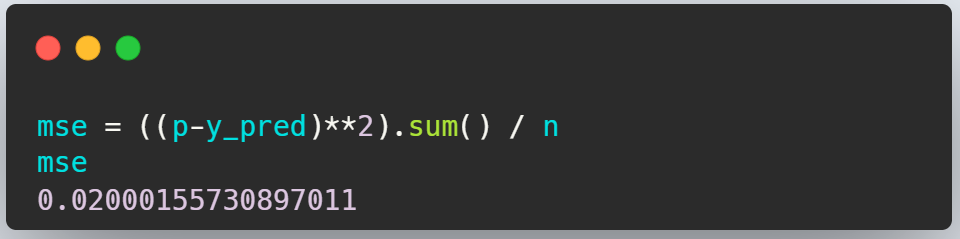
**

**Функция потерь**

Как было сказано ранее, в основе метода линейной регрессии лежит метод наименьших квадратов. Т.е. коэффициенты подбираются таким образом, чтобы минимизировать среднюю квадратичную ошибку, которая находится по формуле:

– число измерений.

**Функция потерь mse - мера измерения ошибок, которые функция делает на нашем наборе данных**.

**

**Матричный метод расчета коэффициентов линейной регрессии**

Чтобы решить матричным методом уравнение вида , нам нужно представить это уравнение в матричном виде*.* Матрица равна произведению матрицы Y и матрицы B.

*Y = X \* B*

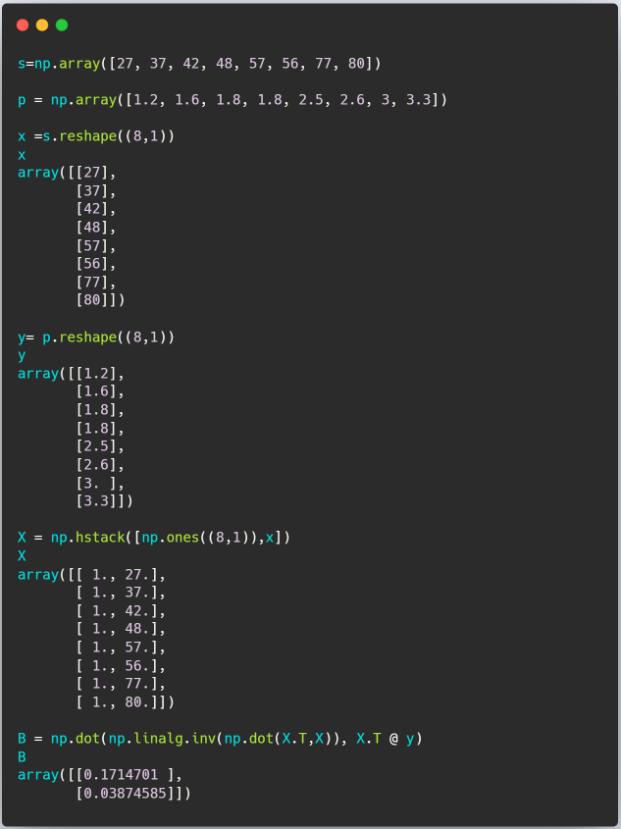
Чтобы создать матрицы, необходимо массив площади и цены превратить в вектор-столбец с помощью атрибута reshape (смотрите ниже). Когда у нас модель с интерсептом, то необходимо добавить в матрицу X столбец единиц. В Python создаем столбец единиц через функцию np.ones() и добавляем к вектору-столбцу X с помощью функции np.hstack().

После этого останется решить матричное уравнение. Не будем на этом курсе заниматься выводом формулы для расчета матрицы B. Вы можете воспользоваться уже выведенной формулой для дальнейшего применения в своих задачах.

Перемножение матриц в Python делается через функцию dot или @. Решив матричное уравнение, получаем матрицу B. Матричный метод дал нам те же значения коэффициентов, что и метод с использованием формул.

Матричный метод мы также можем использовать и при многомерном анализе. Тогда у нас будет больше столбцов (признаков) в матрице X и, соответственно, матрица B даст больше коэффициентов.

Если по какой-то причине мы хотим построить модель без интерсепта, тогда просто не добавляем столбец единиц.

