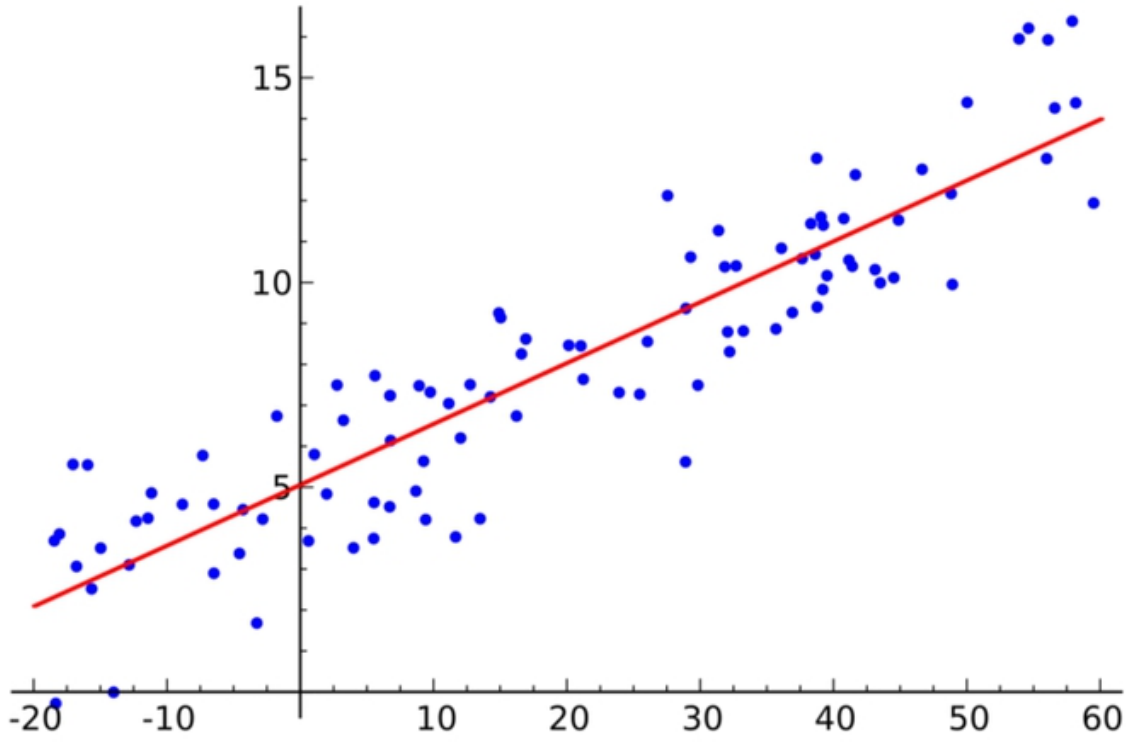
Что такое регрессия?

Регрессионный анализ — это форма метода прогнозного моделирования, которая исследует взаимосвязь между зависимой и независимой переменной.

Приведенное выше определение является книжным определением, простыми словами регрессия может быть определена как «Использование взаимосвязи между переменными для поиска наилучшей линии соответствия или уравнения регрессии, которое можно использовать для прогнозирования».



