# 线性回归

* 概述

**线性回归(Linear Regression)是利用数理统计中的回归分析，来确定两个或两个以上变量间相互依赖的定量关系的一种统计分析方法。**

**在线性回归中，只包括一个自变量和一个因变量，且两者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。**

**如果回归分析中包括两个或两个以上的自变量，且因变量与自变量之间是线性关系，则称为多元线性回归分析。**

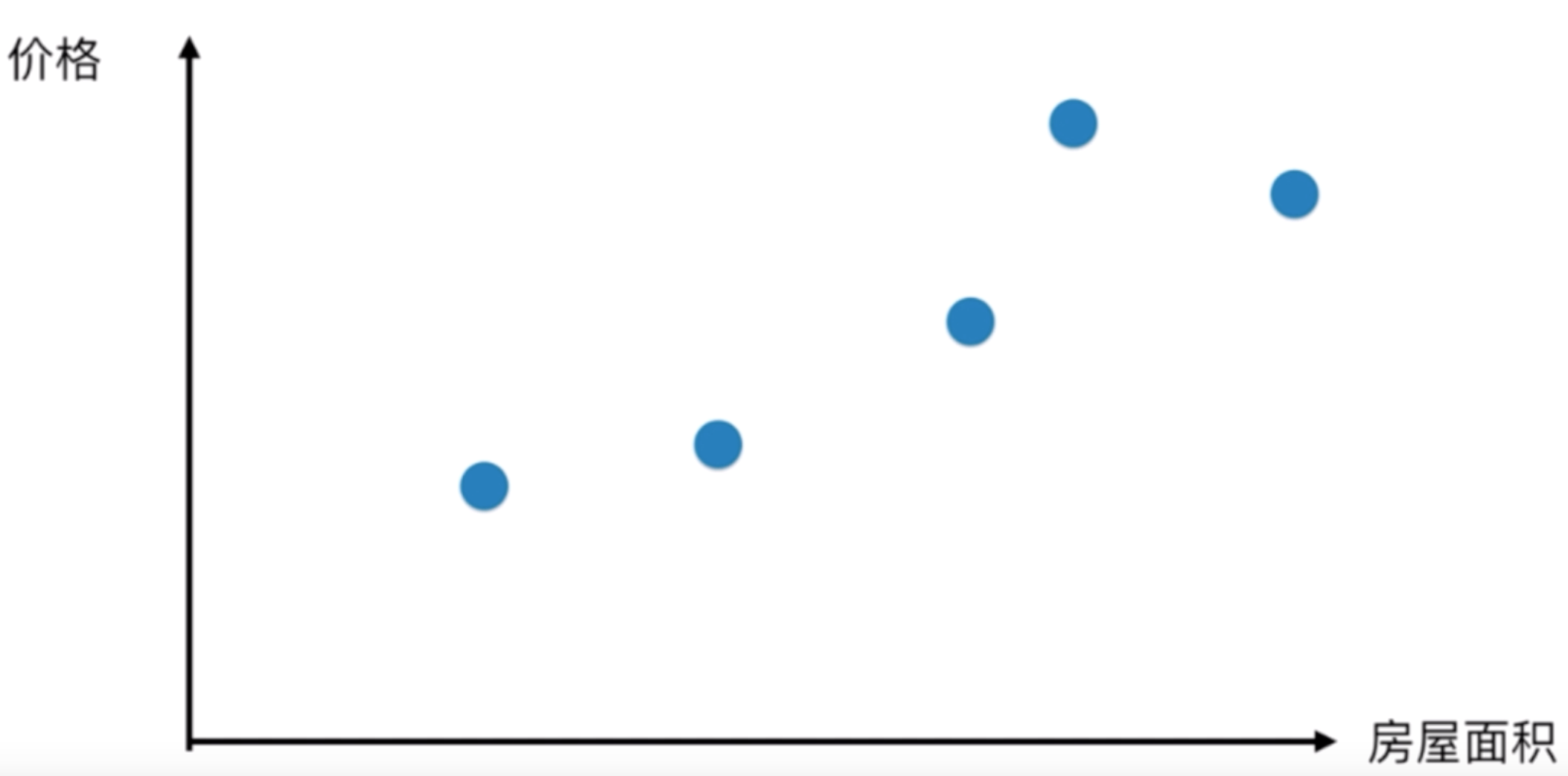
**线性回归算法的特点：**

**1.用于解决回归问题**

**2.思想简单，容易实现**

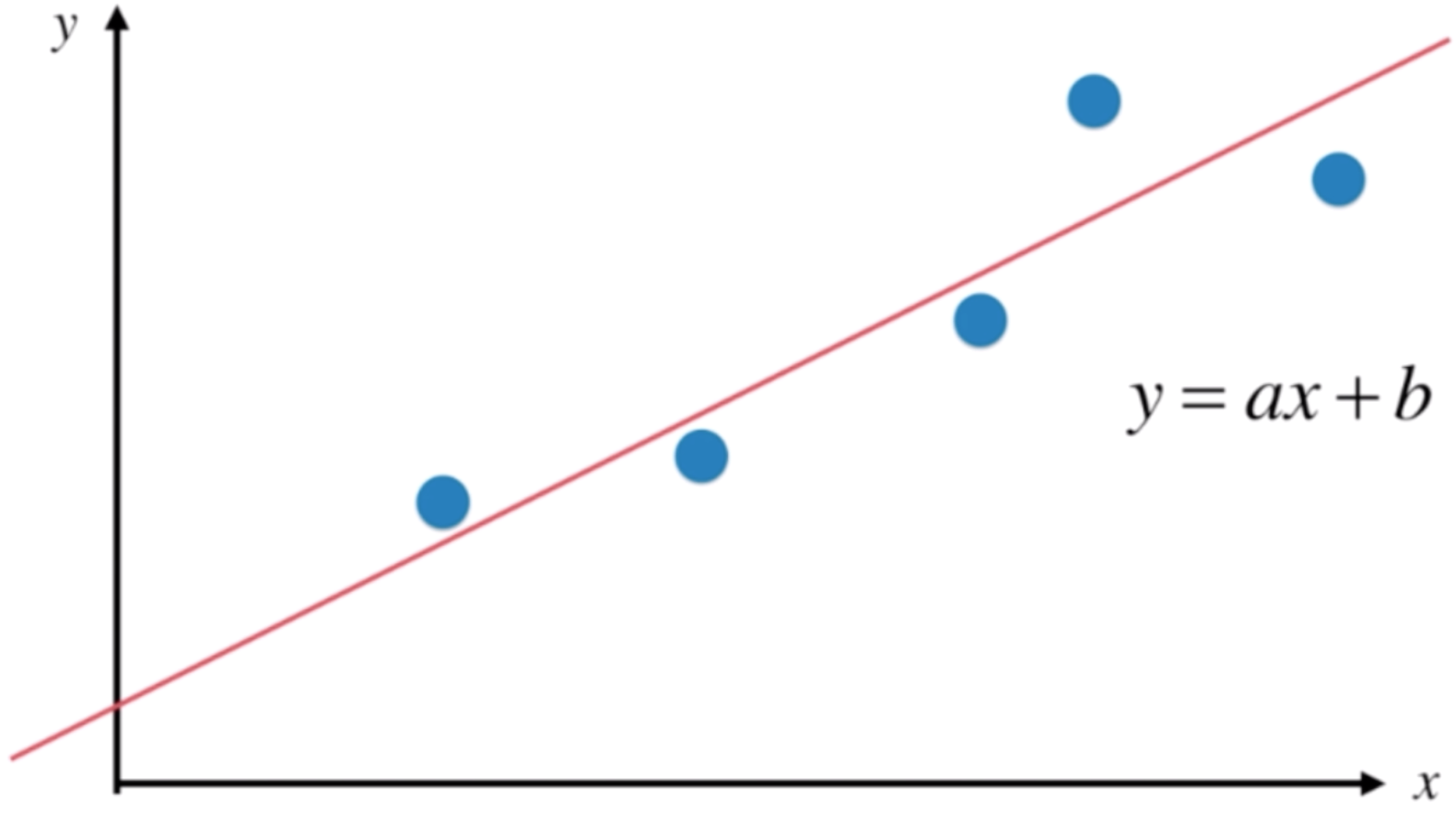
**3.是需要强大的非线性模型的基础**

**4.结果具有很好的解释型**

****

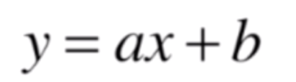
* 简单线性回归

**假设样本特征只有一个，则称为简单线性回归。**

****

**计算方式**

**1.声明假设函数(Hypothesis Function)**

****

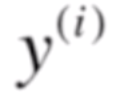
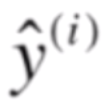
**a、b被称为模型参数，也就需要通过训练集训练得出的数值。**

**假设函数，就是用来进行预测的函数。**

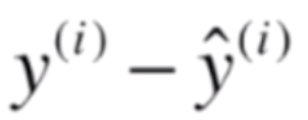
**2.如何使直线更好的拟合数据**

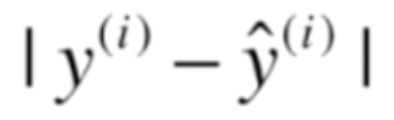
**实际的解决方案就是让预测出的数值与实际的数值更接近。**

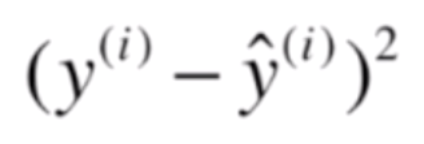
**也就是一个计算最小化的问题。**

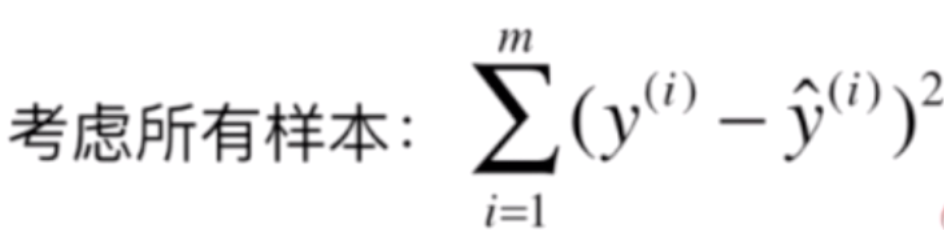
**假设实际值为，预测值为**

**则有如下几种计算预测值与实际值差距的方式**

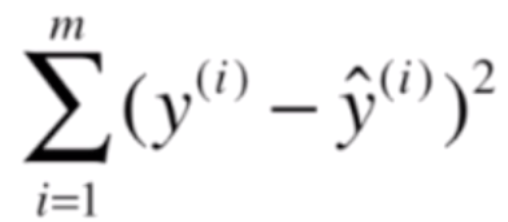
**1.** ****

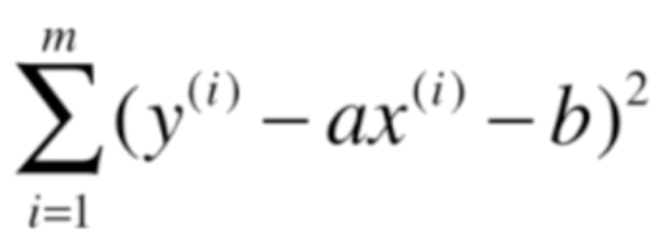
**2.** ****

**3.** ****

****

**3.目标函数(loss function)**

****

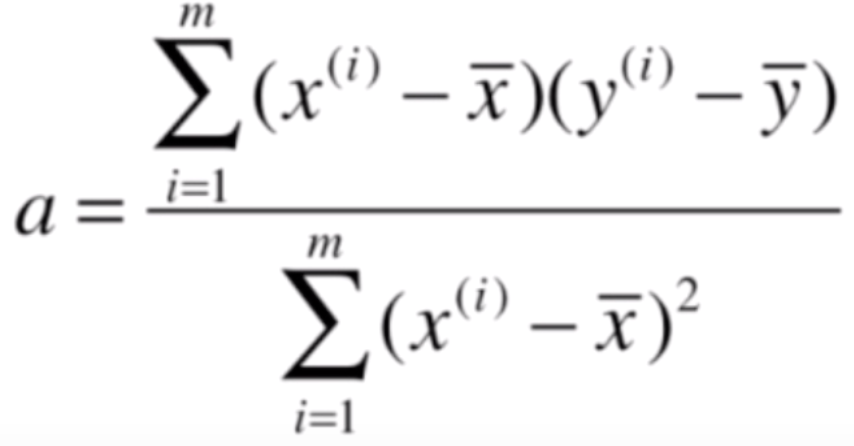
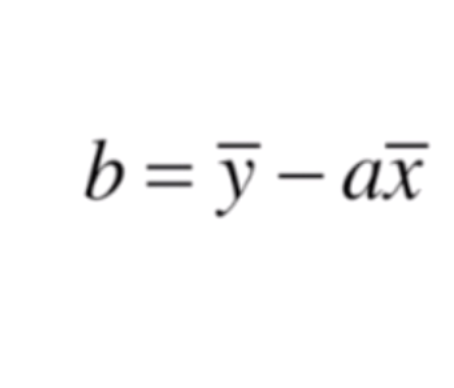
****

**找到a和b，使得目标函数尽可能小**

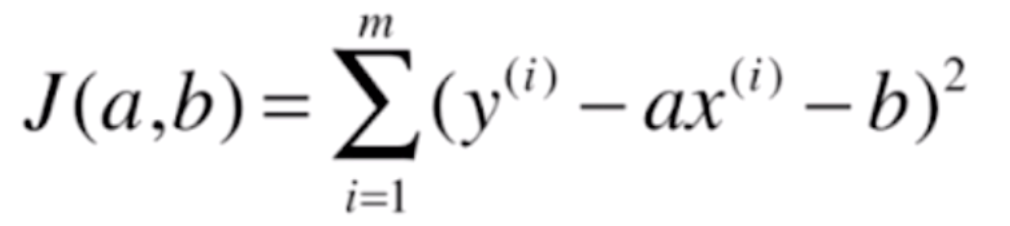
**4.最小二乘法**

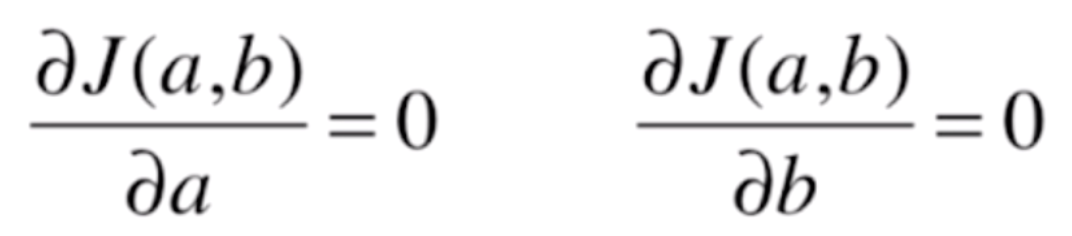
**最小二乘法，又称最小平方法。是一种数学优化技术，它通过最小化误差的平方来**

