Линейные модели. Линейная регрессия.

Часть 1

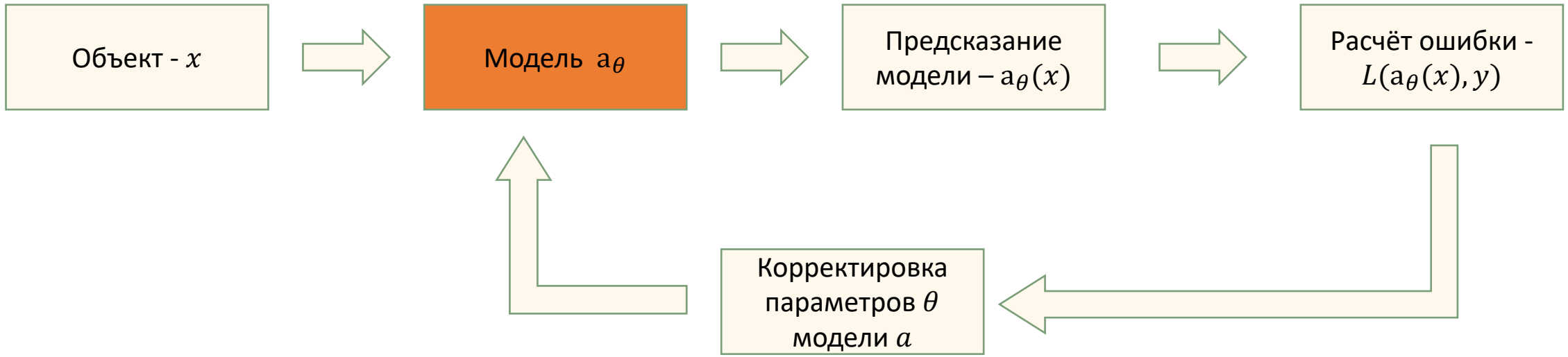
Сбертех, МФТИ

Процесс обучения модели

- Модель – параметрическое семейство функций $A = \{a_\theta(x) | \theta \in \Theta\}$, где a: $X \times \theta \to Y$- функция, Θ – множество допустимых параметров θ. Обозначается как a_θ
- Целевая переменная y. Обладает скрытой зависимостью с x. Задача – аппроксимировать y с помощью a_θ
- $L(a_\theta(x), y)$ – величина ошибки модели a_θ на объекте x относительно целевой переменной y
  - Функция ошибки на задаче классификации – $L(a_\theta(x), y) = [a_\theta(x) \neq y]$
  - Функция ошибки на задаче регрессии $-L(a_\theta(x), y) = (a_\theta(x) - y)^2$
- **Эмпирический риск** – усредненное значение функции $L(a_\theta(x), y)$на подвыборке X размера n

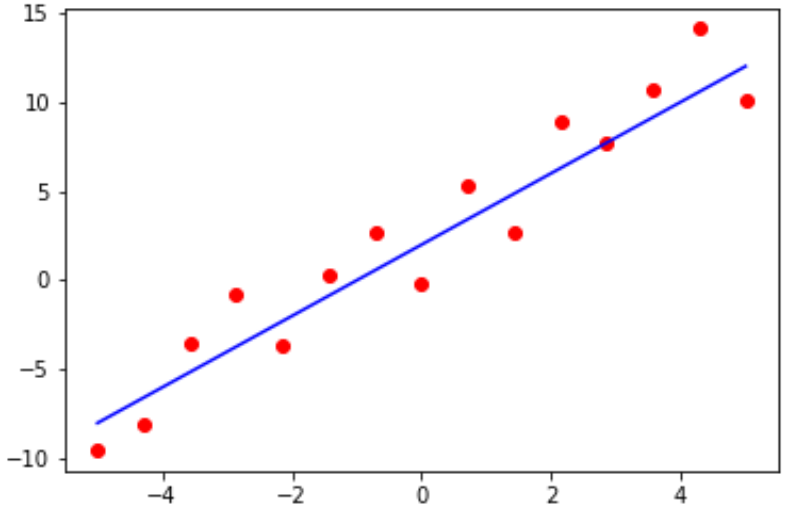
$$\widehat{R_n} = \frac{1}{n}\sum_{i=1}^{n} L(\mathrm{a}_\theta(x_i), y_i) \to min$$

- Задача оптимизации – $\arg\min_{a \in A} \frac{1}{n}\sum_{i=0}^{n} L(a_\theta(x_i), y_i)$



Регрессионные модели

- Пусть дан некоторый вектор $X \in R^n$. Задача регрессионных моделей предсказать некоторое действительное число на основании параметров модели и значений вектора x.
- В вероятностной постановке задачи – задача моделирования распределения $p(t|x)$ – дискриминативная модель
- Как и для всех задач машинного обучения – задача регрессионных моделей минимизация эмпирического риска.



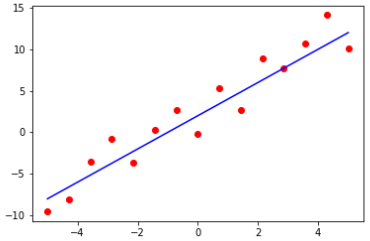
Линейные модели

Пусть есть линейная функция:

$$f(x,\theta) = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + \theta_3 * x_3 + .. + \theta_D * x_D + \epsilon$$

Где $(x_0, x_1, \dots, x_D)$ – объект $x \in R^D$, $(\theta_0, \theta_1, \dots, \theta_n)$ - вектор обучаемых параметров $\theta \in R^d$, а ϵ – нормально распределённая шумовая компонента$(0, \sigma^2)$. Задача алгоритма линейной регрессии – подобрать вектор обучаемых параметров, так чтобы значение $L(f(x,\theta), y)$ было минимальным.

