Среднеквадратичная ошибка (MSE, mean square error) - часто используемый критерий качества. На чем он основывается, кроме здравого смысла и авторитета Карла Фридриха Гаусса?

Может ли MSE быть получен исходя из общих принципов?
Да может. Минимизация среднеквадратичной ошибки (MSE) - это не только неопределенно интуитивный подход, он может быть получен, используя метод максимального правдоподобия для случая линейной гауссовой модели.

Напоминание

Пусть X случайная величина с абсолютно непрерывным распределением.

Обозначим $f(t)$ плотность распределения случайной величины X.

Распределение известно с точностью до нескольких параметров, то есть плотность распределения может быть записана $f(t, a_1, a_2, \dots, a_k)$, где $a_1, a_2, \dots, a_k$ параметры, при этом какие бы значения ни приписать эти параметрам, функция $f(t, a_1, a_2, \dots, a_k)$ остается плотностью.

Известно ***n*** независимых наблюдений случайной величины X

$x_1, x_2, x_3, \dots, x_n$

В случае абсолютно непрерывного распределения мы не можем использовать то же самое определение, что и в случае дискретного распределения. Дело в том, что вероятность наблюдать именно те значения, которые были зарегистрированы равна нулю.

Вопрос 1 : Почему вероятность должна равняться нулю?

Определение функции правдоподобия для абсолютно непрерывного распределения.

Функция правдоподобия $L = f(x_1, x_2, x_3, \dots, x_n)$

где $f(t_1, t_2, t_3, \ldots, t_n)$ - плотность совместного распределения случайных величин $X_1, X_2, X_3, \ldots, X_n$

Если наблюдения случайной величины X независимые, то имеем

$$L = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} f(x_i, a_1, a_2, \ldots, a_k)$$

Вопрос 2 : откуда появились ***n*** случайных величин?

Вопрос 3 : какие аргументы у функции ***L***?

Вопрос 4 : Использовалась независимость наблюдений. Сохраняется ли независимость после стандартизации?

Определение обыкновенной линейной регрессионной модели

Имеем вектор предикторов $\bar{X} = (X_1, X_2, X_3, \ldots, X_k)$

Имеем отклик Y

Имеет вектор параметров $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)^T$

Рассмотрим линейную модель

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots \beta_k \cdot X_k + \epsilon = .$$
$$. = \beta_0 \cdot X_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots \beta_k \cdot X_k + \epsilon = X \cdot \beta + \epsilon$$

где ошибка $\epsilon \in N(0, \sigma^2)$. Предположим, что σ^2 известно, а ошибки независимы и одинаково распределены.

Пусть имеется набор данных. Вектор наблюдений отклика $\bar{Y}=(y_1, y_2, y_3, \ldots, y_n)$ и таблица наблюдений предикторов *X размерности n x k*.

Обозначим X_i i-ю строку таблицы X. Она содержит наблюдения предикторов для i-го наблюдения.

Обозначим e_i ошибку для i-го наблюдения.

Напомним, что правдоподобие - это вероятность наблюдать данные при фиксированных значений параметров модели $\beta=(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)^T$

Доказательство

Выпишем лог правдоподобие для вектора наблюдений ошибок

$$\epsilon=Y-\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\ldots\beta_k\cdot X_k=Y-\bar{X}\cdot\beta$$

Поскольку $\epsilon\in N(0,\sigma^2)$, плотность распределения ε имеет вид

$$\phi(t)=\frac{1}{\sqrt{2\pi\sigma^2}}\cdot\exp\left(\frac{-t^2}{2\cdot\sigma^2}\right)$$

Выпишем лог-правдоподобие для ошибок модели

$$\log L=\log p_Y(\bar{\epsilon})=\log\prod_{i=1}^{n}\phi(\epsilon_i)=\sum_{i=1}^{n}\log\phi(\epsilon_i)=.$$

$$.=\sum_{i=1}^{n}\log\left(\frac{1}{\sqrt{(2\pi\sigma^2)}}\cdot\exp\left(-\frac{(y_i-X_i\cdot\beta)^2}{2\cdot\sigma^2}\right)\right)=.$$

$$.=-\frac{n}{2}\log(2\pi\sigma^2)-\sum_{i=1}^{n}\frac{(y_i-X_i\cdot\beta)^2}{2\cdot\sigma^2}=.$$

Метод максимального правдоподобия предписывает нам выбрать в качестве параметров модели тот вектор $\beta=(\beta_0,\beta_1,\beta_2,\ldots,\beta_k)^T$ **, для которого функция правдоподобия максимальна или**
лог правдоподобие максимально или
- лог правдоподобие минимально

$$\beta=(\beta_0,\beta_1,\beta_2,\ldots,\beta_k)^T=argmin(-\log L)=.$$

$$.=argmin(\frac{n}{2}\log(2\pi\sigma^2)+\sum_{i=1}^{n}\frac{(y_i-X_i\cdot\beta)^2}{2\cdot\sigma^2})=.$$

$$.=argmin(\frac{1}{2\cdot\sigma^2}\cdot\sum_{i=1}^{n}(y_i-X_i\cdot\beta)^2)=.$$

$$.=argmin(\frac{1}{2\cdot\sigma^2}\cdot\sum_{i=1}^{n}(y_i-X_i\cdot\beta)^2)=.$$

$$.=argmin(\sum_{i=1}^{n}(y_i-X_i\cdot\beta)^2)=.$$

$$.=argmin(\frac{1}{n}\cdot\sum_{i=1}^{n}(y_i-X_i\cdot\beta)^2)=MSE$$