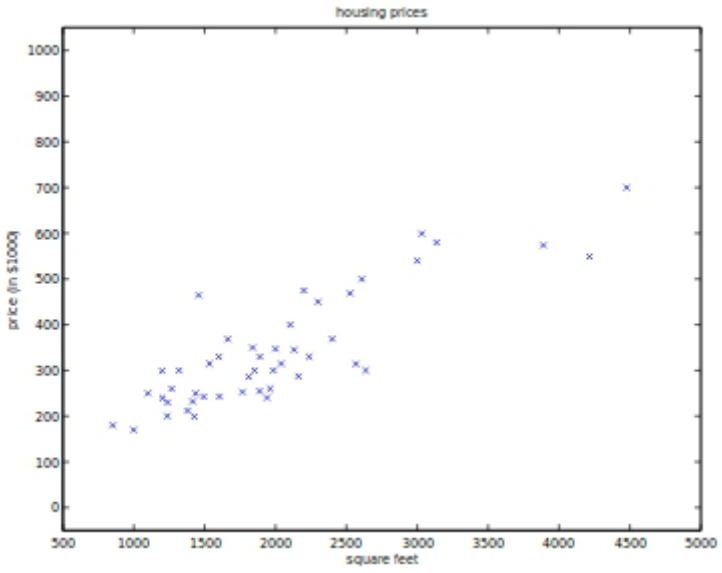
Линейная регрессия

Линейная регрессия — это базовый и широко используемый тип прогнозного анализа, который обычно работает с непрерывными данными. Мы попытаемся понять линейную регрессию на примере:

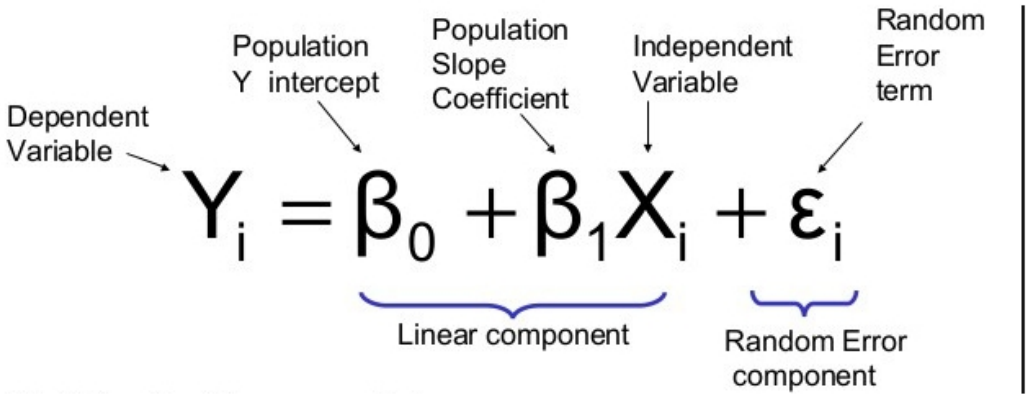
Аарав пытается купить дом и собирает данные о жилье, чтобы оценить «стоимость» дома в соответствии с «жилой площадью» дома в футах.

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

Он наблюдает за данными и приходит к выводу, что данные являются линейными после того, как он строит график рассеяния. Для своего первого точечного графика Аарав использует две переменные: «жилая площадь» и «цена».



Как только он увидел закономерность в данных, он планировал провести линию регрессии на графике, чтобы использовать эту линию для прогнозирования «цены дома».
Используя обучающие данные, т. е. «Цена» и «Жилая площадь», получается линия регрессии, которая даст минимальную ошибку. Для этого ему нужно построить линию, ближайшую к как можно большему количеству точек. Затем это «линейное уравнение» используется для любых новых данных, чтобы он мог предсказать требуемый результат.



Здесь β_1 — это параметры (также называемые весами), β_0 — точка пересечения с осью y, а $Є_i$ — член случайной ошибки, роль которого заключается в добавлении смещения. Приведенное выше уравнение является линейным уравнением, которое необходимо получить с минимальной ошибкой.

Приведенное выше уравнение представляет собой простое «уравнение прямой», то есть

$$Y(predicted) = (\beta 1 * x + \beta o) + Error\ value$$

Где «β1» — наклон, а «βo» — точка пересечения с осью y, аналогичная уравнению прямой. Значения «β1» и «βo» должны быть выбраны так, чтобы минимизировать ошибку. Чтобы проверить ошибку, мы должны вычислить сумму квадратов ошибок и настроить параметры, чтобы попытаться уменьшить ошибку.