В реальной жизни редко можно встретить линейную комбинацию параметров и объектов – чаще всего под x имеется ввиду некоторая базисная функция $\varphi_j(x)$от реального значения x . Например, базовая функция может быть полиномом - $\varphi_j(x) = x^j$ - тогда модель называется полиномиальная регрессия. Здесь Мы будем рассматривать $\varphi_j(x) = x$ – линейную регрессию.



Задача – **подобрать параметры прямой таким образом, чтобы среднее расстояние от всех объектов до прямой было минимальным**

$$L(a,x) = \frac{1}{N}\sum_{i=0}^{N}\left(a_\theta(x_i) - y(x_i)\right)^2 \to min$$

Минимизация отклонений линейной функции от объектов выборки

Как можно подобрать оптимальный набор параметров?

- Метод наименьших квадратов(МНК) – минимизация суммы квадратов отклонений предсказаний модели от истинного значения
- Метод максимального правдоподобия(ММП) – оценка неизвестных параметров модели путем максимизации функции правдоподобия

Линейные модели

- **Интерпретируемость** - с ростом значения признака, вес значения целевого значения будет увеличиваться, при отрицательном – наоборот.
  - Но для интерпретации весов необходим контроль. Если подобранный вес **маленький/большой**, то не всегда значит что признак **не важен/важен**. Для оценки влияния признаков на целевую переменную нужны статистические критерии.
  - Подобранные веса модели меняются с ростом выборки – но в определённый момент сойдутся к оптимальному значению.

- **Простота** – в линейных моделях есть только 1 механизм оценки признаков – вес для каждого из них.
  - Можно самому генерировать много признаков – модель определить какие из них важны/не важны.
  - Отсутствие защиты от мультиколлинеарности - если в модели есть линейная зависимость между признаками – подобранные коэффициенты модели могут потерять смысл и выдать некорректное значение на тестовой выборке.

Основные понятия статистики

http://cs229.stanford.edu/notes2021spring/notes2021spring/friday_lecture2.pdf

Математическое ожидание – среднее значение случайной величины, f_X - функция задающая распределение этой случайной величины

$$E(x) = \sum xp(x) \qquad E(x) = \int_{-\infty}^{\infty} x f_X(x) dx$$

Дисперсия – мера разброса случайной величины относительно математического ожидания.

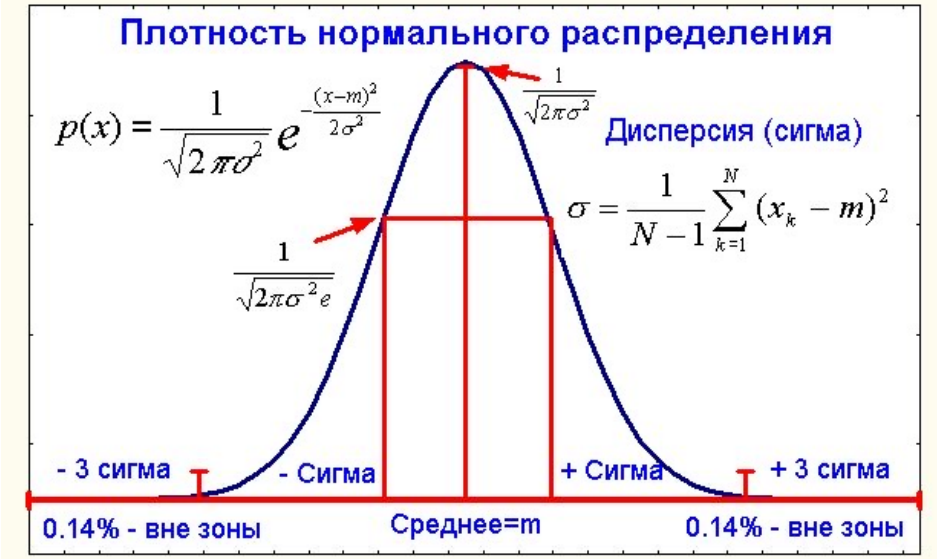
$$D(x) = E[(x\ - E(x)^2] = E(x^2) - \big(E(x)\big)^2$$

Ковариация – мера линейной зависимости двух величин.

$$cov(X, Y) = E[(X\ - E(X)(Y - E(Y)))]$$

Если X_n и Y_n – выборки множеств X и Y тогда:

$$cov(X, Y) = \frac{1}{n} \sum_{t=1}^{n} (X_t\ - \bar{X}) * (Y_t\ - \bar{Y})$$



Корреляция – статистическая взаимосвязь двух величин, когда изменение одной из величин соответствует к закономерному изменению другой величины. Область значений – от -1 до 1.

$$\mathrm{R}(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_y} \qquad \sigma = \sqrt{D(x)}$$

Решение линейной регрессии в матричном виде

Пусть есть - матрица X: $n * d$(n-кол-во объектов, d – количество признаков)(**design matrix**), вектор параметров θ: $1 * d$,
y: $n * 1$(n-количество объектов)

$$\mathrm{X} = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix} \qquad \theta = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_n \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$$