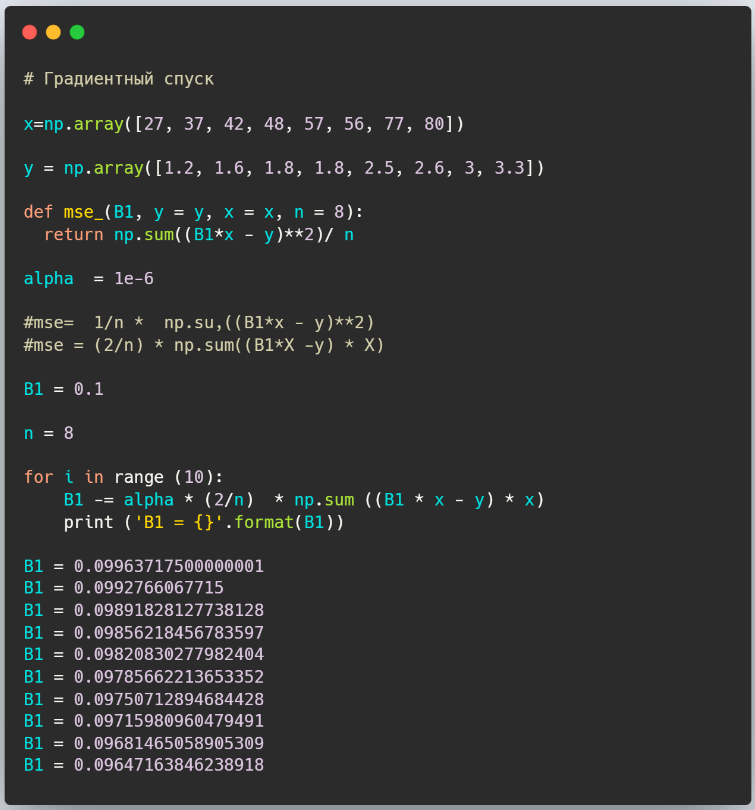
**

**Расчет коэффициентов методом градиентного спуска**

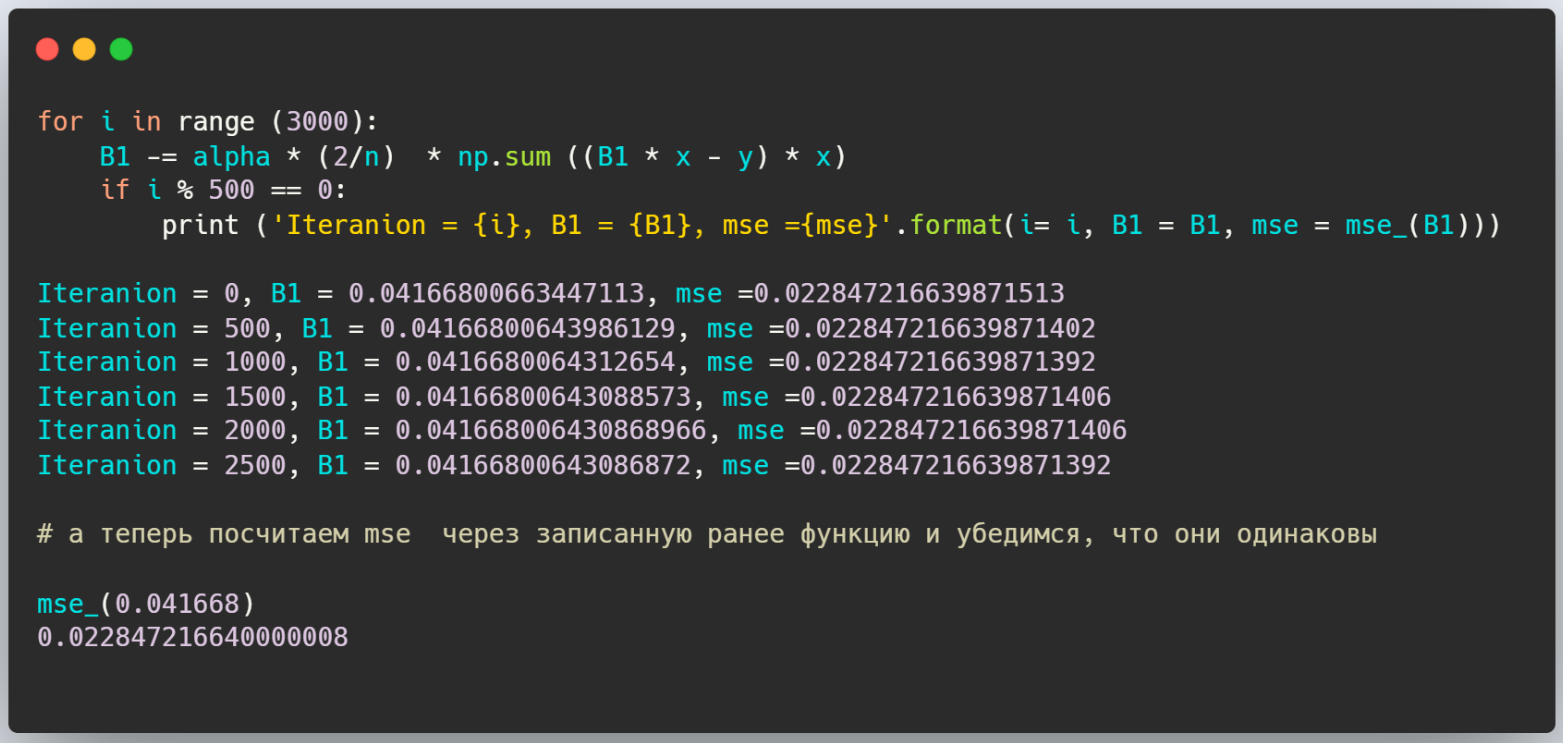
Градиентный спуск – итеративный метод машинного обучения. Мы рассмотрим этот алгоритм поиска коэффициентов для модели вида без интерсепта.