Error = Σ (actual output — predicted output)²

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2.$$

Cost function

Ключ:

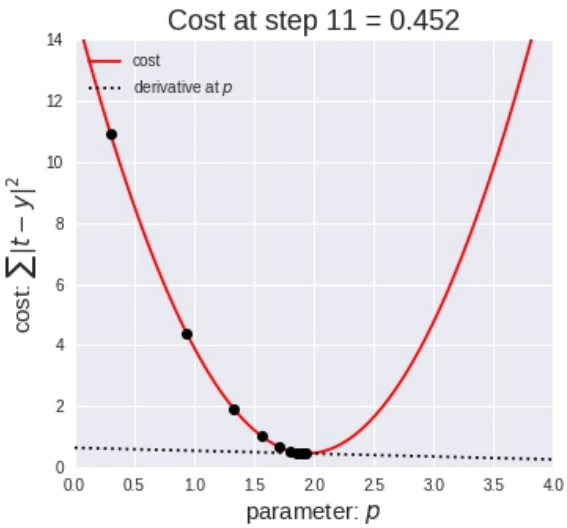
1. **Y(предсказанный)** также называется функцией гипотезы.

2. **J(θ)** — функция стоимости, которую также можно назвать функцией ошибок. Наша главная цель – минимизировать стоимость.

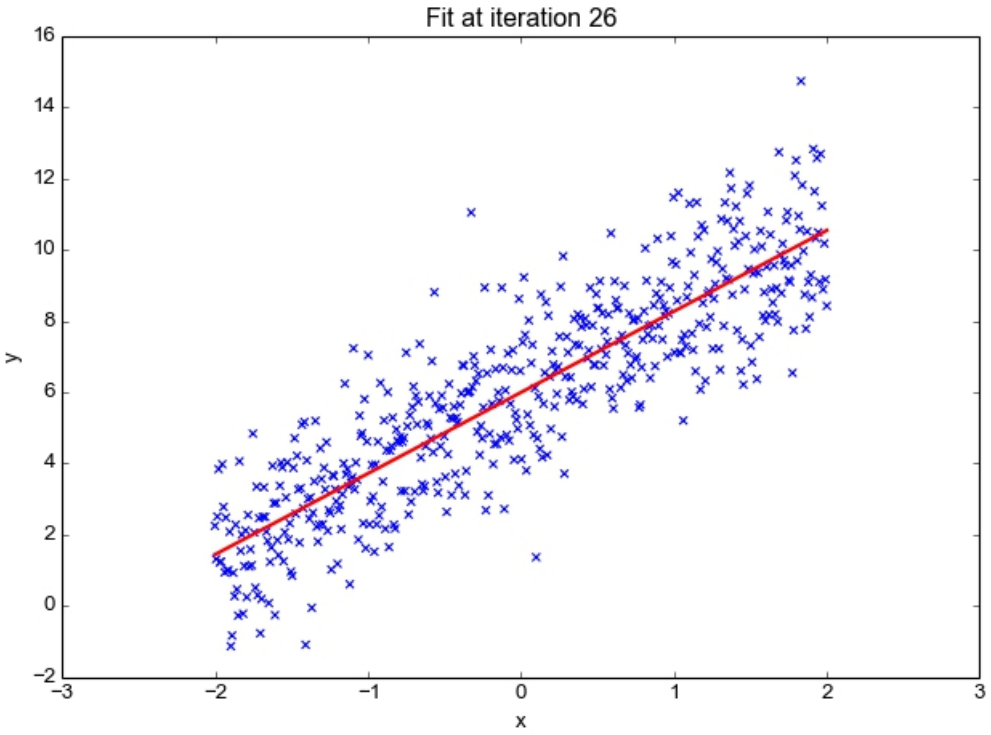
3. **y(i)** — прогнозируемый результат.

4. **hθ(x(i))** называется функцией гипотезы, которая в основном представляет собой значение Y (прогнозируемое).