**寻找拟合数据的最佳参数。**

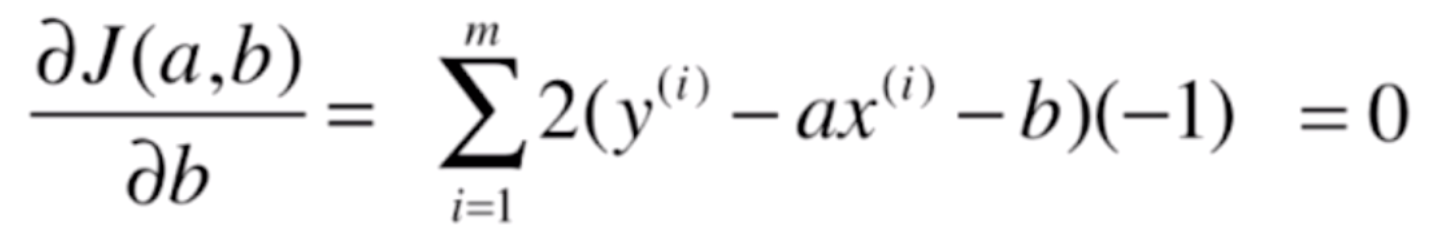
**** ****

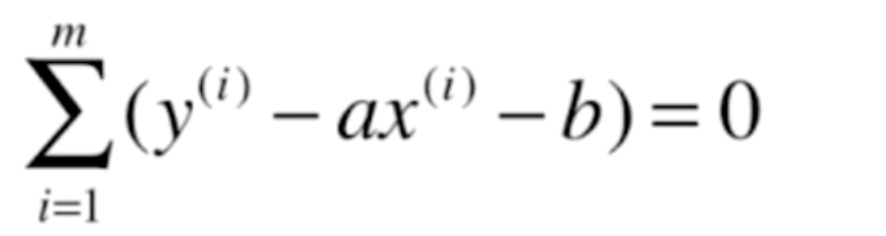
* 最小二乘法

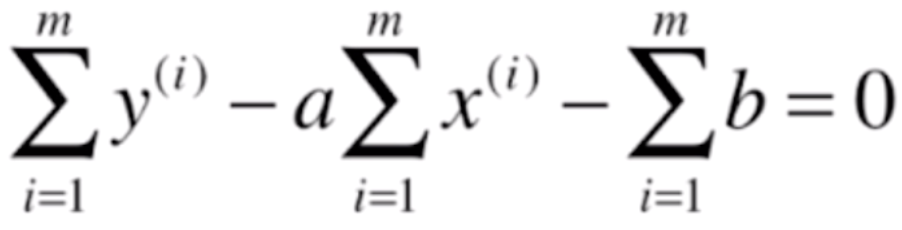


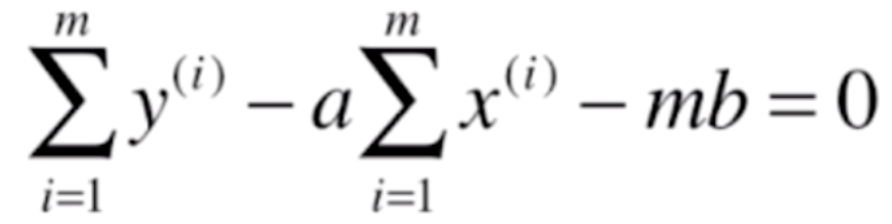


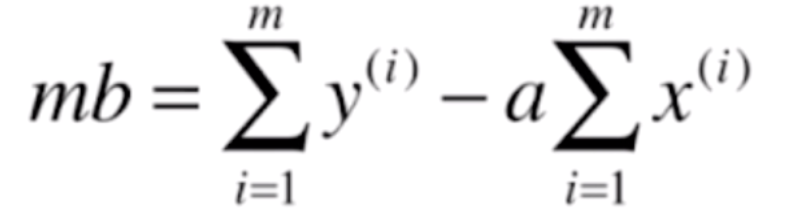
**对b求偏导**

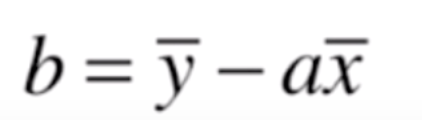




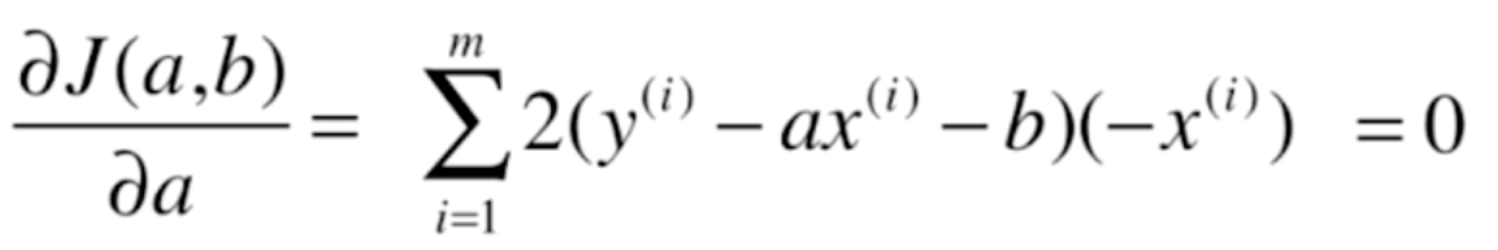


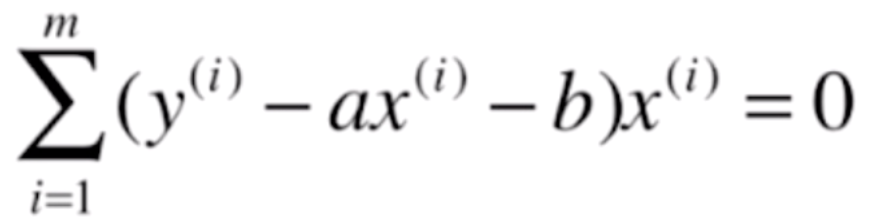




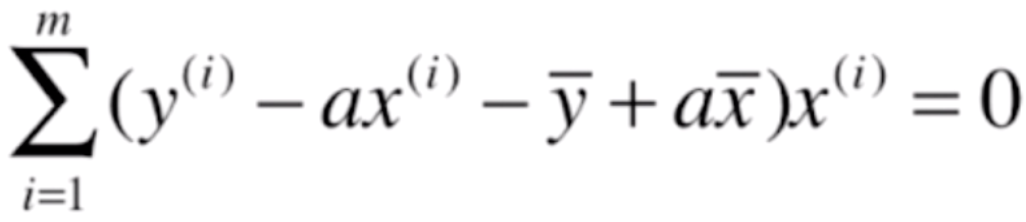


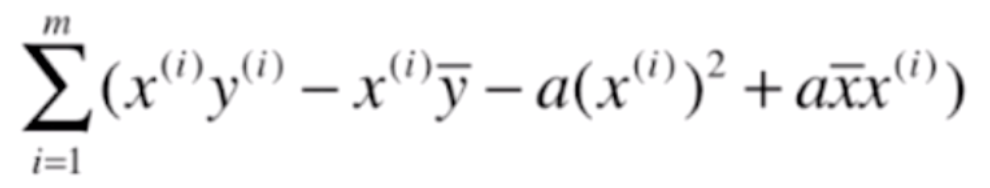
**对a求偏导**

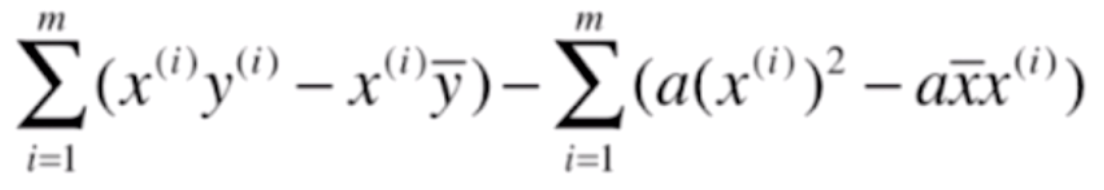




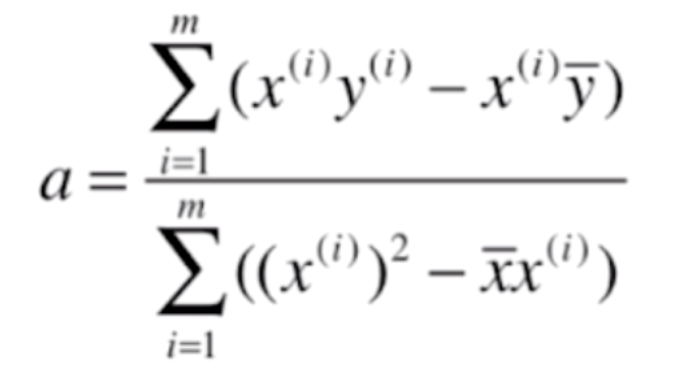