Для начала напишем в Python функцию mse , чтобы потом сравнить расчеты mse через функцию с вычислениями mse с помощью метода градиентного спуска. Коэффициенты подбираются так, чтобы минимизировать эту ошибку. В этом методе мы работаем с производной mse, причем производную берем по коэффициенту, который ищем, а все остальные переменные рассматриваем как константы.

Нам понадобится установить скорость обучения alpha. Нет четких правил по выбору этого параметра. Если мы возьмем слишком большой, то не сможем посчитать коэффициент, проскакивая через минимум, а если слишком маленький, то вычисления могут занять очень много времени. Также надо определиться со стартовыми значениями коэффициентов. У нас модель без интерсепта, поэтому только определяем стартовое значение для = 0.1. (Обычно значение берут из нормального стандартного распределения). В примере ниже взяли 10 итераций и скорость обучения alpha = 0.000001. Коэффициент на 10-й итерации продолжает изменяться.



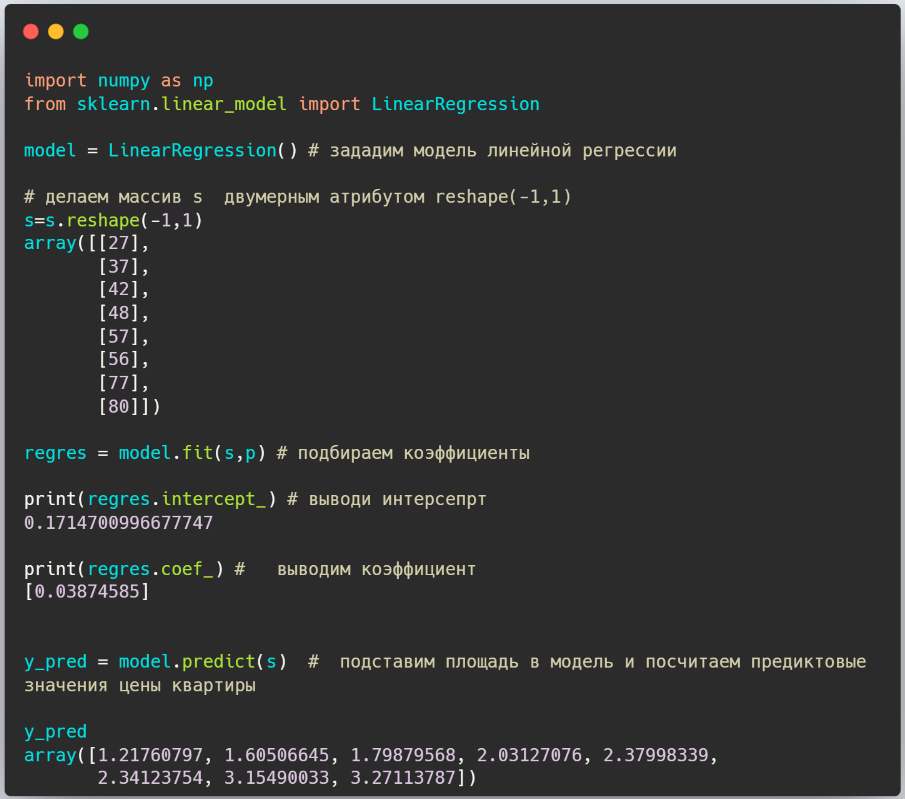
Теперь увеличим число итераций до 3000 и будем выводить каждую 500 итерацию и помимо коэффициента будем выводить и mse.