Уравнение линейной регрессии задается в матричном виде задается следующим образом:

$$y = X\theta$$

Значит,

$$X\theta - y = \begin{bmatrix} x^{(1)}\theta^T \\ \dots \\ x^{(n)}\theta^T \end{bmatrix} - \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x^{(1)}\theta^T - y_1 \\ \dots \\ x^{(n)}\theta^T - y_n \end{bmatrix}$$

Так как $\sum_{i=1}^{N} z_i^2 = z^T z$, мы можем представить квадратичную ошибку линейной регрессии так:

$$L_\theta(a, x) = \sum_{i=0}^{N} \big(a_\theta(x_i) - y(x_i)\big)^2 = \frac{1}{2}(X\theta - y) * (X\theta - y)^T$$

Решение линейной регрессии в аналитическом виде

Продифференцируем нашу функцию ошибки по вектору параметров θ:

$$\nabla L_\theta = \frac{1}{2}\nabla_\theta (X\theta - y) * (X\theta - y)^T$$

$$= \frac{1}{2}\nabla_\theta \left((X\theta)^T X\theta - (X\theta)^T y - y^T X\theta + y^T y\right)$$

$$= \frac{1}{2}\nabla_\theta (\theta^T (X^T X)\theta - y^T X\theta - y^T X\theta + y^T y)$$

$$= \frac{1}{2}\nabla (\theta^T (X^T X)\theta - 2(X^T y)^T \theta + yy^T)$$

$$= \frac{1}{2}\left(2(X^T X)\theta - 2(X^T y)\right)$$

$$= X^T X\theta - X^T y$$

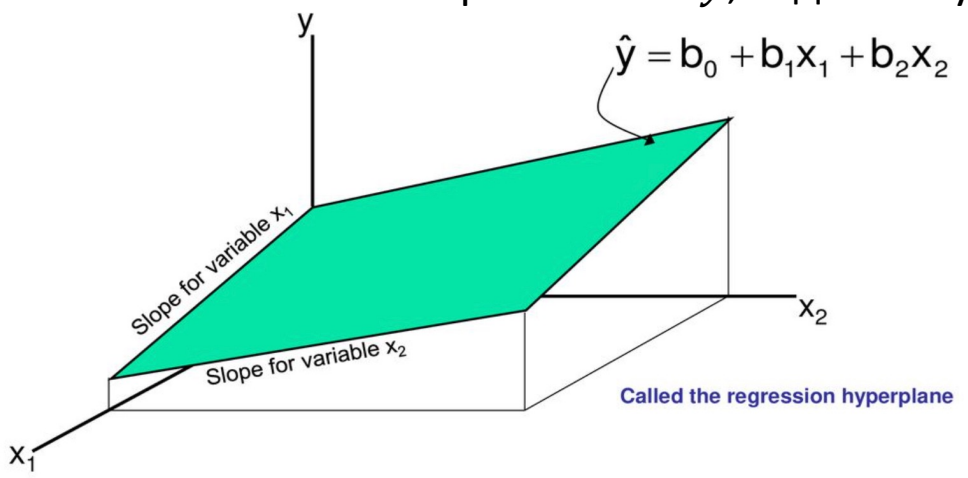
Приравняем полученное выражение к 0(для минимизации функции) и получаем выражение для расчёта параметров линейной регрессии :

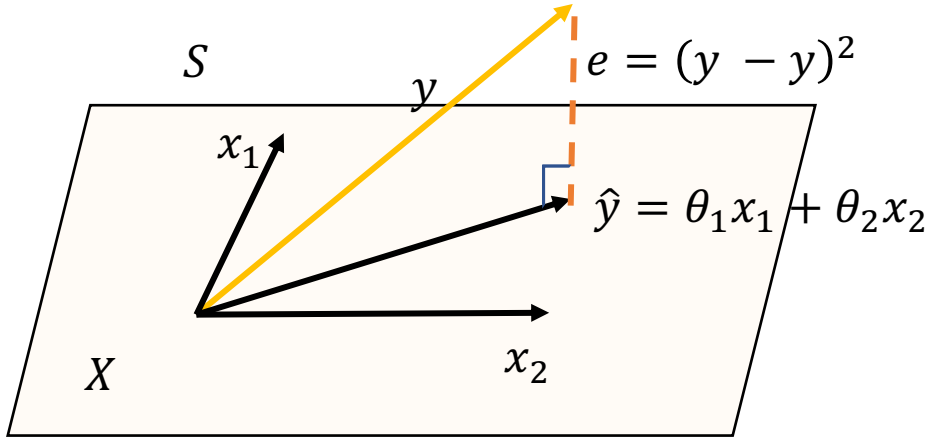
$$X^T X\theta - X^T y = 0$$

$$\boldsymbol{\theta^T = \left(X^T X\right)^{-1} X^T y}$$

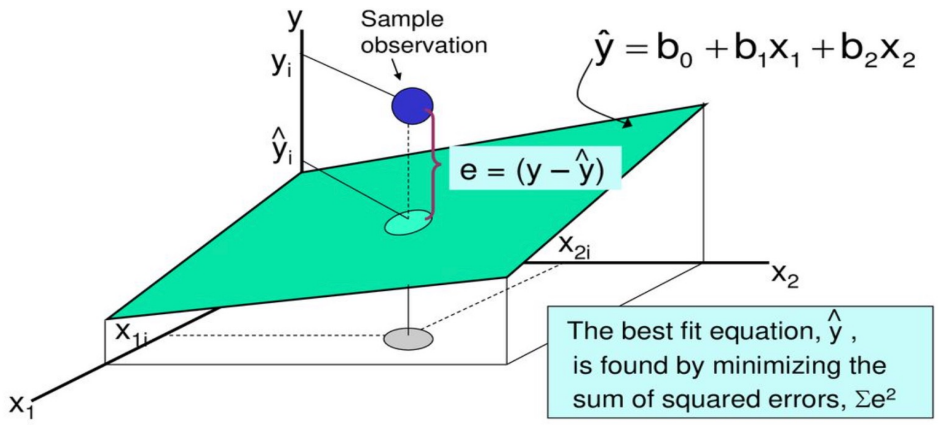
Геометрическая интерпретация линейной регрессии