**代入b**

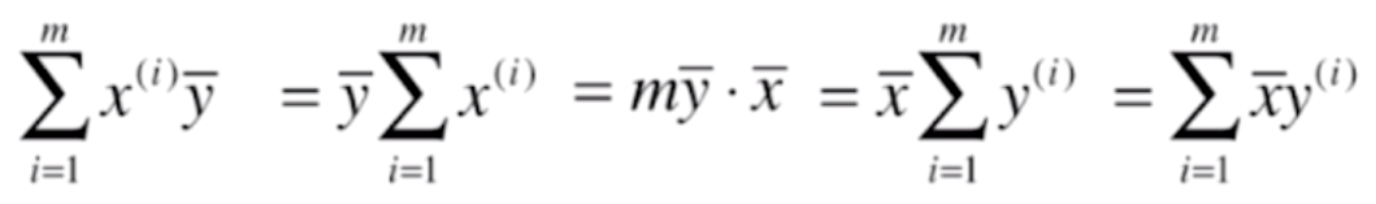
****

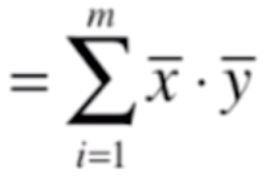
****

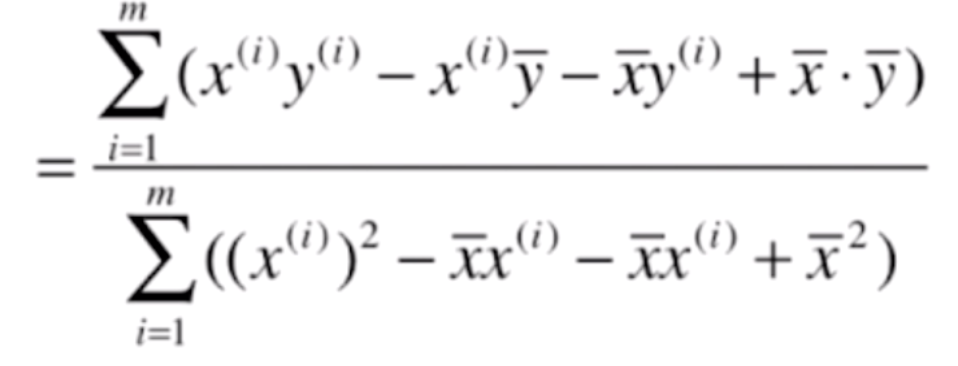
****

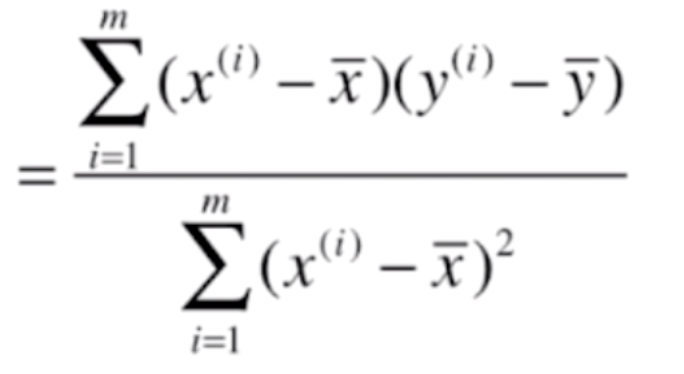




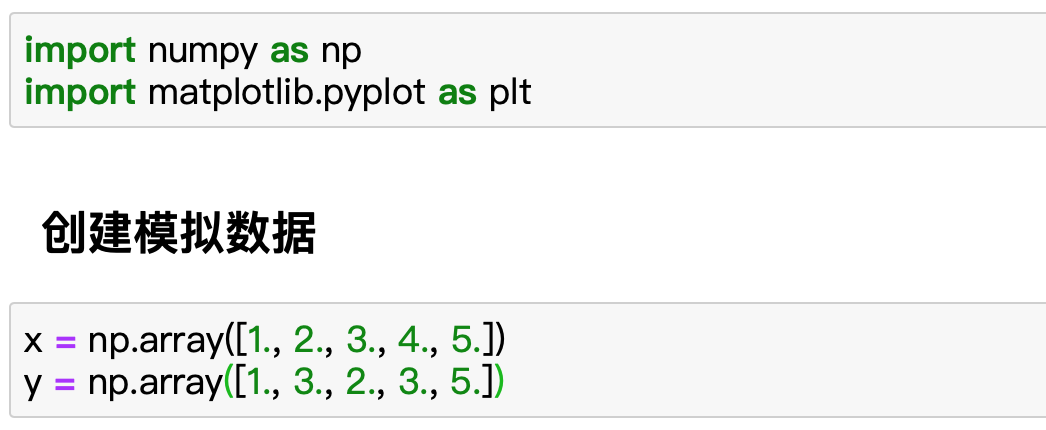


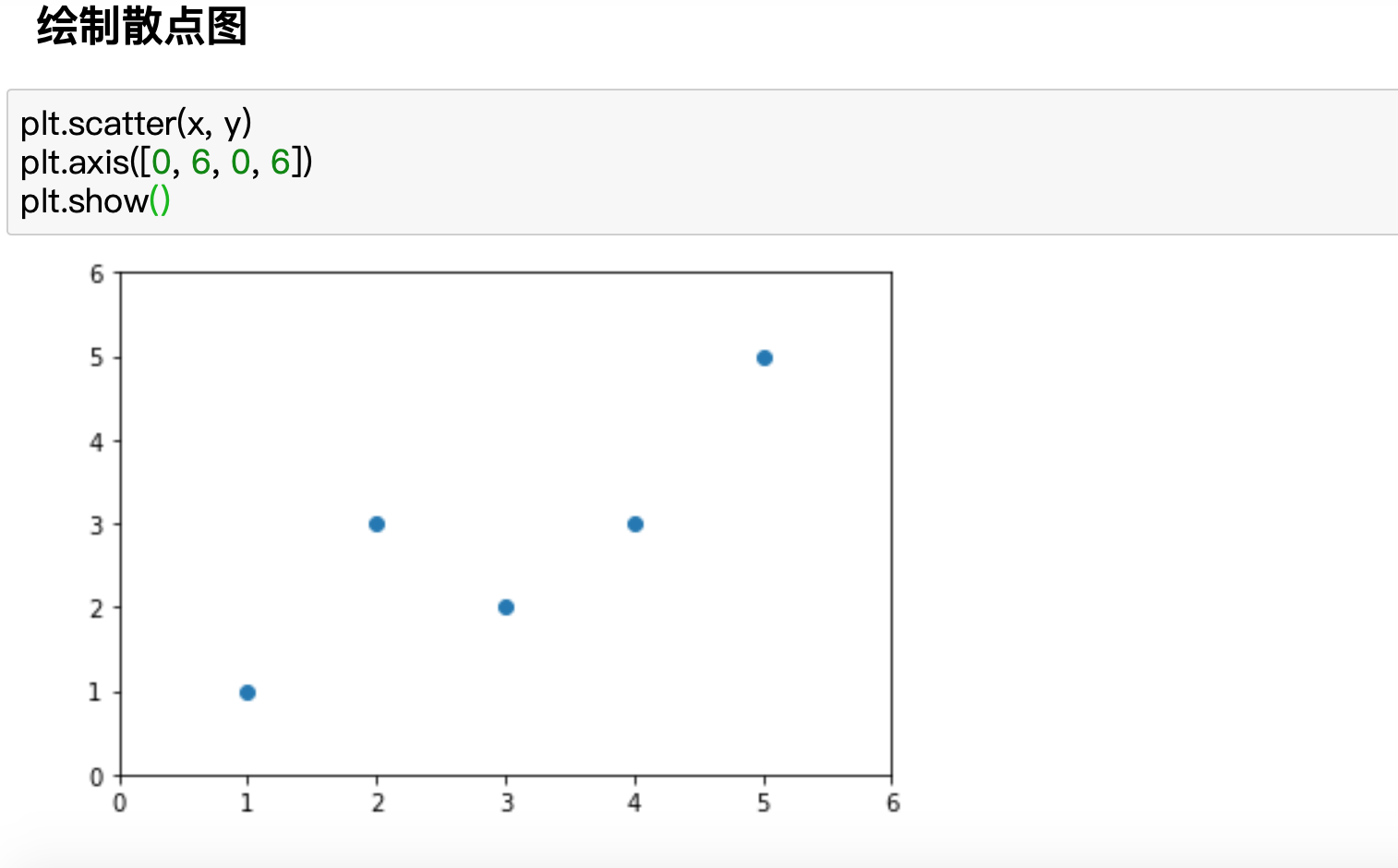


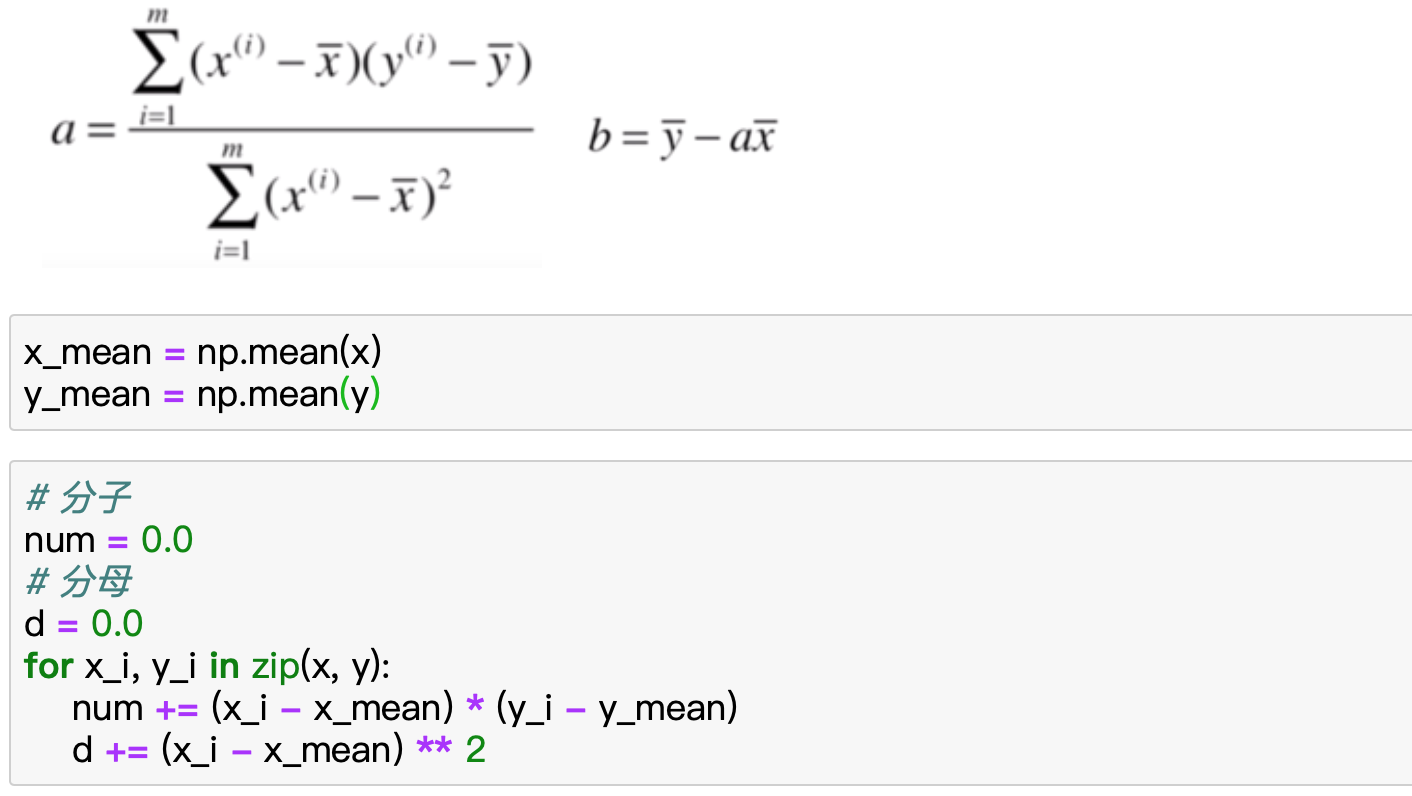


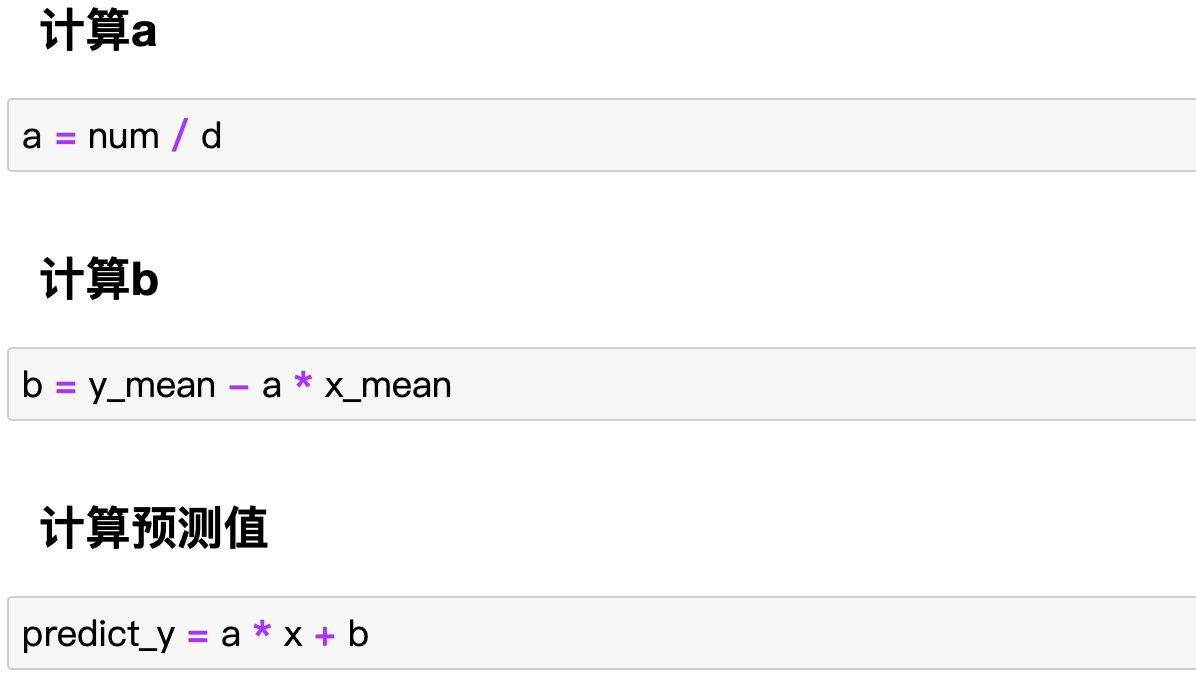


* 简单线性回归实现

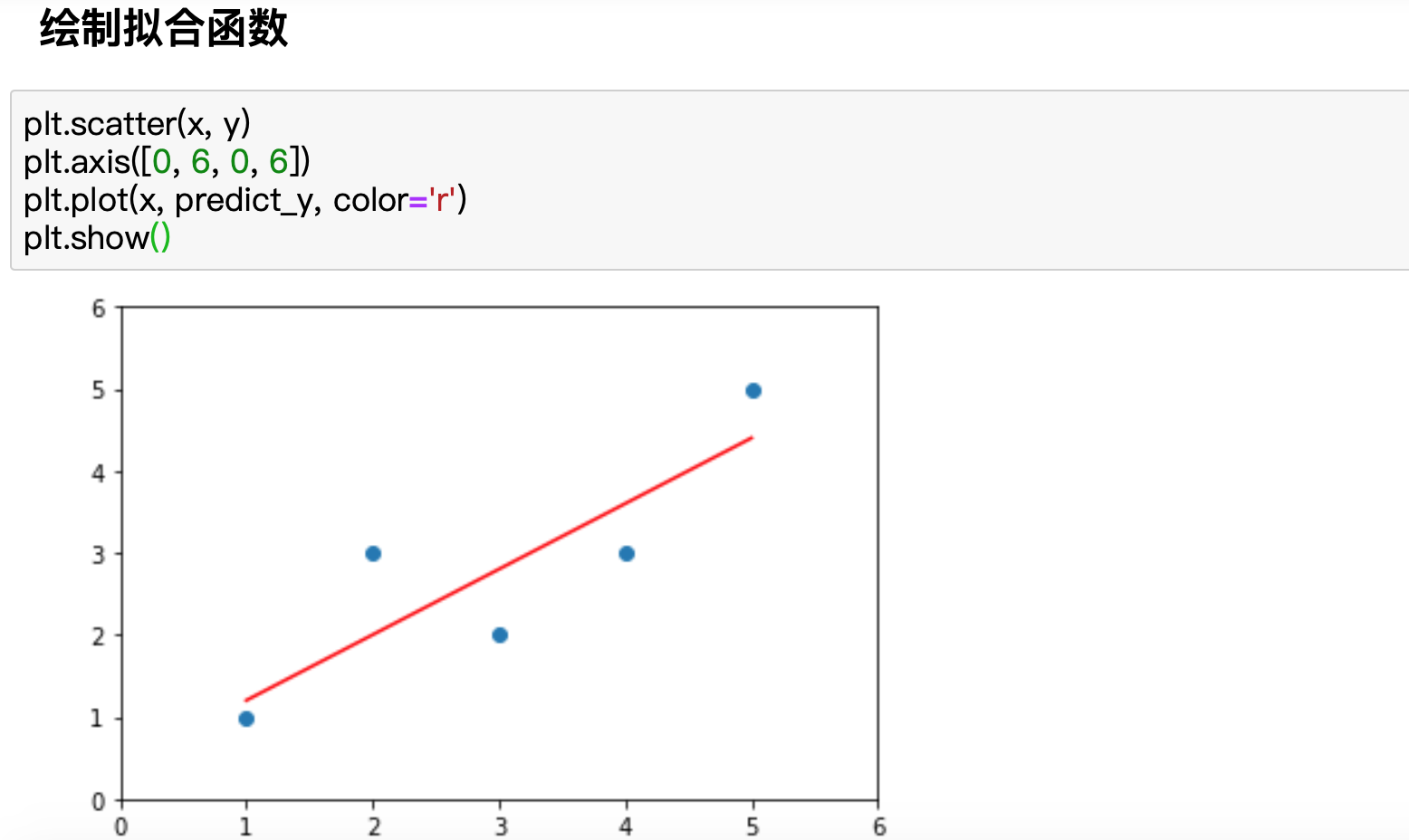


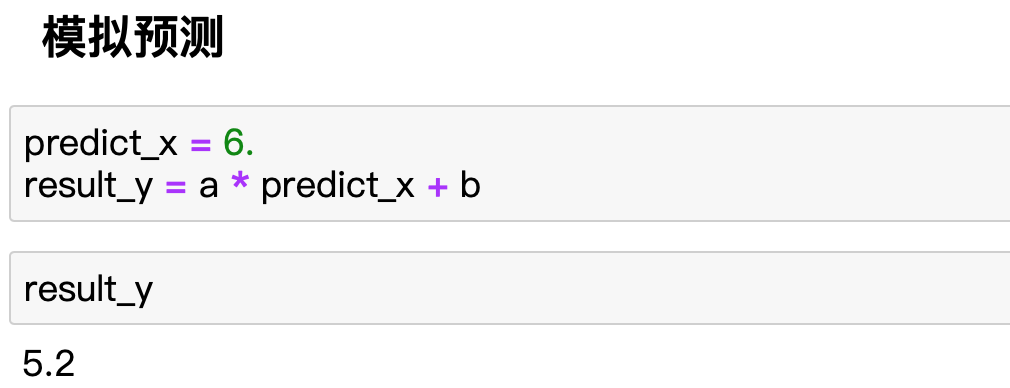




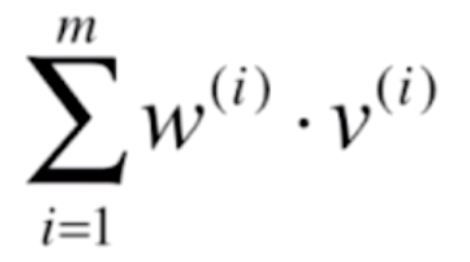
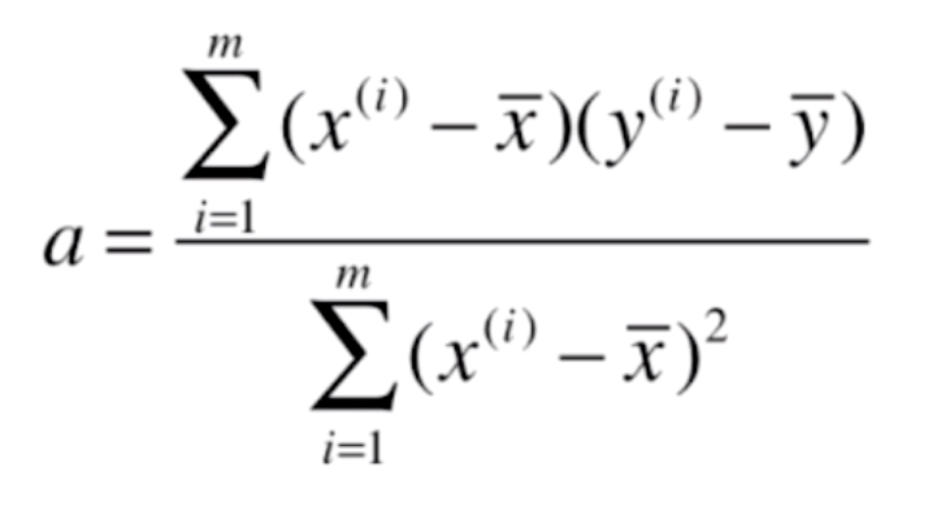


