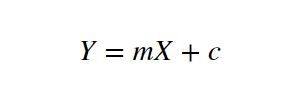
линейная регрессия

В статистике линейная регрессия - это линейный подход к моделированию взаимосвязи между зависимой переменной и одной или несколькими независимыми переменными. Пусть **X**-независимая переменная, а **Y**-зависимая переменная. Мы определим линейную зависимость между этими двумя переменными следующим образом:



чтобы найти регрессию с минимальными ошибками - ГРАДИЕНТ

Рост=a + возраст \* b

В этом случае “А” и " Б " называются соответственно перехватом и наклоном. В том же примере “А " или перехват-это значение, с которого вы начинаете измерение. Новорожденные дети с нулевым месяцем не обязательно имеют нулевые сантиметры; это функция перехвата. Наклон измеряет изменение высоты по отношению к возрасту в месяцах. В общем, с каждым месяцем взросления ребенка его рост будет увеличиваться на “б".

Нормализация – метод масштабирования, при котором значения сдвигаются таким образом, что в конечном итоге они попадут в диапазон от 0 до 3

## **Значение p: проверка статистической значимости**

 Достаточно ли этого, чтобы на самом деле использовать эту модель? Нет! Прежде чем использовать регрессионную модель, вы должны убедиться, что она статистически значима. Мы можем считать линейную модель статистически значимой только тогда, когда оба этих p-значения меньше заранее определенного уровня статистической значимости, который в идеале равен 0,05. Это визуально интерпретируется звездами значения в конце ряда. Чем больше звезд рядом с P-значением переменной, тем значительнее переменная.

**Когда коэффициенты равны нулю – плохо. Это означает что предсказывать даже не надо, так как нет связи между результатом и переменными.**

#### **Нулевая и альтернативная гипотеза**

Когда есть p-значение, есть нулевая и альтернативная гипотеза, связанная с ним. В линейной регрессии нулевая гипотеза состоит в том, что коэффициенты, связанные с переменными, равны нулю. Альтернативная гипотеза состоит в том, что коэффициенты не равны нулю (то есть существует связь между рассматриваемой независимой переменной и зависимой переменной).

**R-SQUARE**

Другой способ тестирования модели это коэф детерминации R квадрат (Это отношение общей изменчивости к предполагаемой, если хорошо соответствует данными, то R sq = 1 плохо)

F-statistika принять или отклонить нулевую гипетезу

#### **t-значение**

Мы можем интерпретировать значение t примерно так. Чем больше t-значение, тем меньше вероятность того, что коэффициент не равен нулю чисто случайно. Таким образом, чем выше значение t, тем лучше

Pr (> | t |) или p-значение - это вероятность того, что вы получите значение t, такое же высокое или большее, чем наблюдаемое значение, когда нулевая гипотеза (коэффициент β равен нулю или отсутствует связь) верна. Таким образом, если Pr (> | t |) низкое, коэффициенты значимы (значительно отличаются от нуля). Если Pr (> | t |) высокий, коэффициенты не имеют значения.

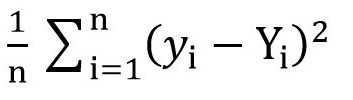
Когда значение p меньше уровня значимости (<0,05), мы можем смело отвергнуть нулевую гипотезу о том, что коэффициент β предиктора равен нулю

Абсолютно важно, чтобы модель была статистически значимой, прежде чем мы сможем использовать ее для прогнозирования (или оценки) зависимой переменной, в противном случае доверие к прогнозируемым значениям из этой модели уменьшается и может быть истолковано как случайное событие.

**Остатки**

Хороший способ проверить качество соответствия модели - посмотреть на остатки или различия между реальными и прогнозируемыми значениями. Прямая линия на изображении выше представляет предсказанные значения. Красная вертикальная линия от прямой к наблюдаемому значению данных - это остаток.

Представим, что у нас есть данные о некоторых n магазинах (торговая площадь и годовой товарооборот) — наш датасет и k торговых площадей(X), для которых мы хотим предсказать годовой товарооборот(Y) — наша задача.  
Выдвинем гипотезу, что наше значение Y зависит от X в виде: Y = a + b \* X.  
Чтобы решить нашу задачу, мы должны подобрать коэффициенты a и b.  
Для начала зададим a и b случайные значения. После этого нам нужно определиться с функцией потерь и оптимизационным алгоритмом.  
Для этого мы можем воспользоваться среднеквадратичной функцией потерь([MSELoss](https://pytorch.org/docs/stable/nn.html" \l "mseloss)). Вычисляется по формуле:



Где y[i] = a + b \* x[i] после a = rand() и b = rand(), а Y[i] – правильное значение для x[i].  
На данном этапе мы имеем среднеквадратичное отклонение(некую функцию от a и b). И очевидно, что, чем меньше значение этой функции, тем более точно подобраны параметры a и b относительно тех параметров, которые описывают точную зависимость между площадью торгового помещения и товарооборотом в этом помещении.  
  
Теперь мы можем приступить к использованию градиентного спуска (как раз для минимизации функции потерь).

## **Стандартная ошибка и F-статистика**

Как стандартные ошибки, так и F-статистика являются показателями хорошей подгонки.



Минимизация любой функции означает поиск самой глубокой впадины в этой функции. Имейте в виду, что функция используется, чтобы контролировать ошибку в прогнозах модели машинного обучения. Поиск минимума означает получение наименьшей возможной ошибки или повышение точности модели. Мы увеличиваем точность, перебирая набор учебных данных при настройке параметров нашей модели (весов и смещений).

Мы знаем, что наклон графика функции – производная от функции по переменной. Это наклон всегда указывает на ближайшую впадину! Если функция от нескольких переменных, то поиск идет локального минимума.

Вот есть функция ошибок и сумма по всем ошибкам дает функционал ошибки, которую хотим минимизировать. Чем меньше, тем меньше ошибка. Градиентный спуск один из методов минимизации.

Детерминированное уравнение и тут сразу формула. А у градиента там алгоритм. минус это обратная матрица долго считать