Гиперплоскость $\hat{y}$, заданная уравнением





S – векторное пространство признаков заданное признаками X
y – вектор зависимых переменных



1) Мы не можем описать y нашими признаками, поэтому мы аппроксимируем значение y взвешенными значением наших признаков в нашем векторном пространстве $\hat{y}$
2) Расстояние e между $\hat{y}$ и y будет задаваться евклидовым расстоянием
3) Нам нужно оптимизировать e по параметрам θ так чтобы e было минимально

Мультиколлинеарность

$$Salary = \theta_0 + \theta_1(Age) + \theta_2(YoE) + \epsilon$$

- С увеличением YoE увеличивается Age
- Изменение в параметре θ_2, ведут к изменению в параметре θ_1
- В таком случае, параметры модели θ_1 и θ_2 могут быть очень большие в масштабах
- Большие параметры θ_1 и θ_2 ведут к большим изменениям в $Salary$ при небольших изменениях в Age и YoE
- Получаем большую дисперсию в предсказаниях, как следствие неуверенность на тесте(переобучение)
- Обычно у нас зависимо несколько признаков – частичная мультиколлинеарность

При решении линейной регрессии в аналитическом виде, наличие мультиколлинеарности (полной/частичной) ведет к нестабильному расчёту θ

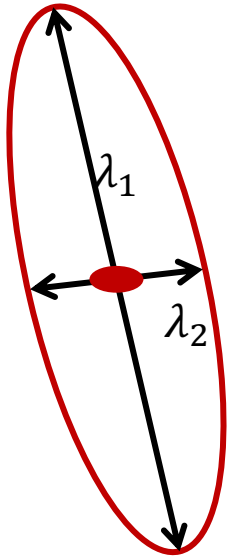
Мультиколлинеарность

$$A^{-1} = \frac{1}{|A|} A_*^T$$

$$\boldsymbol{\theta^T} = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T y}$$

Нормализация - $x_i = \frac{x_i - \mu_i}{\sigma}$

- Пускай признаки $\boldsymbol{X}$ – нормализованы ($X \in N(0,1)$), значит $\left(\boldsymbol{X^T X}\right)$ - матрица корреляции между признаками
- При мультиколлинерности матрица $\left(\boldsymbol{X^T X}\right)$ - становится приблизительно вырожденной, хотя без нее $\left(\boldsymbol{X^T X}\right)$ - полного ранга
- Признаки мультиколлинерности - $X\theta \approx 0$:
  - **Высокая корреляция между признаками** – скорей всего есть линейная зависимость
  - **Высокое число обусловленности** - $k(X) = \frac{\lambda_{max}(X)}{\lambda_{min}(X)}$(чем больше корреляция, тем меньше собственные значения -> тем выше веса)
    - Если мы рассмотрим квадратичную функцию ошибку, то она будет соответствовать эллипсоидным линиям уровня, форма которого будет задаваться квадратичной формой $\boldsymbol{X^T X}$



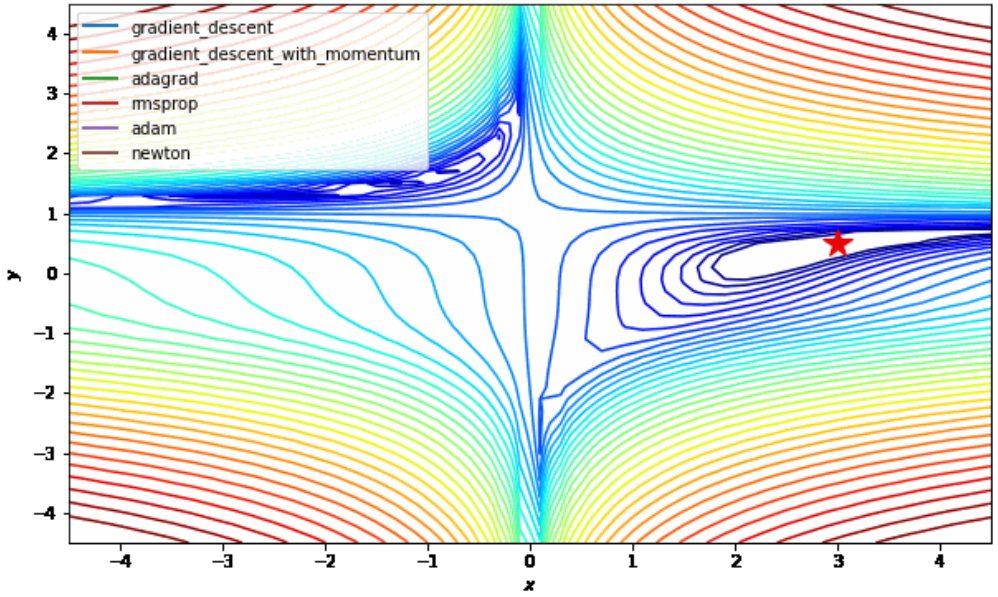
- Наличие частичной мультиколлинерности ведет высокой оценки дисперсии значений признаков модели(переобучению)
- Решение линейной регрессии в матричной форме используется редко, из-за вычислительной нестабильности(+ при большом количестве объектов сложно считать A^{-1}

https://www.ime.unicamp.br/~dias/John%20Neter%20Applied%20linear%20regression%20models.pdf, стр 216