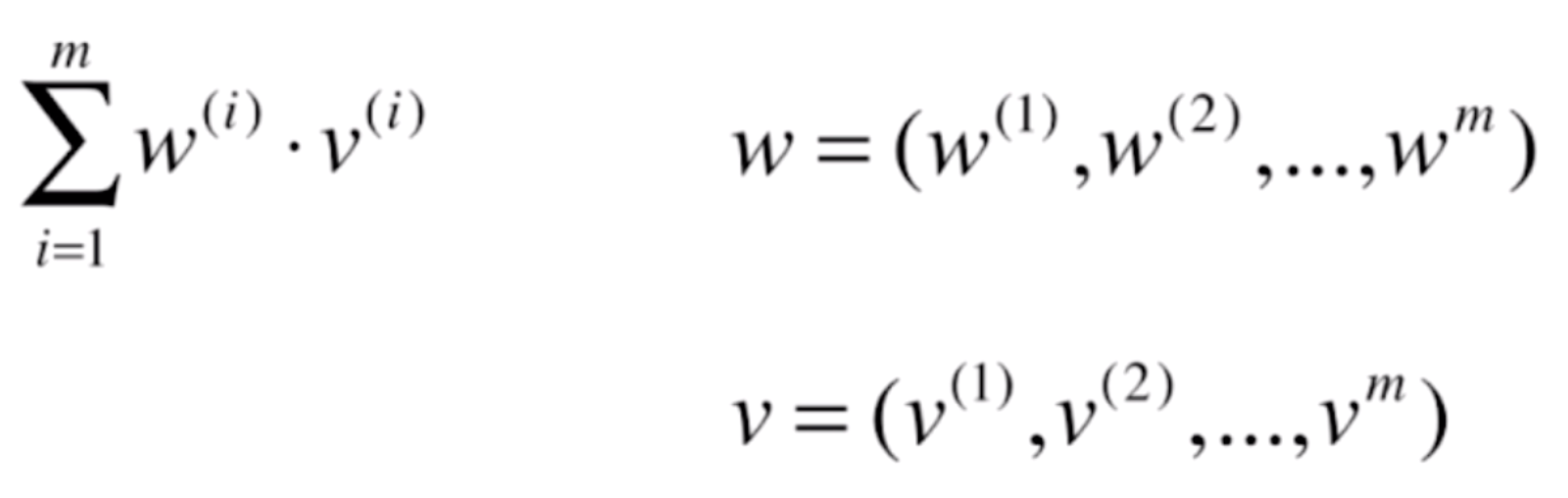


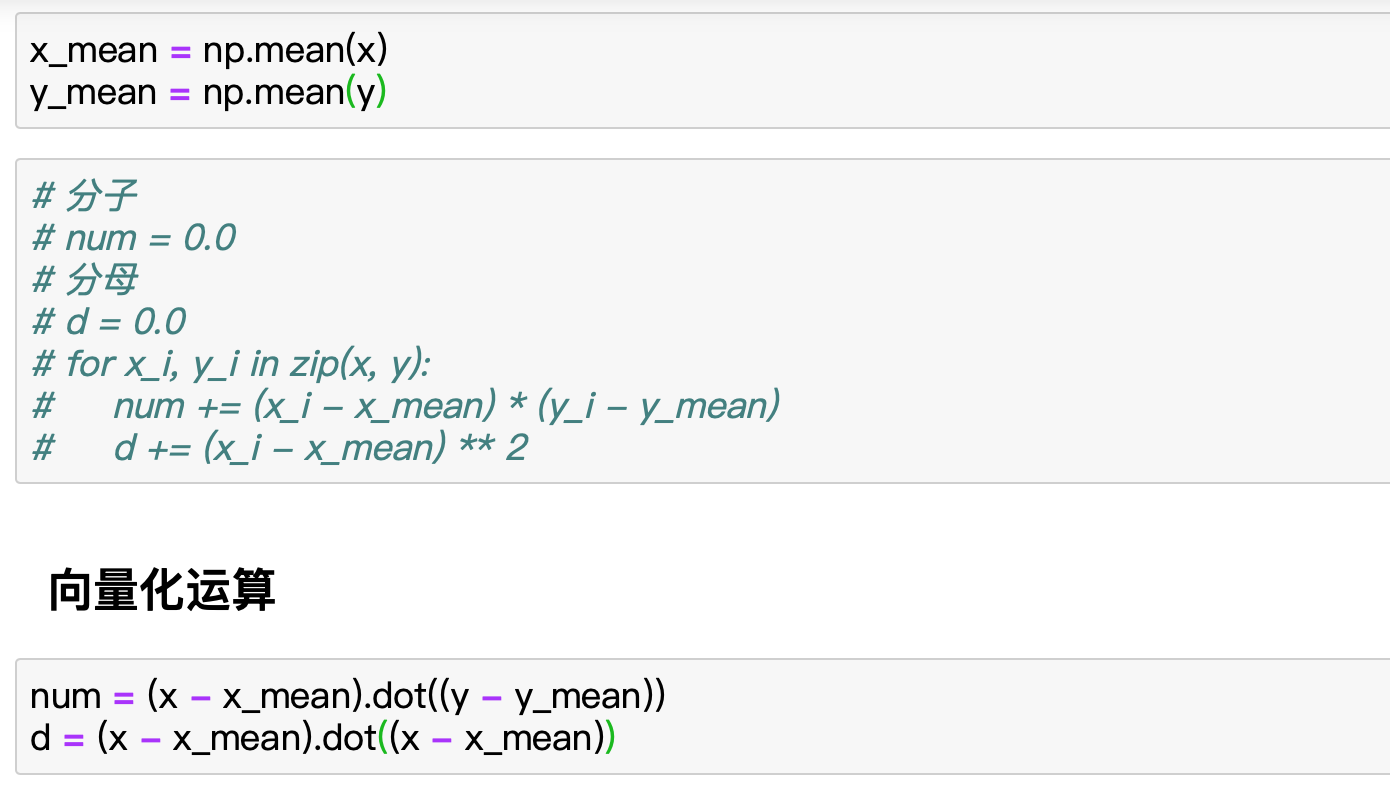




* 向量化运算

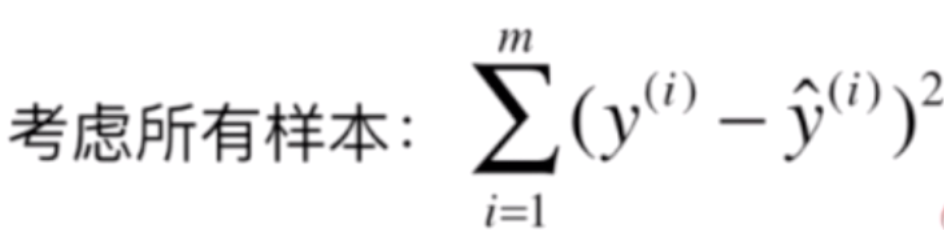
****





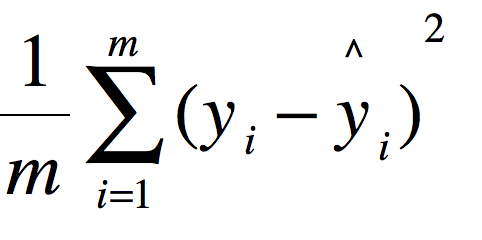
* 回归算法评测指标

**用目标函数评测回归算法的问题**

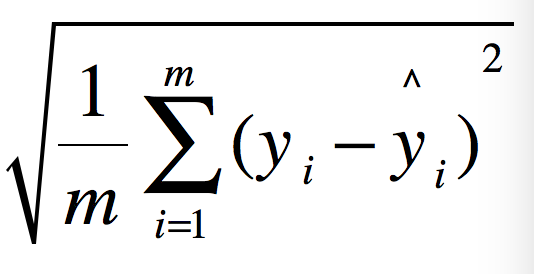
****

**样本数m是大还是小？**

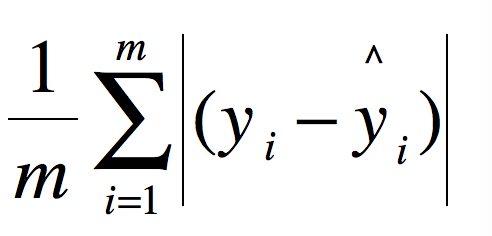
**MSE(Mean Squared Error)均方误差**



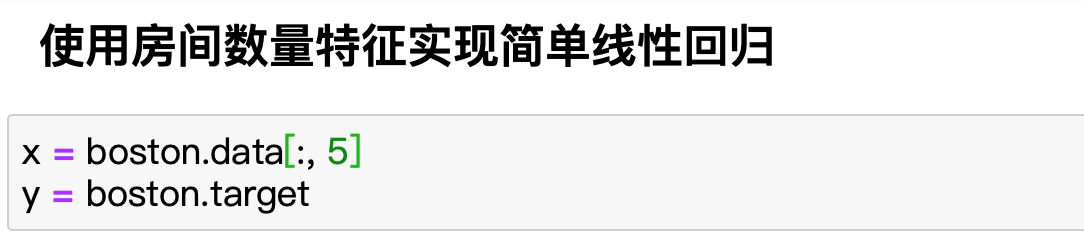
**RMSE(Root Mean Squared Error)均方根误差**

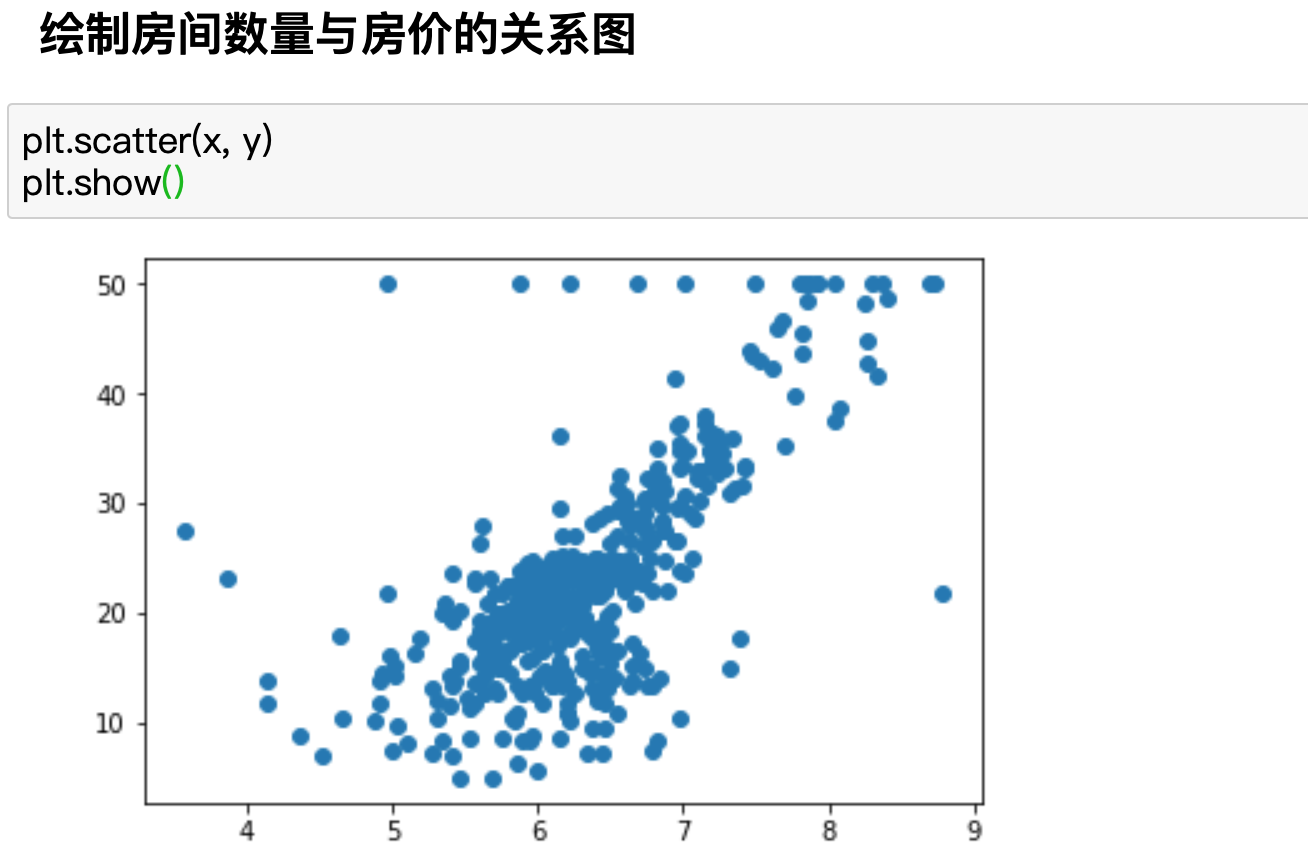


**MAE(Mean Absolute Error)平均绝对误差**

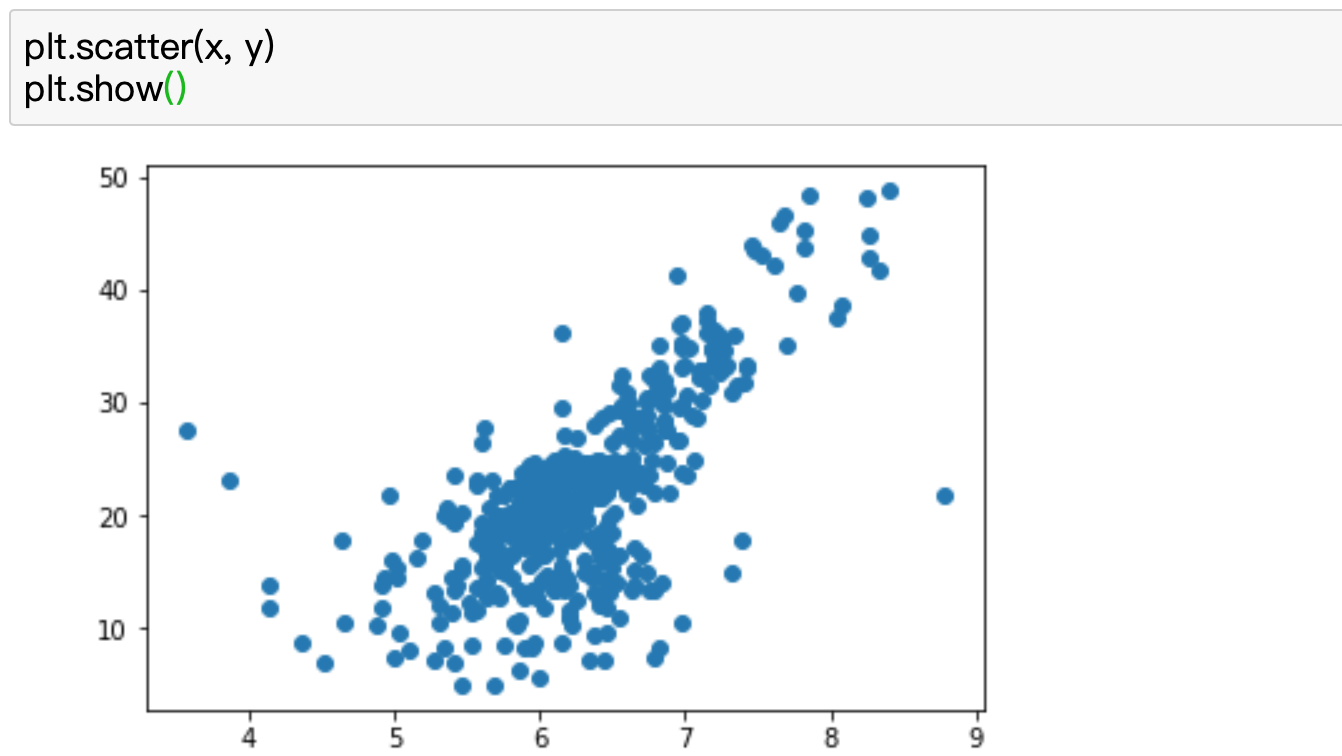