Теперь возникает вопрос, как нам уменьшить значение ошибки. Что ж, это можно сделать с помощью градиентного спуска. Основная цель градиентного спуска — минимизировать стоимость. т.е. мин J(θo, θ1)

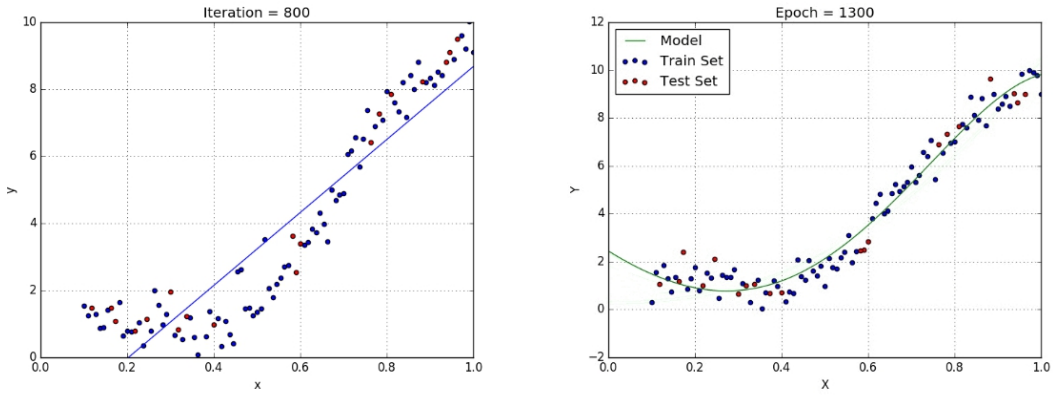


Выбор идеальной скорости обучения — очень важная задача, поскольку она зависит от того, насколько большой шаг мы делаем вниз во время каждой итерации. Если мы сделаем слишком большой шаг, мы можем перешагнуть через минимум. Однако, если мы будем делать небольшие шаги, потребуется много итераций, чтобы достичь минимума.



Полиномиальная линейная регрессия

В предыдущем разделе мы видели, что две переменные в вашем наборе данных коррелированы, но что произойдет, если мы знаем, что наши данные коррелированы, но связь не выглядит линейной? Следовательно, в зависимости от того, как выглядят данные, мы можем выполнить полиномиальную регрессию для данных, чтобы подобрать к ним полиномиальное уравнение.



Left: Linear Regression, Right: Polynomial regression | GIF: Towards Data Science

Следовательно, если мы попытаемся использовать простую линейную регрессию на приведенном выше графике, то линия линейной регрессии не очень хорошо подойдет. Очень сложно уместить линию линейной регрессии на приведенном выше графике с низким значением ошибки. Следовательно, мы можем попытаться использовать полиномиальную регрессию, чтобы подобрать полиномиальную линию, чтобы мы могли достичь минимальной ошибки или функции минимальной стоимости. Уравнение полиномиальной регрессии для приведенных выше графических данных будет выглядеть так:

$$y = \theta o + \theta_1 x_1 + \theta_2 x_1^2$$

Это общее уравнение полиномиальной регрессии:

$$Y = \theta o + \theta_1 X + \theta_2 X^2 + \ldots + \theta_m X^m + \text{residual error}$$

Преимущества использования полиномиальной регрессии:

Полином обеспечивает наилучшее приближение отношения между зависимой и независимой переменной.

Под него можно поместить широкий спектр функций.

Многочлен в основном соответствует широкому диапазону кривизны.

Недостатки использования полиномиальной регрессии

Наличие одного или двух выбросов в данных может серьезно повлиять на результаты нелинейного анализа.

Они слишком чувствительны к выбросам.

Кроме того, к сожалению, существует меньше инструментов проверки модели для обнаружения выбросов в нелинейной регрессии, чем для линейной регрессии.