Итеративное решение линейной регрессии

Вычисление больших матриц за раз может быть вычислительно сложно для больших матриц. Поэтому возможно вычислять матрицу итеративно. Например алгоритмами стохастического градиентного спуска или методами второго порядка.

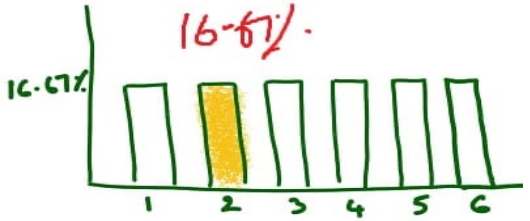
$t \leftarrow 0$
$InitialValue\ \theta_t = 0$
$while\ \theta_{t+1} - \theta_t \geq \epsilon$:
  Выбор N семплов $- (x_i, y_i)_{i=1}^N$
  $AverageGrad = \frac{1}{N} \sum_{i=0}^{N} \nabla_\theta L\ (a_\theta(x_i), y_i)$
  $StepCalculation = A(\mu, AverageGrad)$
  $\theta_{t+1} \leftarrow \theta_t - StepCalculation$
  $t \leftarrow t + 1$



При высоком числе обусловленности – скорость сходимости также уменьшается.

Основные понятия статистики

Распределение плотности вероятности(probability density function, PDF) – закон, который описывает область значений случайной величины и соответствующие вероятности появления этих значений. Является производной от CDF



Функция распределения(cumulative distribution function, CDF) – функция задающая вероятность того что некоторая случайная величина X примет значение меньшее или равное x.

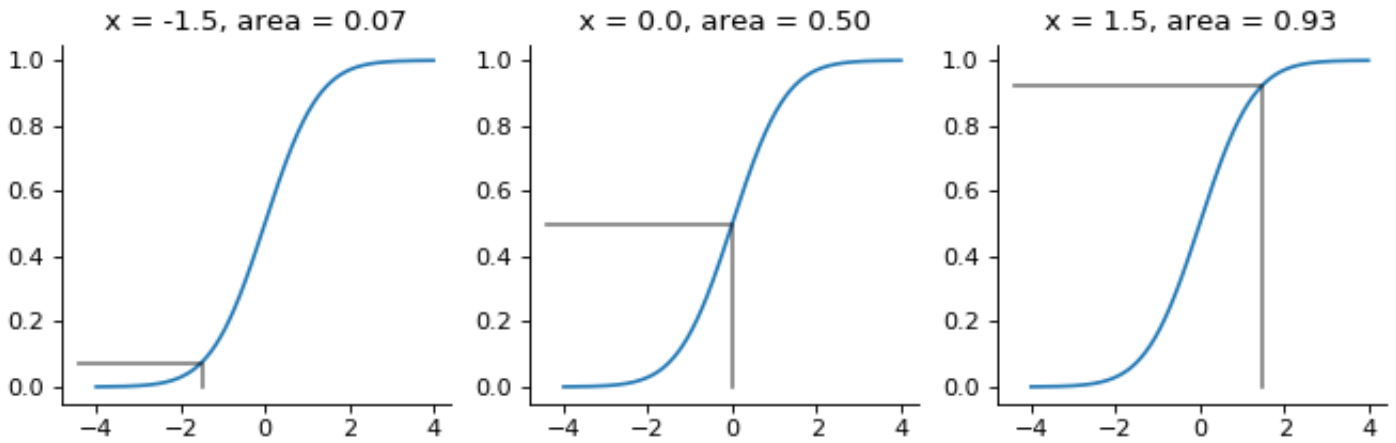
Статистический критерий – предположение о виде распределения и свойствах случайной величины, которое можно либо подтвердить либо опровергнуть с помощью статистических методов.

1. Пускай есть 2 гипотезы – нулевая гипотеза и альтернативная
   - H_0 : Среднее данной выборки равно m
   - H_1: Среднее данной выборки не равно m
2. На основании параметров выборки – размер, мат. ожидание выборки, станд. отклонение считаем статистику выбранного статистического критерия
3. Вычисляем для данной статистики ее статистическую значимость
   - Если статистическая значимость ниже определённого значения – мы отклоняем нулевую гипотезу.