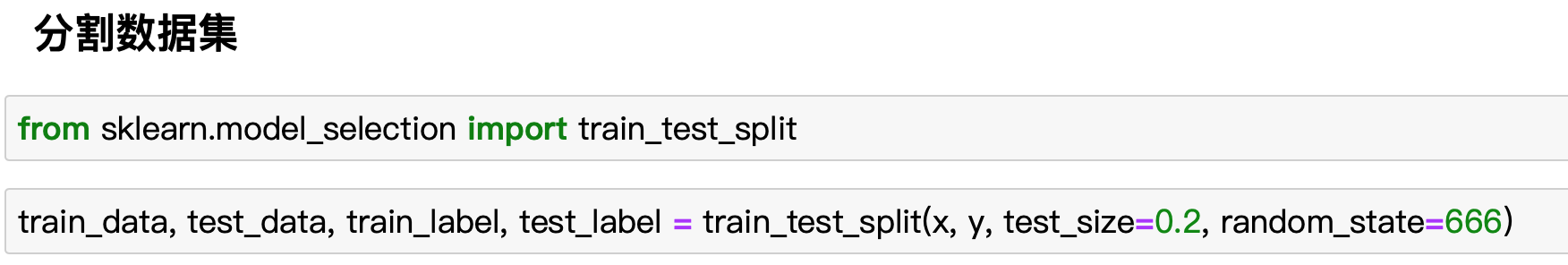
****

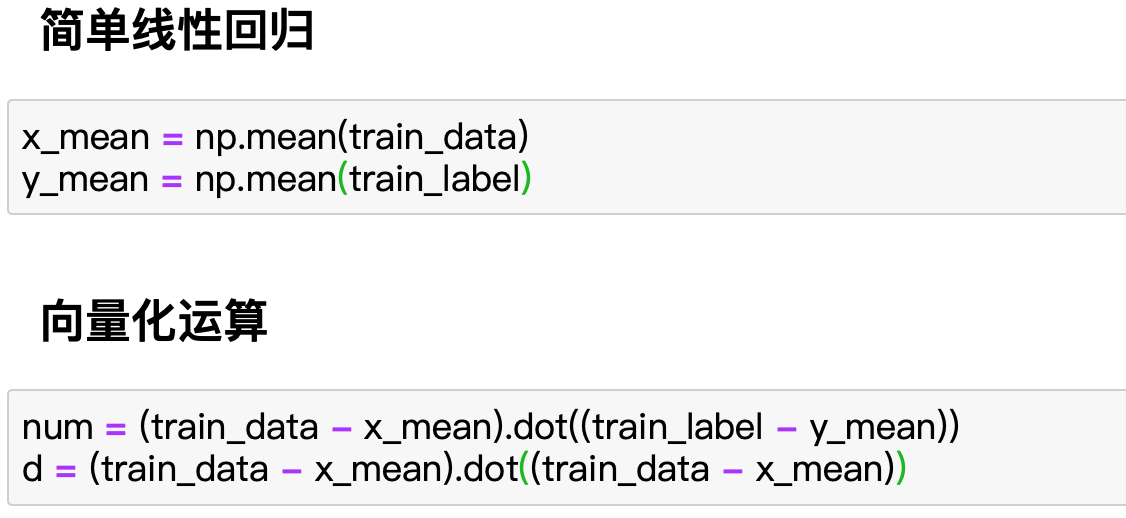
****

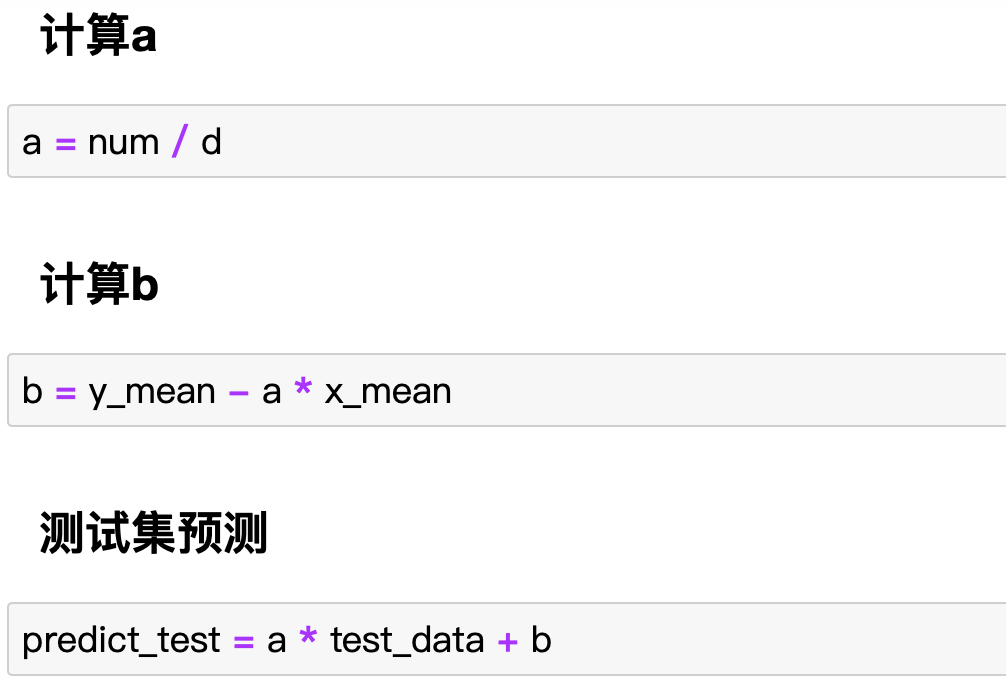
****

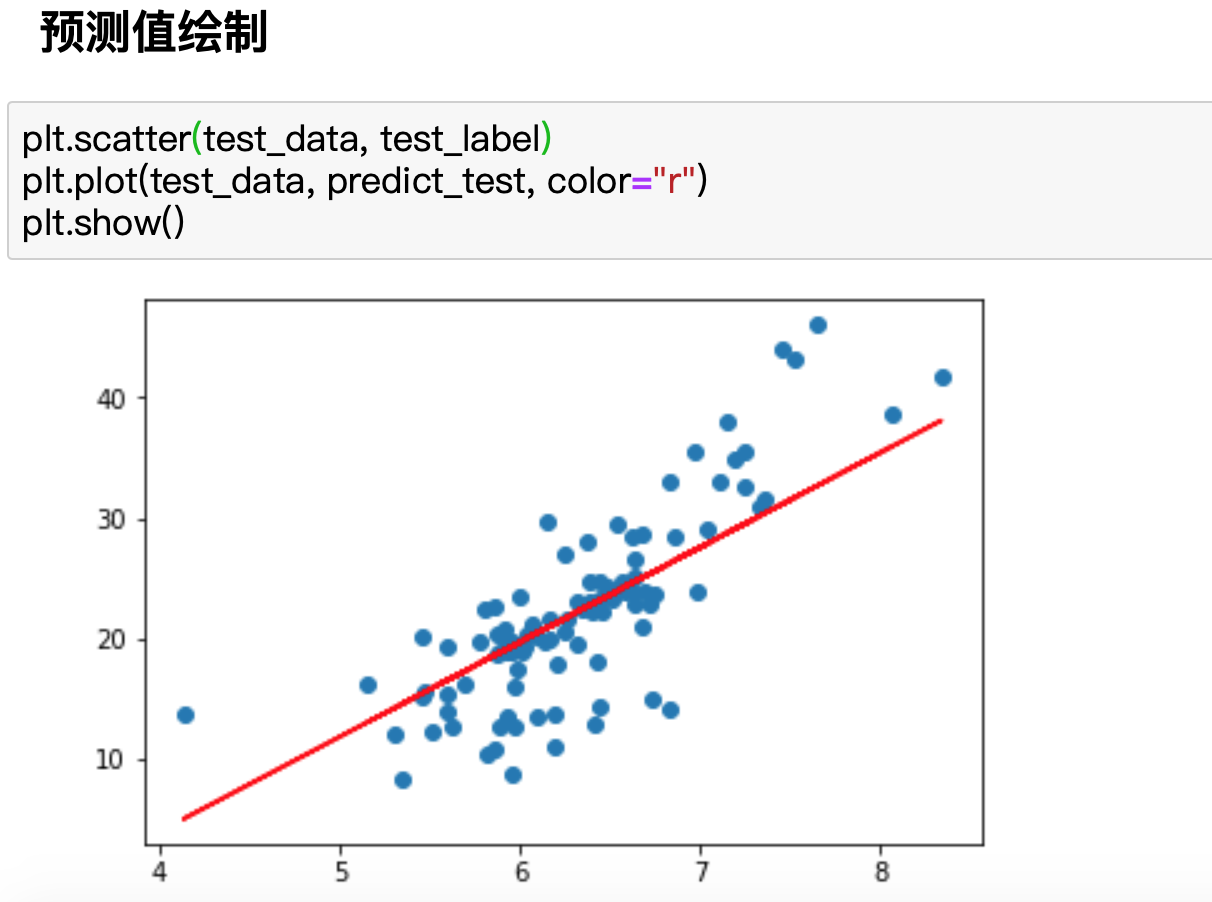
****

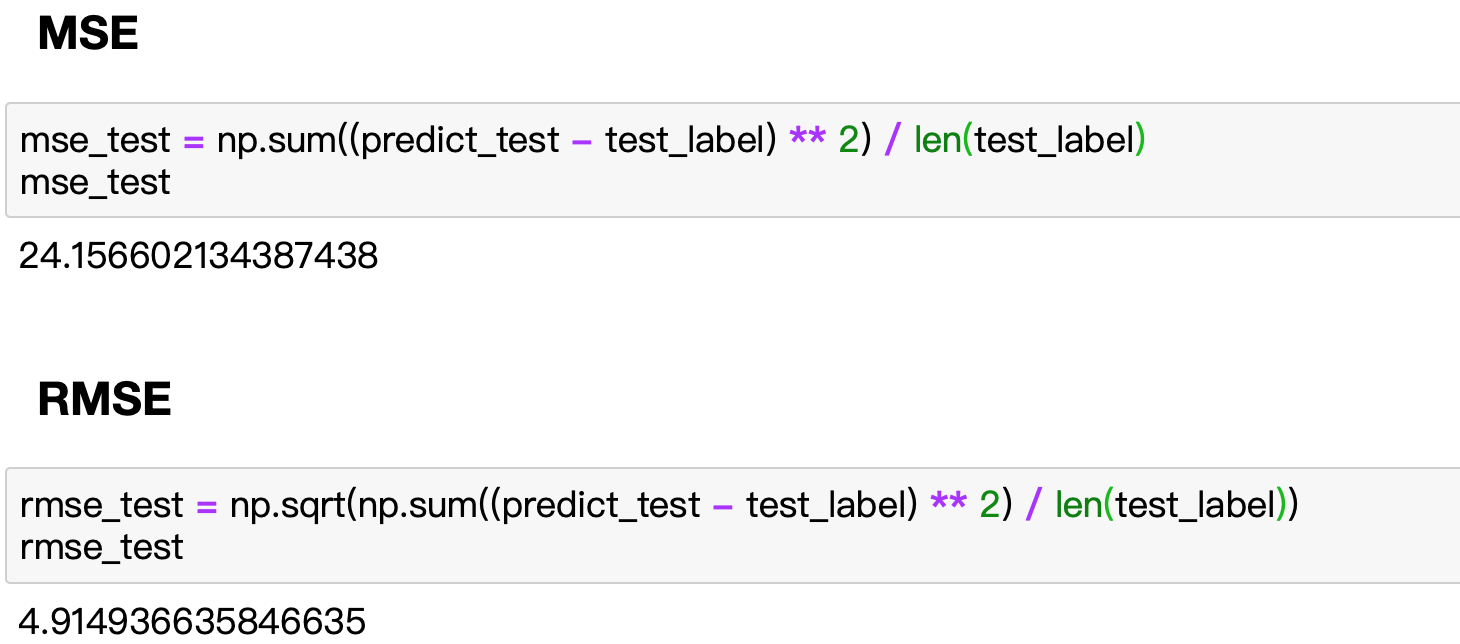
****

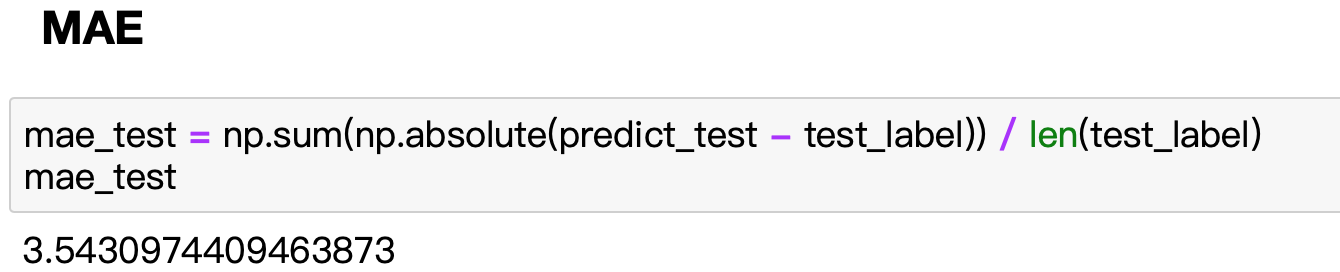
****

****

****

****

****

****

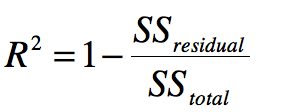
**RMSE vs MAE**

**RMSE相当于L2范数，MAE相当于L1范数。次数越高，计算结果就越与较大的值有关，而忽略较小的值，所以这就是为什么RMSE针对异常值更敏感的原因（即有一个预测值与真实值相差很大，那么RMSE就会很大）。**

**局限性**

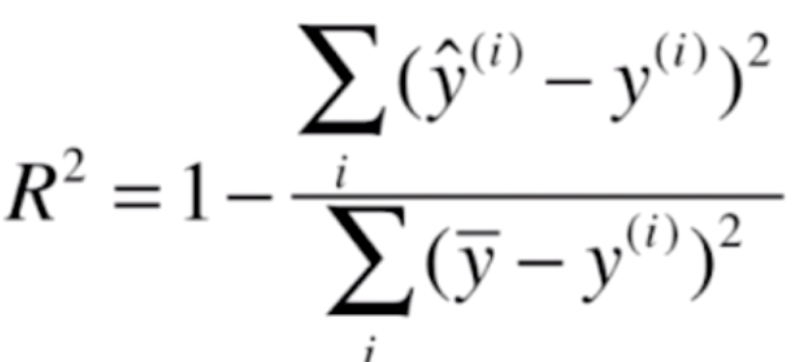