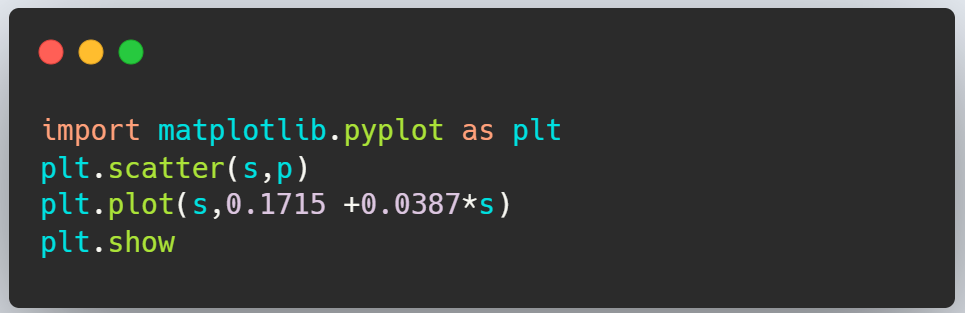
Мы видим, что в последних 2-х выведенных итерациях mse уже не меняется до 14-го знака после запятой и изменение коэффициента также остановилось. Мы получили модель вида:\*

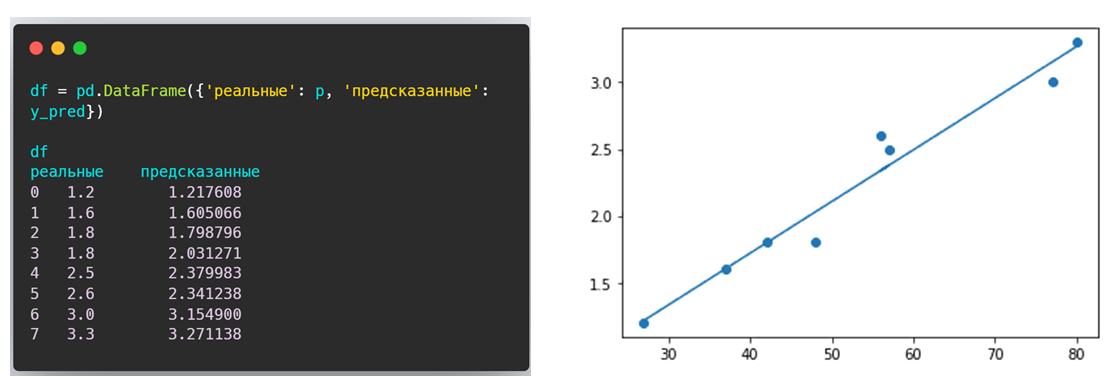
**Функция в Python для построения линейной регрессии**

В Python также есть метод для построения линейной регрессии. Сначала мы задаем математическую модель. Затем превращаем независимую переменную s в двумерный массив, и после этого подбираем коэффициенты. Комментарии также оставлены в коде.



Построим график, где линия – модель линейной регрессии, а точки – истинные значения.