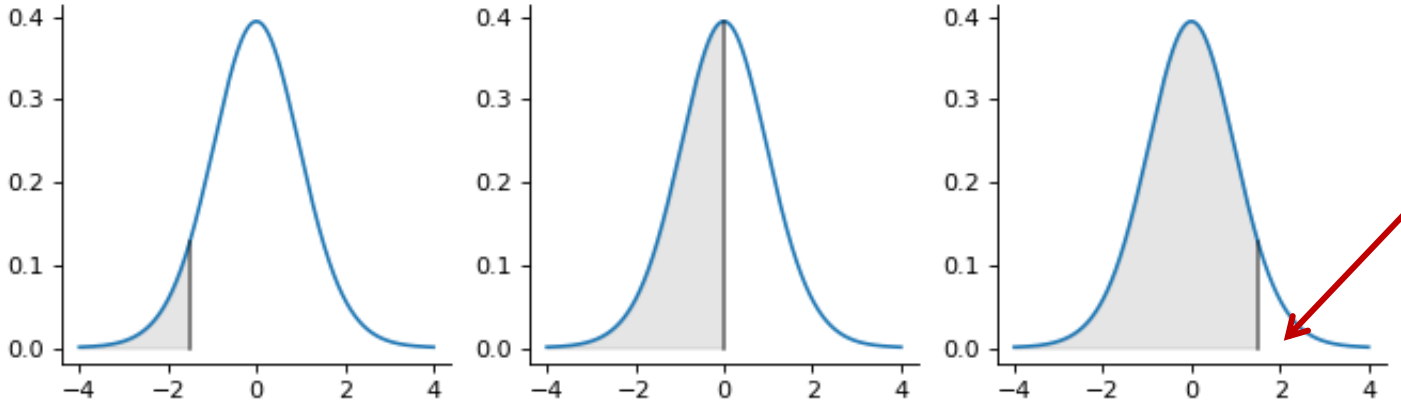
$$t = \frac{\overline{X} - m}{s_X/\sqrt{n}}$$

PDF vs CDF

Функция распределения(CDF)



Распределение плотности вероятности(PDF)



Вероятность для значения принять значение $x < 1.5$, то

$$P(x > 1.5) = 1 - CDF_t(1.5)$$

Статистическая значимость(p-value)

Статистическая значимость(p-value) – вероятность того что отвергнув нулевую гипотезу – мы окажемся не правы. Иными словами – **p-value** – наша уверенность в том что нулевая гипотеза верна.

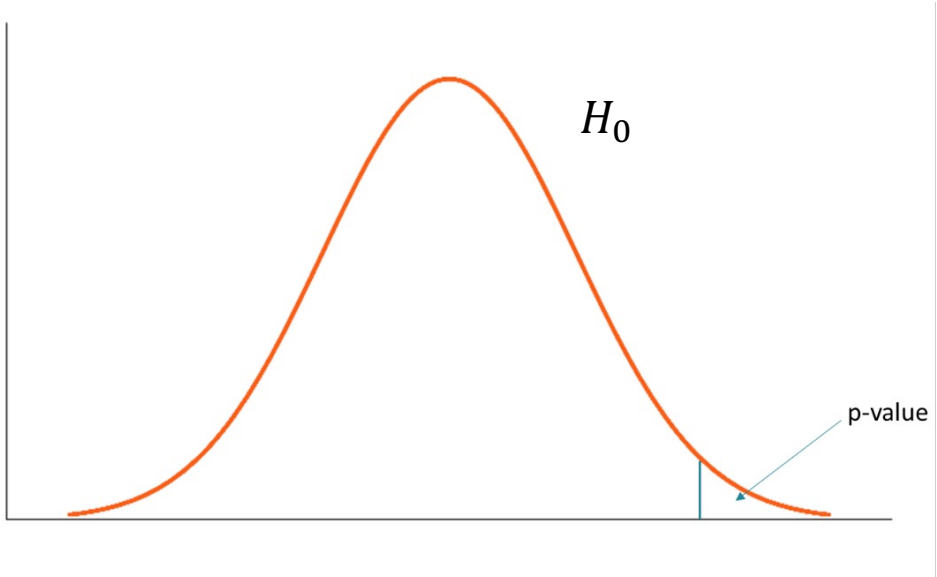
- **p-value** – вероятность отвергнуть правильную H_0 - вероятность совершить ошибку первого рода.

Уровень значимости – допустимое значение вероятности для отвержения нулевой гипотезы. Обычно принимается значение **0.05**.

Алгоритм принятия гипотез

1. Выбор статистической гипотезы и определение H_0
2. Расчёт статистики гипотезы S и определение p-value из распределения статистики(в зависимости от стат. Критерия)
3. Использование распределения статистики для определения статистической значимости:
   - Если p-value маленькое, значит уверенность в нулевой гипотезе маленькая(<= **0.05**) следовательно отвергаем H_0 гипотезу и принимаем альтернативную H_1
   - Если p-value большое значит наша уверенность в H_0 большая следовательно принимаем H_0.

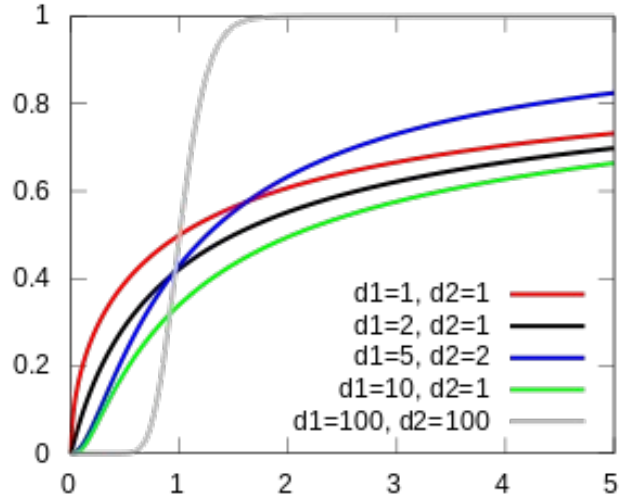
Если у нас будет значение статистики равное или большее чем S, его вероятность, в рамках принятия H_0 очень мала – значит мы отвергаем H_0 и принимаем H_1