**不论是RMSE还是MAE，都无法评判相同的算法，在不同业务上的好坏。**

**R Squared(R方)**



**residual=Residual Sum of Squares(残差平方和)**

**total=Total Sum of Squares(总平方和)**

****

**分子:预测值与实际值差的平方和(模型预测产生的误差)**

**分母:实际值的平均值与每个实际值差的平方和(使用实际值的平均值预测产生的误差)**

**最终的结果相当于衡量了训练出的模型拟合住数据的比例。**

**R^2结果说明**

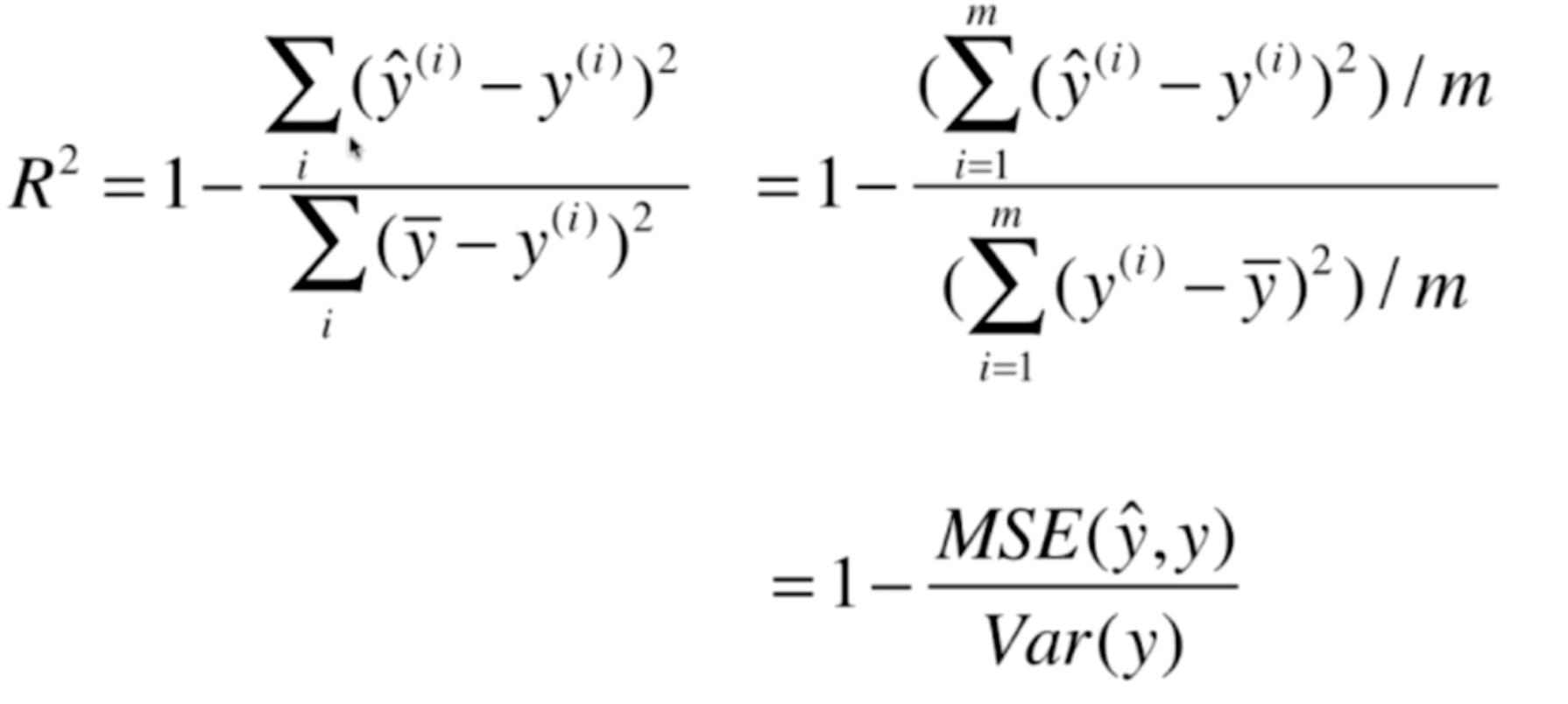
**1.R^2<=1**

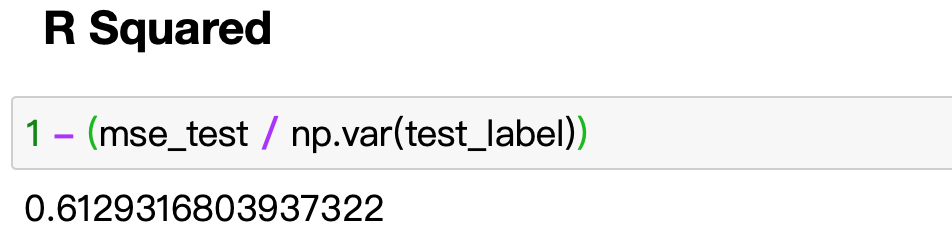
**2.R^2越大越好。当模型不犯任何错误时，R^2得到的最大值为1**

**3.当模型等于基准模型时，R^2的值为0**

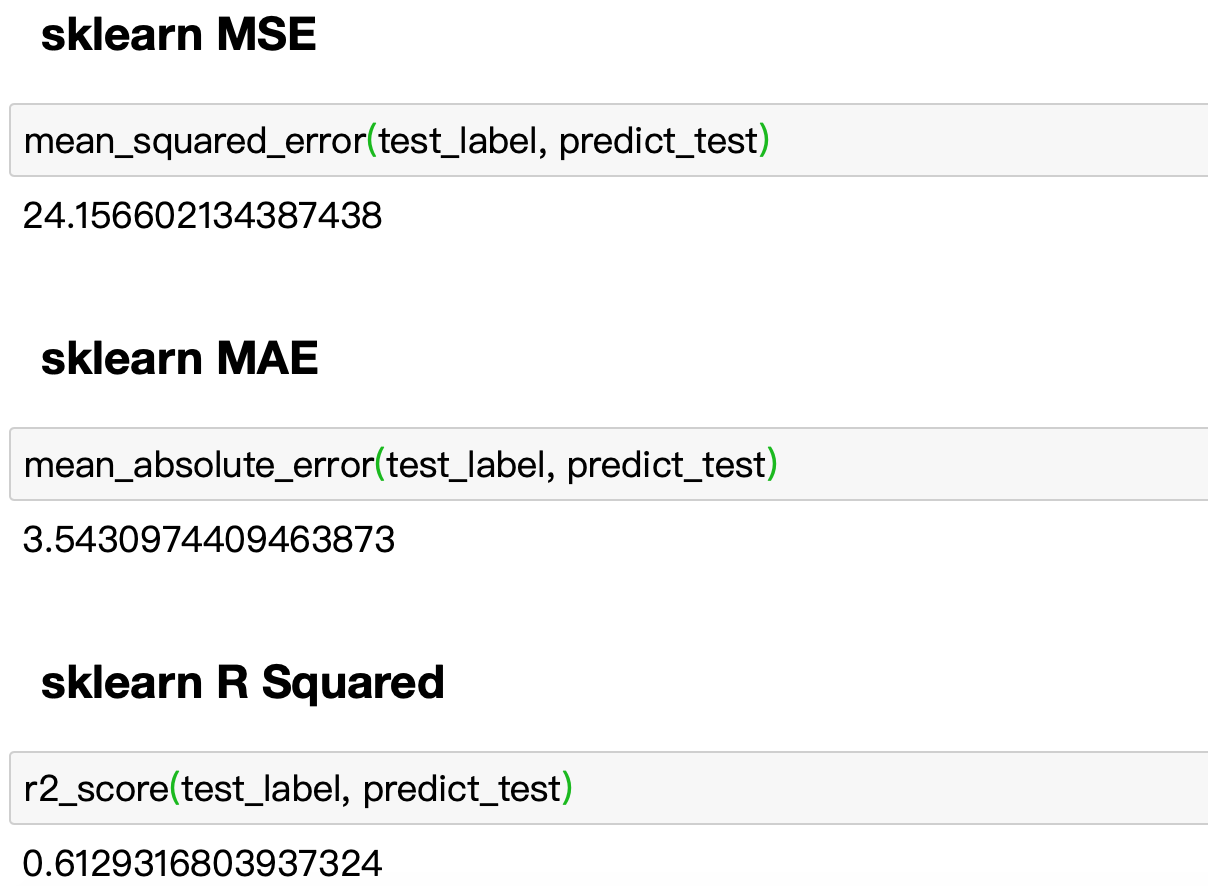
**4.如果R^2<0,说明训练出的模型还不如基础模型。此时，很有可能我们的数据不存在 任何线性关系**