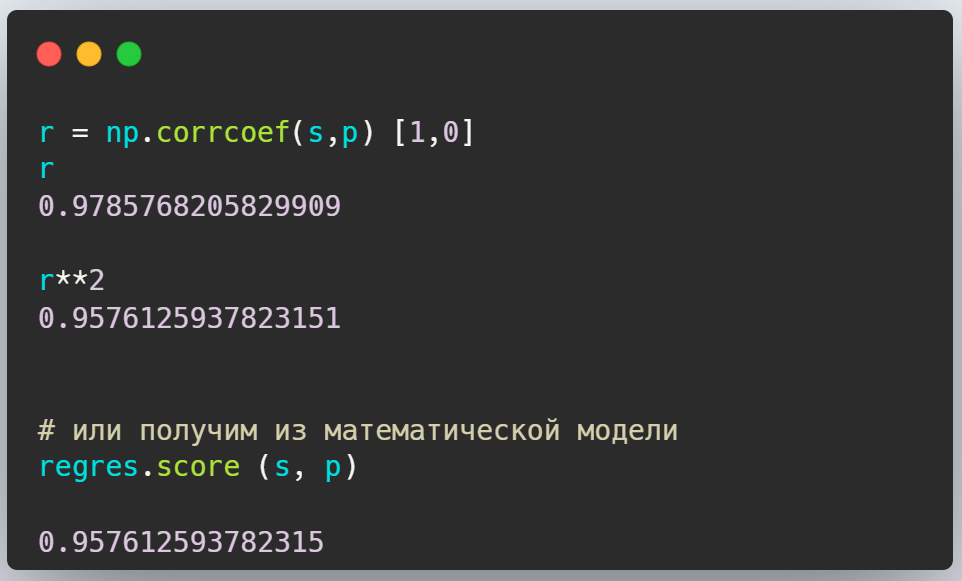




После того как построили модель линейной регрессии мы должны оценить полученную модель. Нам понадобятся такие параметры как коэффициент детерминации, критерий Фишера и критерий Стьюдента.

**Коэффициент детерминации**

Коэффициент детерминации показывает, какую долю изменчивости описала подобранная математическая модель. Коэффициент детерминации равен квадрату коэффициента корреляции и обозначается .



Это означает, что приблизительно 95.8% изменчивости описано с помощью подобранной математической модели, что является очень хорошим показателем.

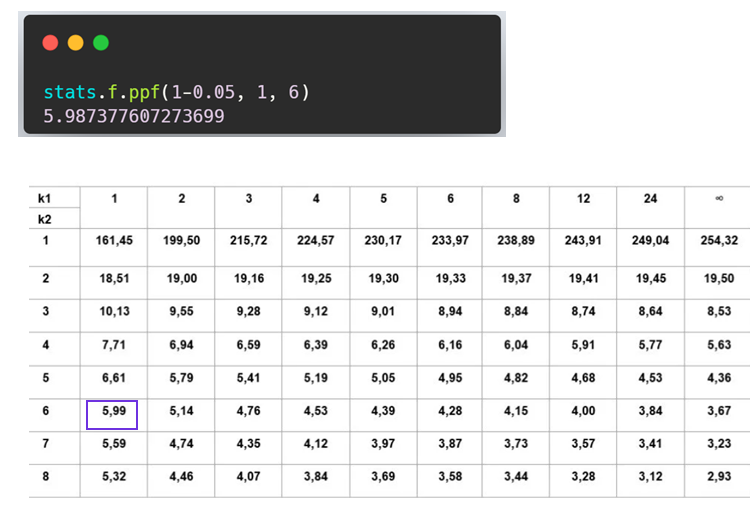
**Оценка значимости математической модели**

Для оценки значимости математической модели используется критерий Фишера. Критерий Фишера имеет ассиметричное распределение с правым длинным хвостом. Как и в любом тестировании гипотезы, нам надо сравнить табличное значение коэффициента с расчетным. Табличное значение выбираем по таблице значений F - критерия (таблица Фишера- Снедекора) для заданного уровня статистической значимости. Установим заранее уровень статистической значимости α, как обычно, 0.05.

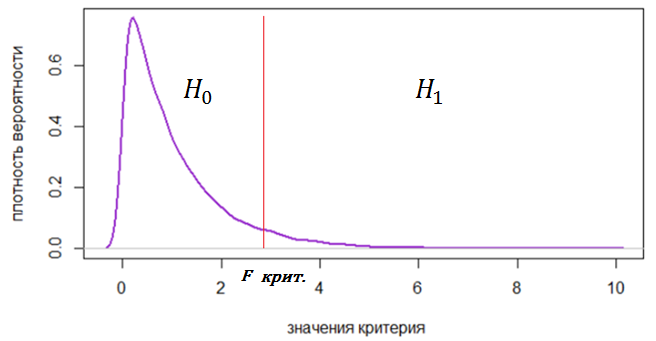
Рассчитаем критерий Фишера F в Python



Сравним расчетное значение с табличным значением, которое можем получить либо с помощью функции f.ppf, либо с помощью таблицы значений критерия Фишера.