Вероятность события H_0 маленькая, значит можно отказаться от H_0 в пользу H_1

Степень свободы

1. *Размер выборочной дисперсии не равен дисперсии генеральной совокупности.*
   1. Допустим есть выборка из некоторой генеральной совокупности. Мы оцениваем ее с помощью среднего. Так как наши объекты находятся вокруг этого среднего – среднее смещено относительно истинного среднего.
   2. С помощью нашего среднего попробуем оценить дисперсию генеральной совокупности. Так как наше среднее смещено относительно истинного, у нас в том числе сместится и дисперсия относительно истинной дисперсии. **В результате дисперсия относительного выборочного среднего смещается в меньшую сторону.**
   3. Нам нужно скорректировать расчёт этой дисперсии
2. *Если мы знаем выборочное среднее, то для подсчета других статистик этой выборки нам будет достаточно n-1 элемент выборки.*
   1. Тогда для оценки несмещенной дисперсии нам нужно разделить сумму отклонений не на n, а на n-1 элементов
3. *Среднее – способ оценки нашей выборки состоящий из одного параметра.*
   1. В случае с линейной регрессией у нас не 1 параметр – среднее – а $n + 1$ параметров $\theta_{1..n}$ которые мы оцениваем – кол-во независимых переменных модели + параметр смещения θ_0. В нашем случае, параметры модели – **степени свободы**.
   2. Расчёт несмещенной дисперсии в линейной регрессии – $n - k - 1$



[* Доказательство необходимости сдвига степеней свободы]

Коэффициент детерминации

$$R^2 = \frac{SS_{mean} - SS_{model}}{SS_{mean}}$$

1) **Интерпретация R^2**
   - насколько хорошо модель “ объясняет “ дисперсию целевой переменной
   - R^2 отвечает за процент наблюдаемой дисперсии в дисперсии целевой переменной

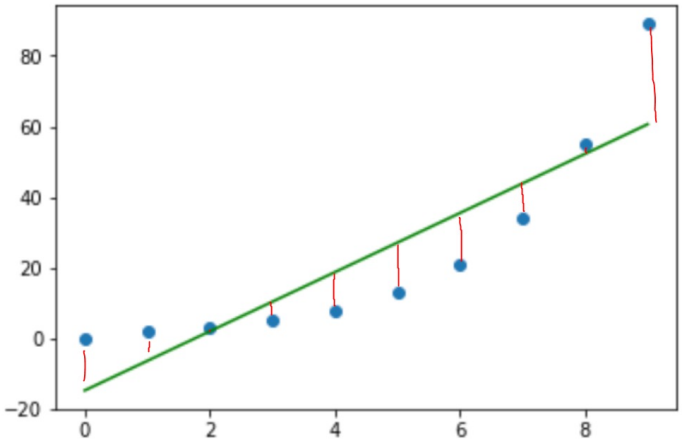
2) Дисперсия выборки $SS_{mean} = \frac{1}{n}\sum(y_i - \bar{y})^2$

3) Среднеквадратичное отклонение предсказаний - $SS_{model} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$

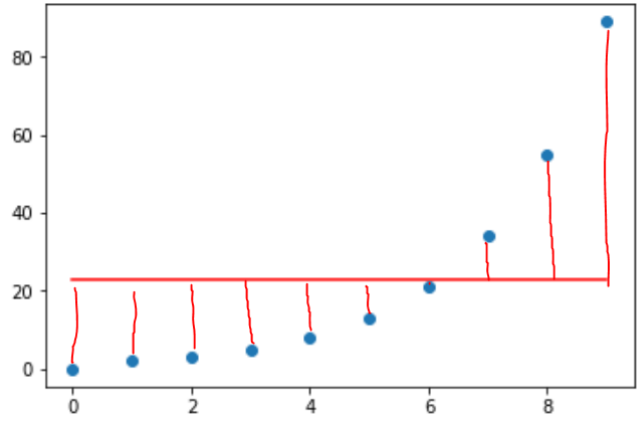
4) $R^2 \in [0; 1],\ R \in [-1; 1]$
   - Чем R^2 ближе к 1, тем лучше модель соответствует данным.

5) Модель линейной регрессии с набором “бесполезных “ параметров не будут ухудшать R^2, только улучшать. **Как статистически оценить важность признаков?**

$$R^2 = 1 - \frac{df_{mean}}{df_{model}}(1 - R^2) = 1 - \frac{(N-1)}{N-k-1}(1 - R^2)$$