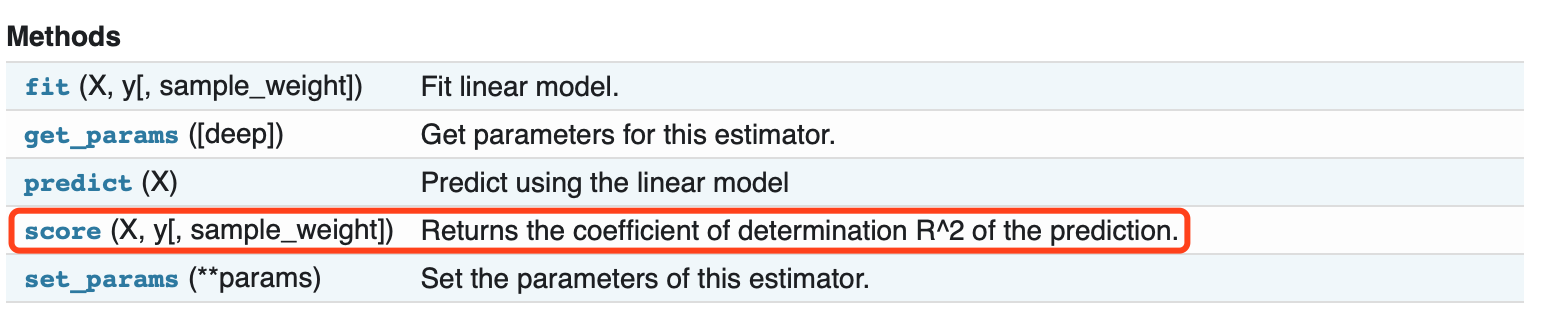
**算法变换**

****

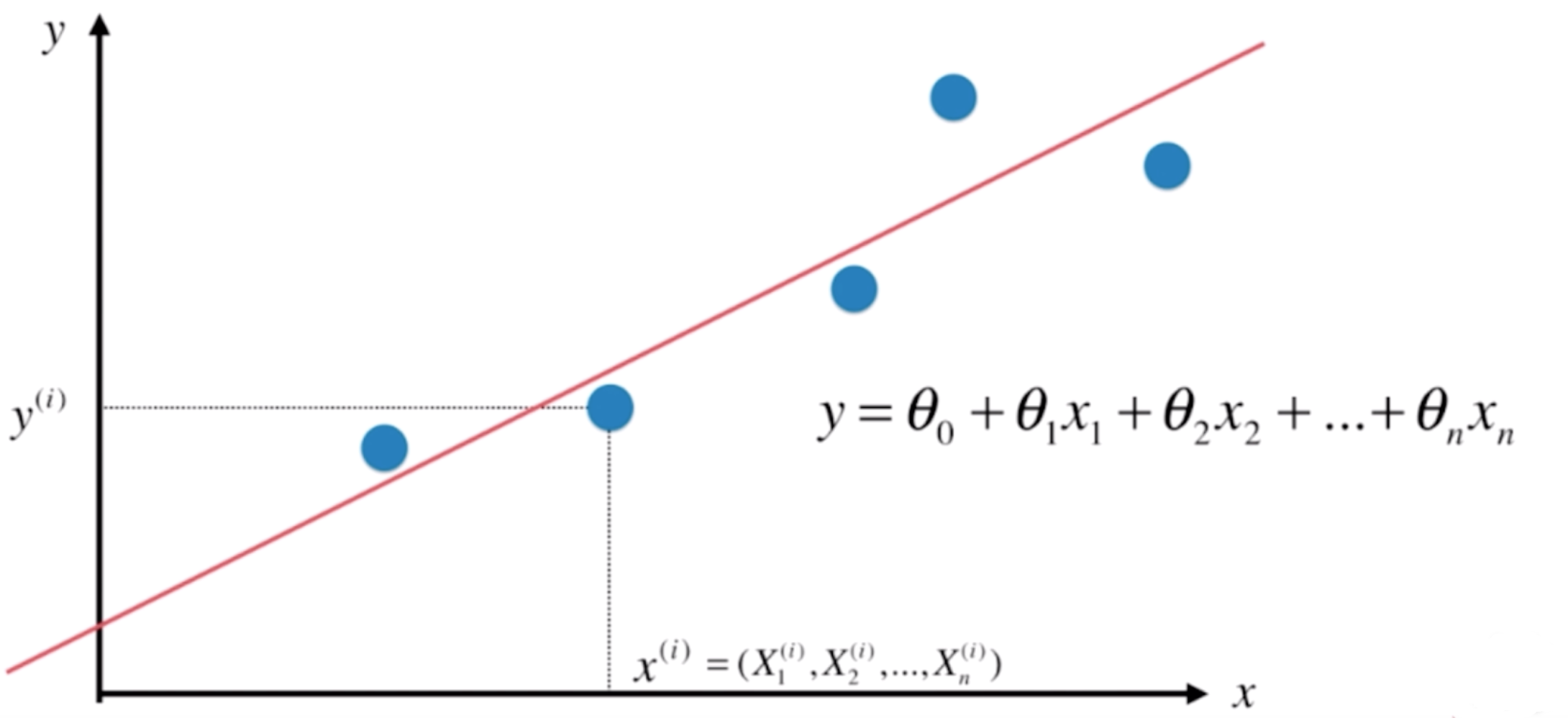
****

****

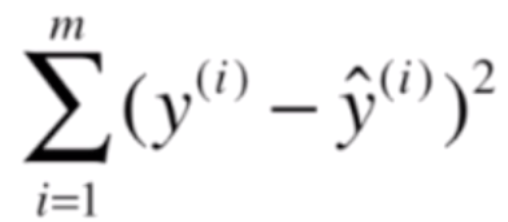
****

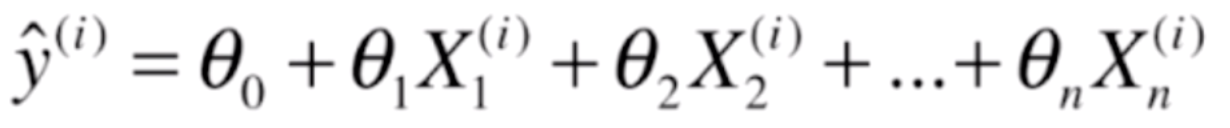
****

* 多元线性回归

****

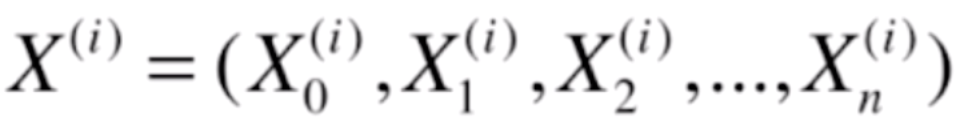
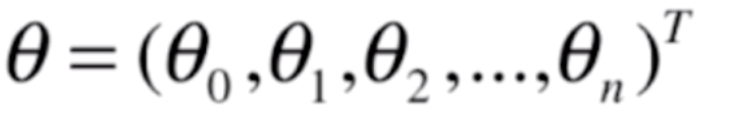
**目标函数(loss function)**

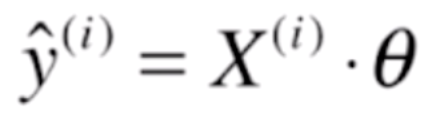
**尽可能小**

****

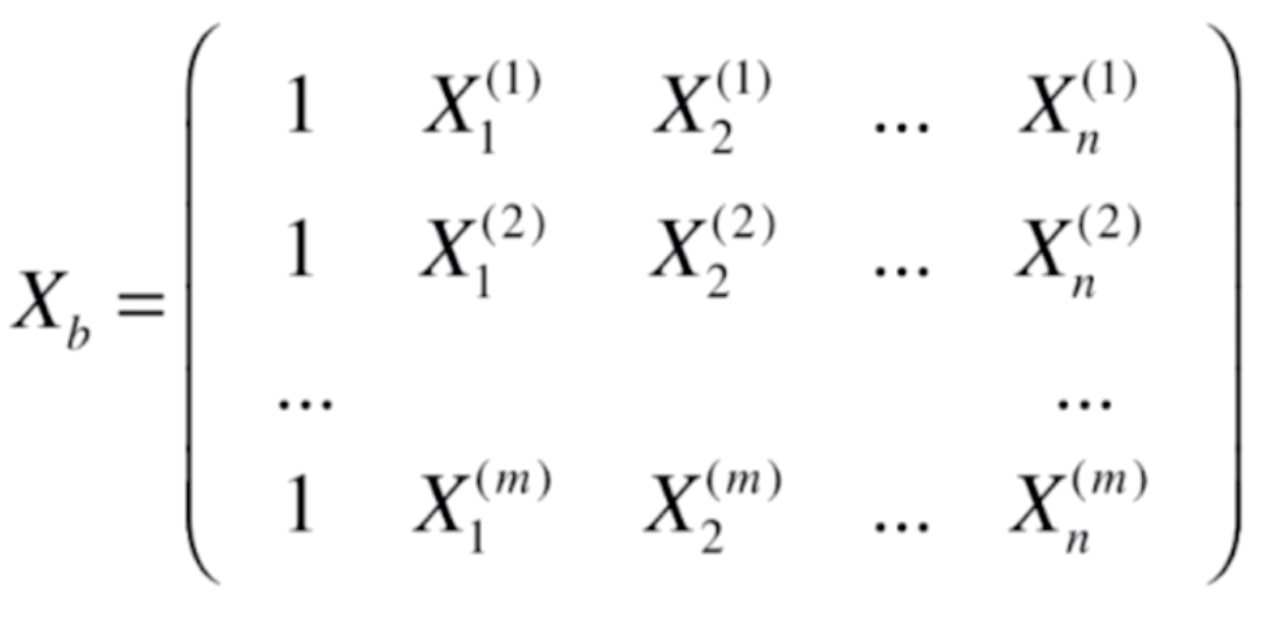
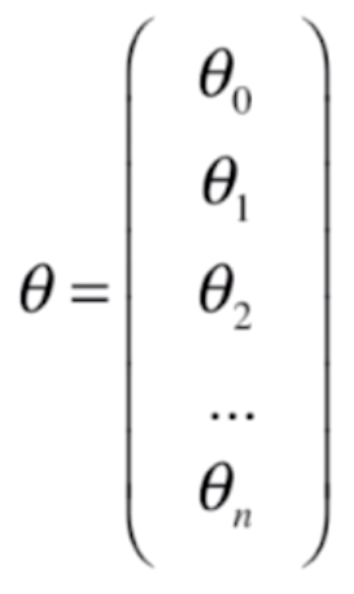
****