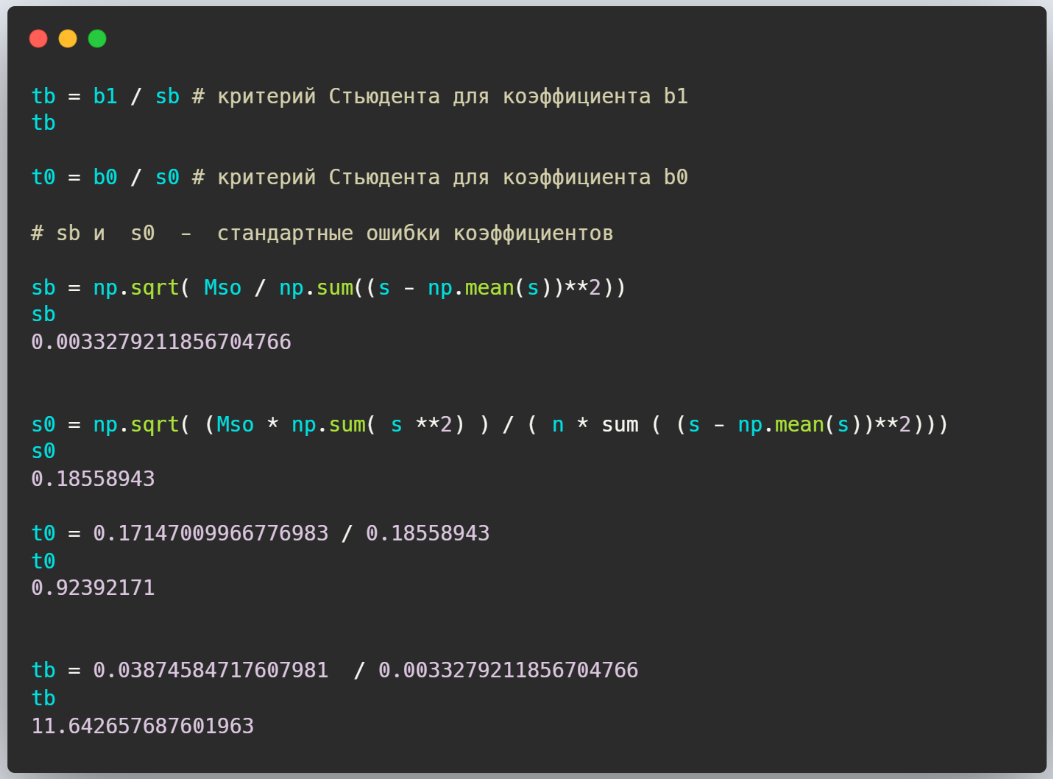


Осталось сделать вывод. Табличное значение разбивает график на область принятия нулевой гипотезы и альтернативной. Критическое значение 5.99, а расчетное 135.55, т.е. попадает в область принятия гипотезы . Т.е. наша модель признана статистически значимой на уровне α= 0.05

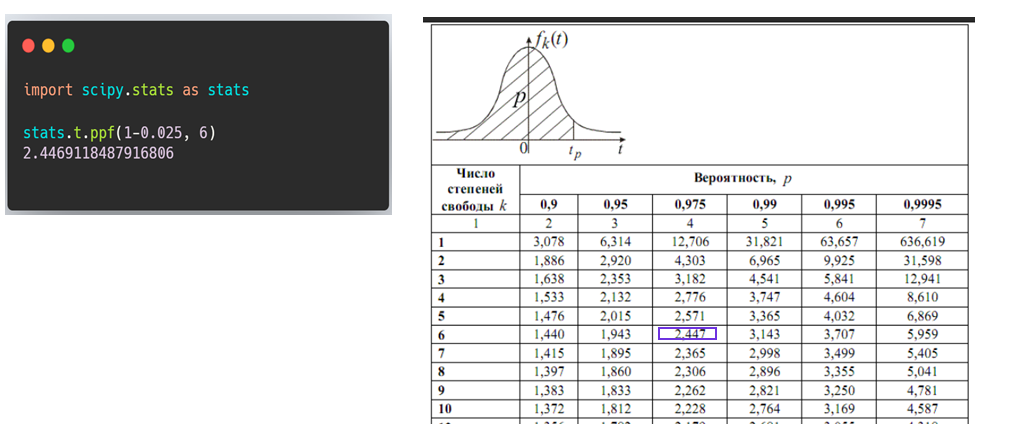
****

**Оценка значимости отдельных коэффициентов**

Помимо оценки статистической значимости модели в целом, также используют оценку значимости отдельных коэффициентов. Для этого воспользуемся критерием Стьюдента. Критерий Стьюдента рассчитываем: значение коэффициента разделить на его стандартную ошибку. Вычисления приведены ниже. Mso находили выше при расчете критерия Фишера.