$$SS_{model} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$



$$SS_{mean} = \frac{1}{n}\sum(y_i - \bar{y})^2$$

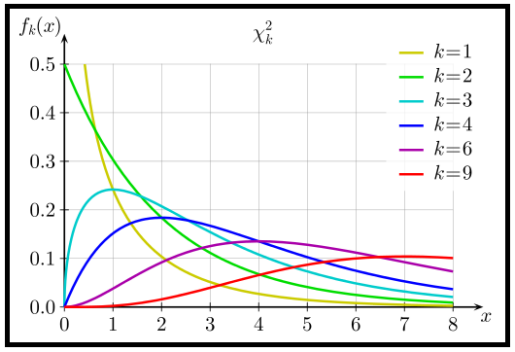
Коэффициент детерминации

F – test / критерий Фишера

$$F = \frac{R^2/(d_{model} - d_{mean})}{(1 - R^2)/(n - d_{model})}$$

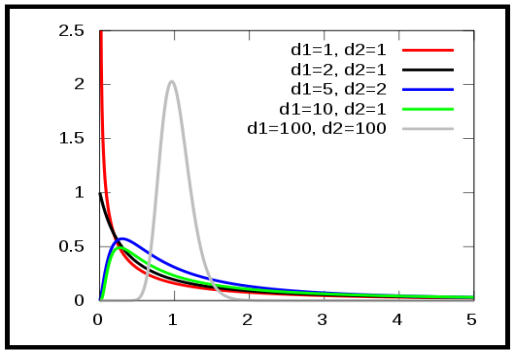
d_{model} - количество параметров модели
d_{mean} - количество параметров оценки дисперсии

1) Гипотеза H_0 – не существует зависимости между целевой переменной и регрессионными параметрами
2) Гипотеза H_1– зависимость между хотя бы одним признаком и целевой переменной существует

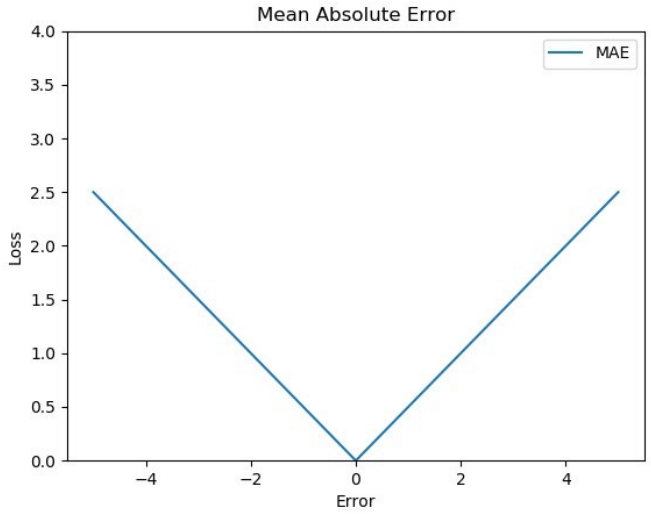


1) Параметры(независимые случайные величины) F-статистики попадают под распределение - χ^2 .
   1) $R^2 \sim \chi^2$
   2) $1 - R^2 \sim \chi^2$
2) F-статистика имеет следующее распределение:

$$X \sim F(d_1, d_2) \sim \frac{\frac{R^2}{d_{model} - d_{mean}}}{\frac{1 - R^2}{n - d_{model}}} \sim F(d_{model} - d_{mean}, n - d_{model})$$



Функции ошибки регрессии



$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i|$$

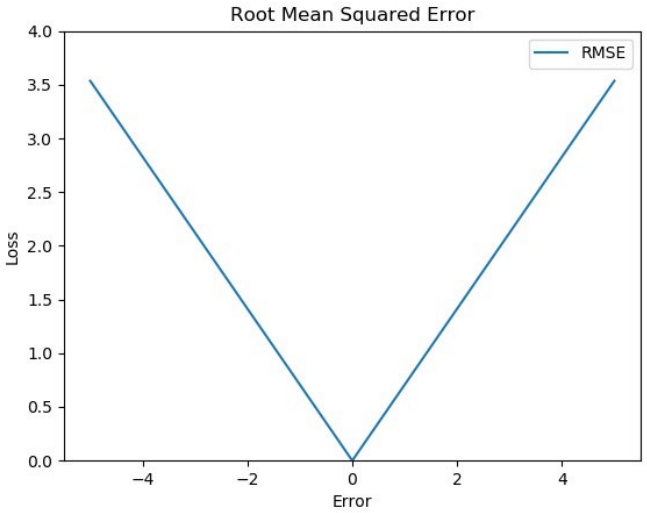
\+ Прост в понимании и вычислении
\+ Устойчива к выбросам
\+ Оптимальное решение - медиана

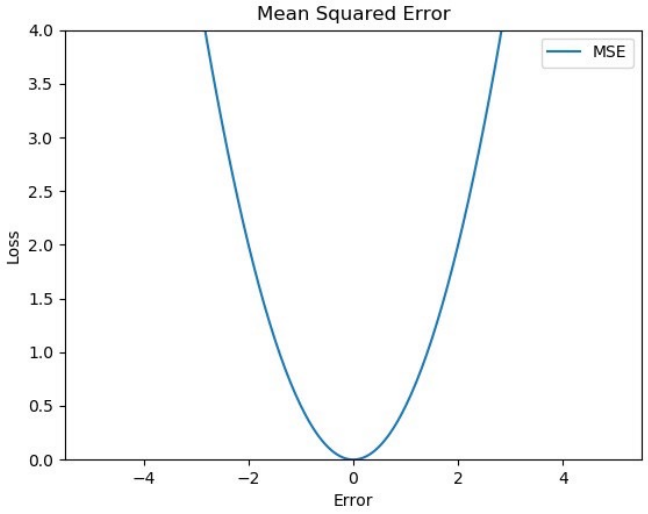
\- В 0 функция не дифференцируема

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$$

\+ Устойчива к выбросам
\+ Интерпретируема

\- В 0 функция не дифференцируема





$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

\+ При небольших ошибках и нормальном LR сойдемся к минимуму
\+ Оптимальное решение - среднее

\- Функция чувствительна к выбросам