****

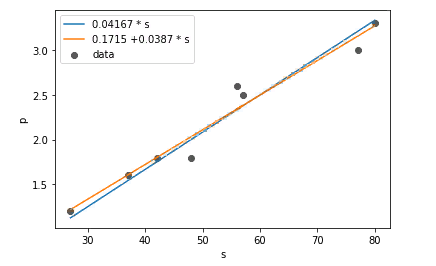
Найдем табличный критерий Стьюдента для α = 0.05 и степеней свободы n - 2 = 6. Помним, что здесь тест двусторонний, поэтому в функции указываем значение . Табличный критерий 2.447.

****

Критерий Стьюдента для коэффициента попадает в область принятия нулевой гипотезы, т.е. коэффициент не является статистически значимым, а вот tb = 11.64 для коэффициента является статистически значимым.

Давайте сравним две графические модели с интерсептом и без. Несмотря на то, что интерсепт в данной модели не является статистически значимым, мы его оставляем. И вообще на практике, всегда рекомендовано строить модели с интерсептом.

Если мы взглянем на mse для модели без интерсепта (0.02285) и на mse для модели с интерсептом (0.020), то видим, что все-таки для модели с интерсептом mse меньше.



Итак, окончательная модель имеет вид

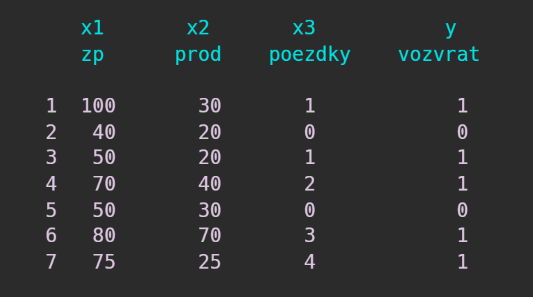
Модель является статистически значимой на уровне

Помимо линейной регрессии, встречаются и нелинейные регрессии. Одним из примеров таких моделей является логистическая регрессия.

**Логистическая регрессия**

Логистическая регрессия применяется, когда является бинарной переменной (0 или 1). Т.е. с помощью этого метода мы можем решить задачу бинарной классификации.

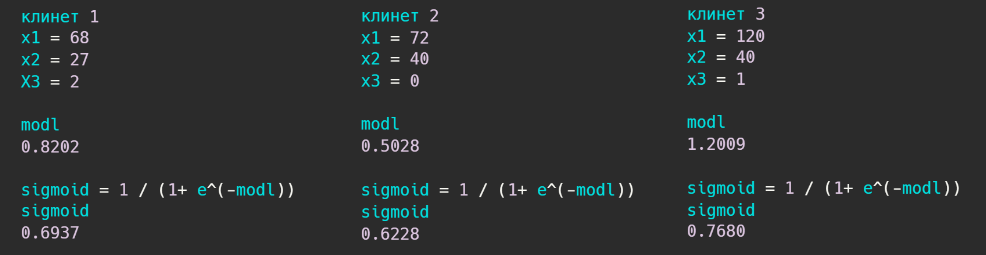
Рассмотрим применения логистической регрессии в банковском деле, когда принимается решение о выдаче кредита. У нас есть уже некоторые данные, где мы знали о клиентах их заработную плату zp, среднюю сумму, которую они оставляют на продукты, prod, количество поездок за рубеж poezdky и результирующую (вернул - 1 или не вернул кредит- 0).



Построим предварительную линейную модель modl.



Если мы будем подставлять данные по новым клиентам, то эта модель не обеспечит нас значениями в пределах от 0 до 1, поэтому мы должны применить трансформацию данных для , полученного с помощью model. Применим сигмоидную функцию.

**

Теперь – это вероятность, что клиент вернет кредит. Банк выбирает отсечку, выше которой будет превращаться в 1, т.е. такому клиенту кредит дадут.

Сегодня мы познакомились с регрессионным анализом. На следующем уроке мы продолжим знакомство с критерием Фишера. Научимся его применять в регрессионном анализе, который позволяет делать множественные сравнения.