**向量化计算**

****

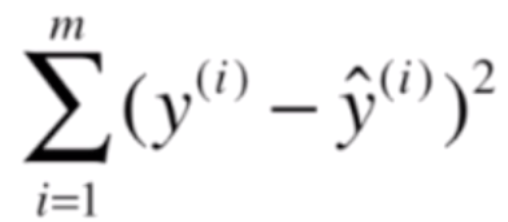
****

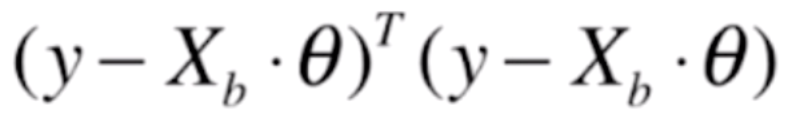
**矩阵化计算**

**** ****

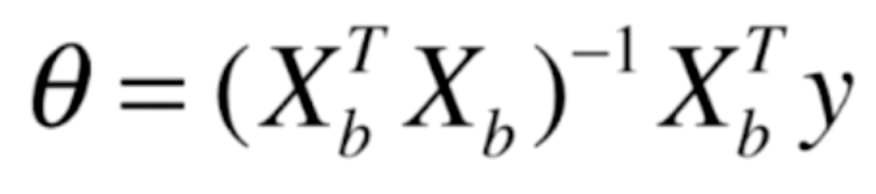


**目标函数向量化**

**尽可能小**

**尽可能小**

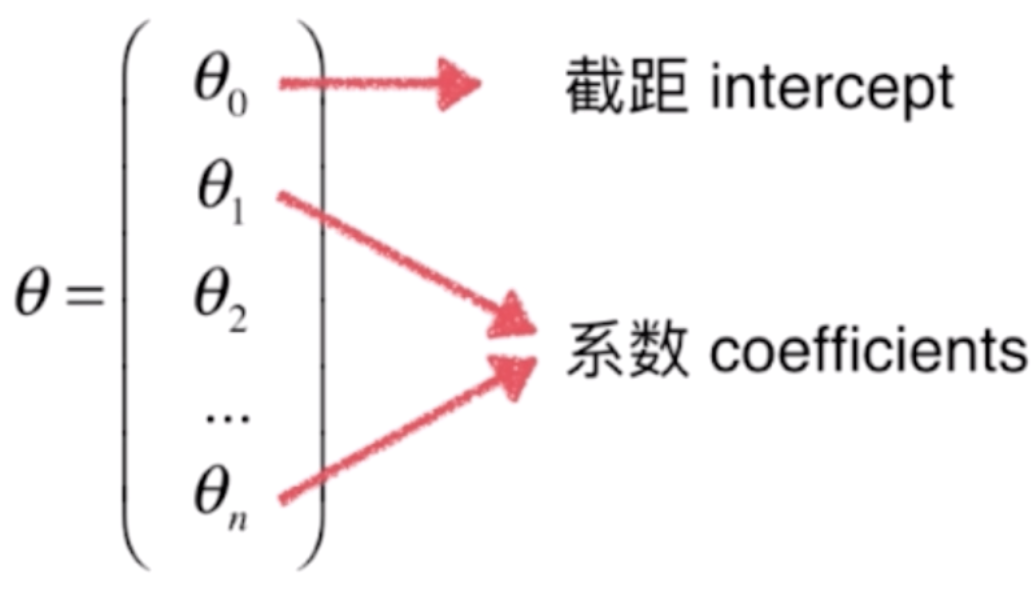
**正规方程解(Normal Equation)**

****

**优点：不需要对数据进行归一化处理**

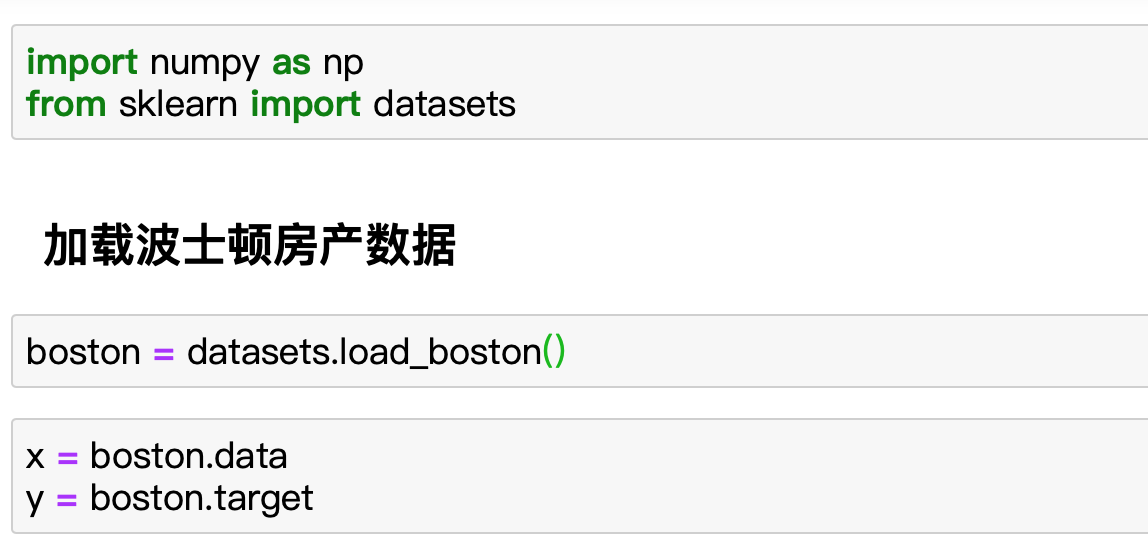
**缺点：时间复杂度高**

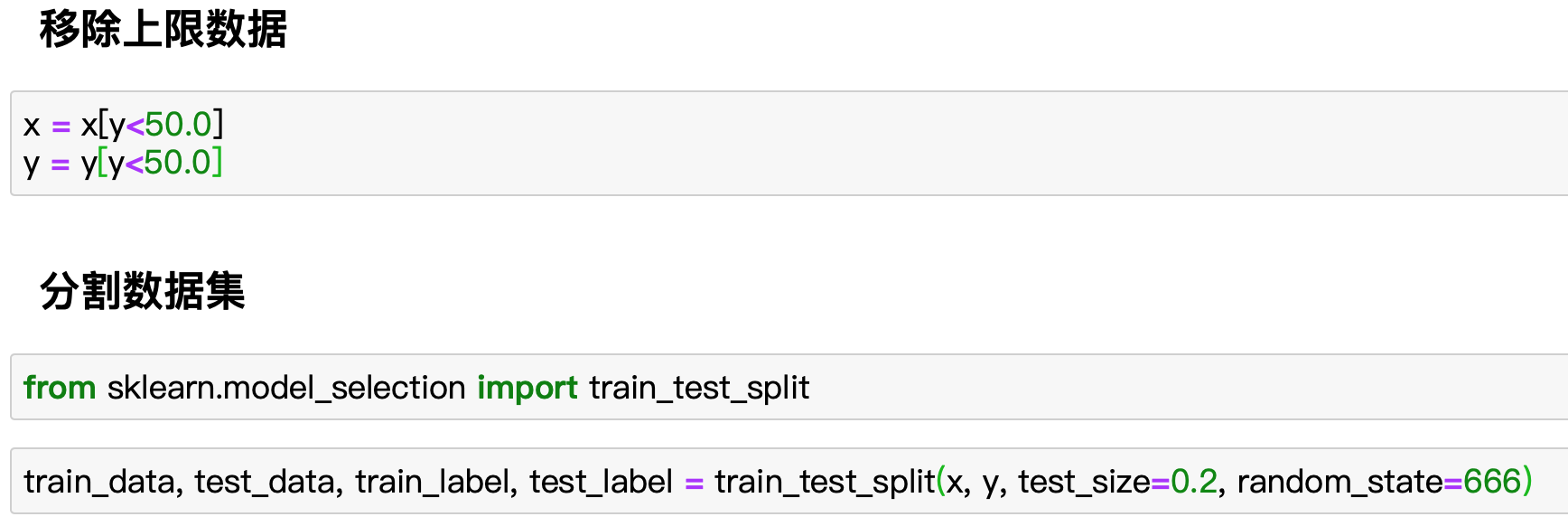
* 多元线性回归实现

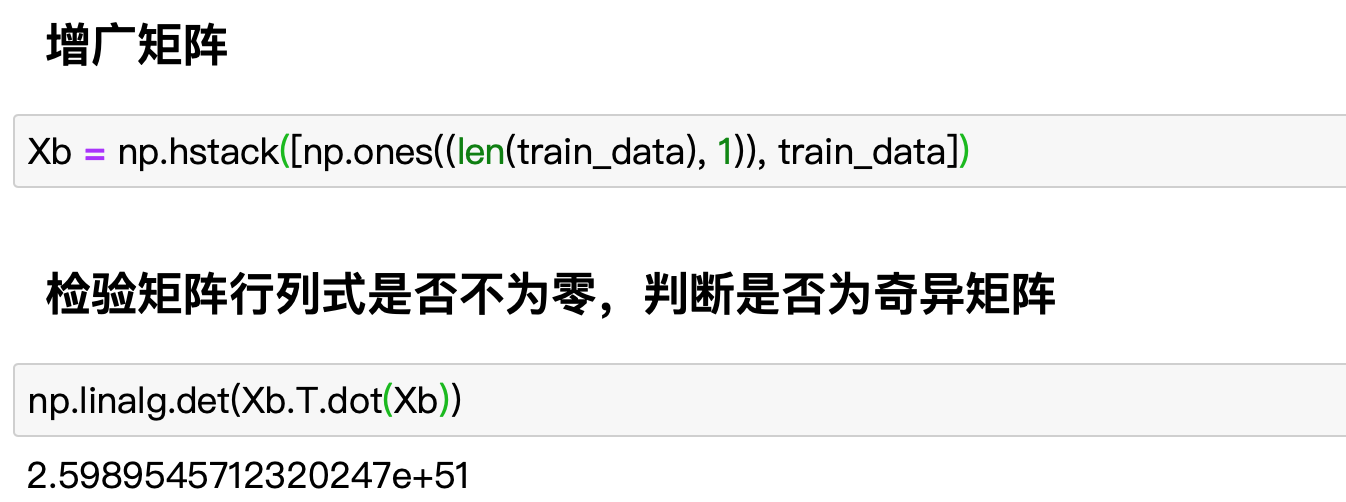
****

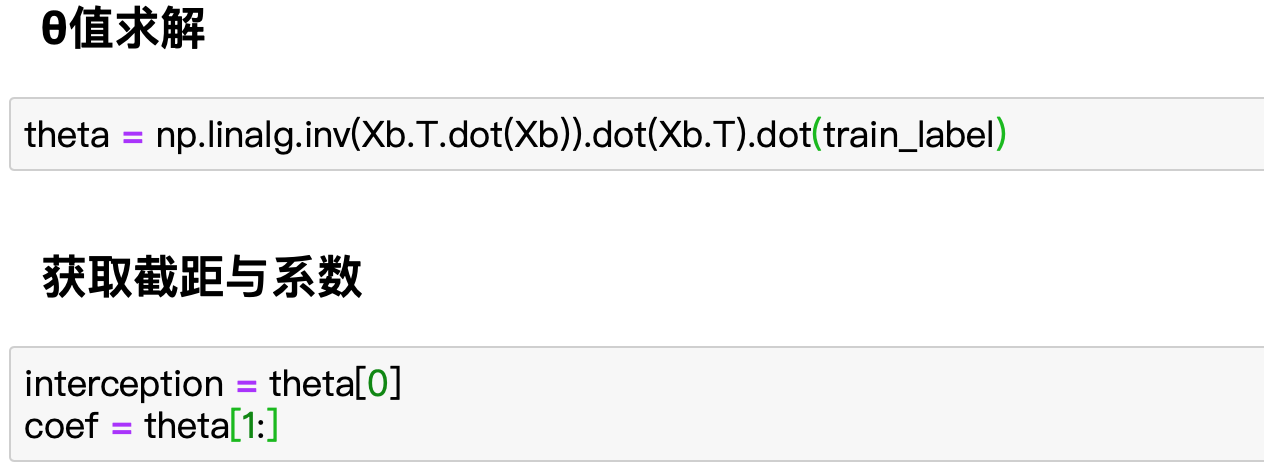
**系数：每一个θ值对应一个样本值，通常用于表示不同特征的权重，通过对于最终的预测结果有很好的说明。**

**截距：与最终的预测结果没有关系，只是代表模型预测的偏移量。**

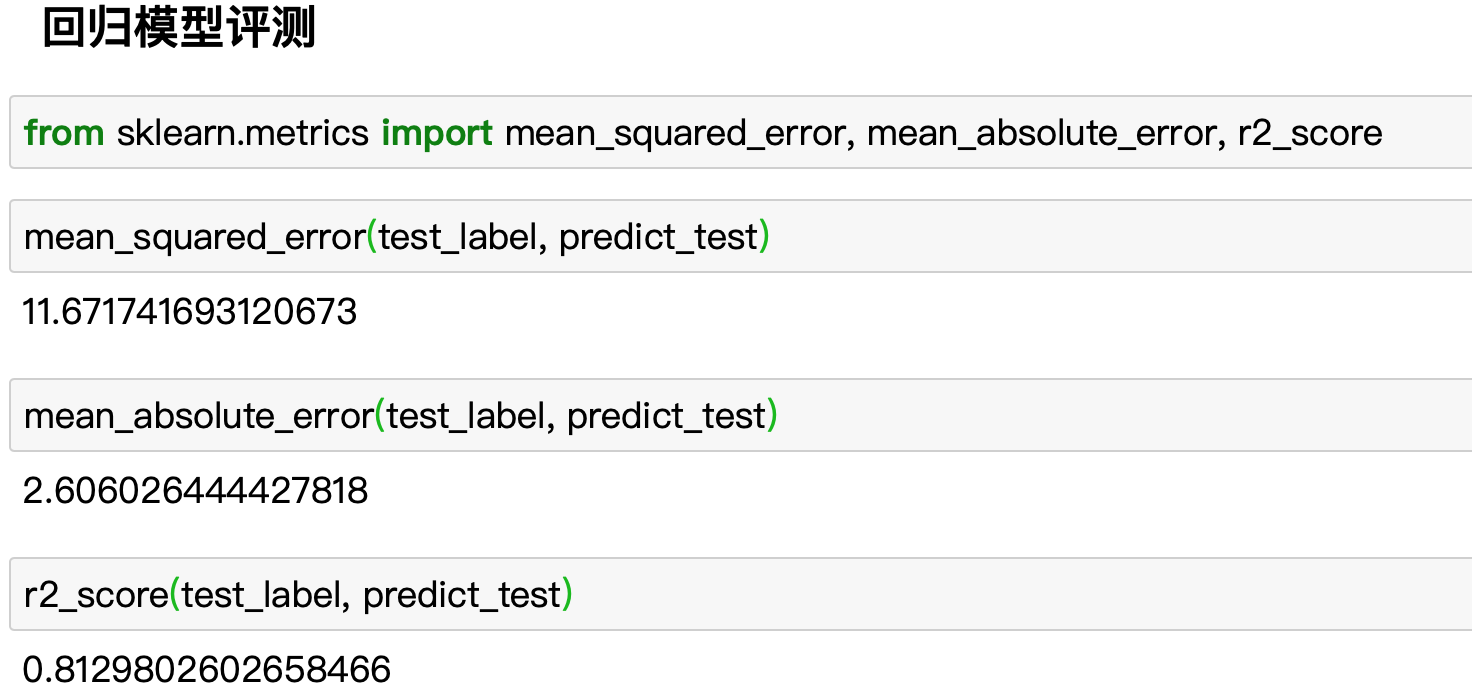
****

****

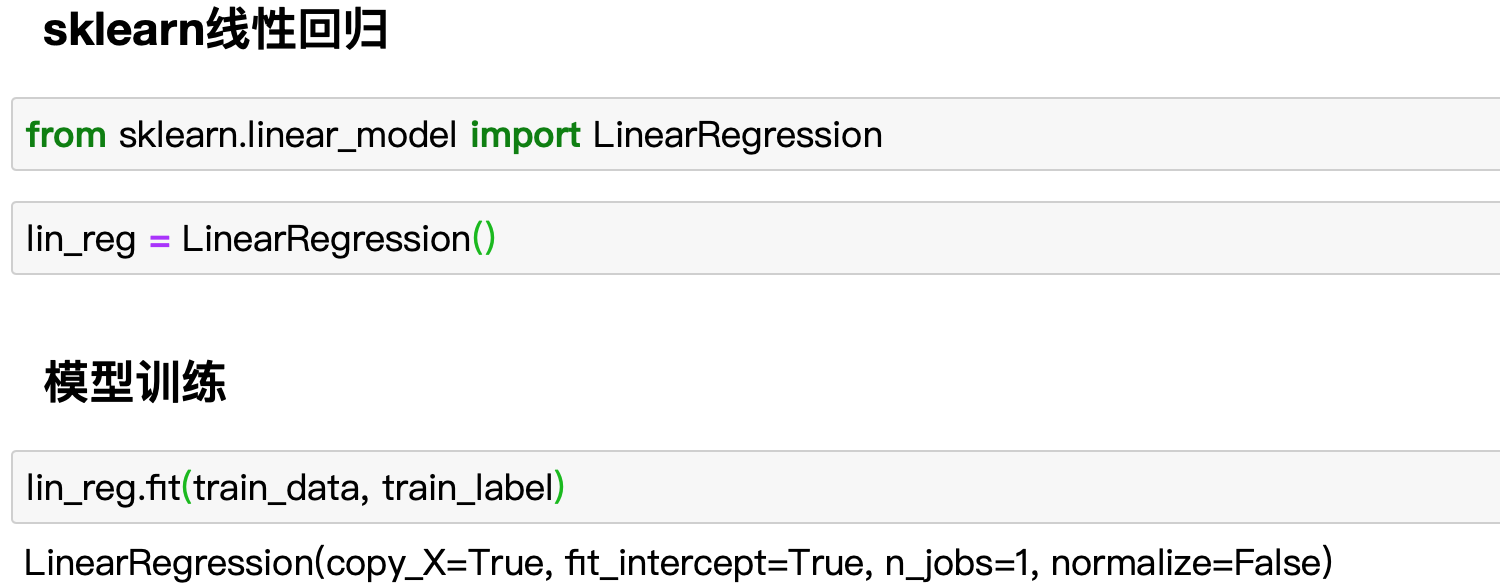
****

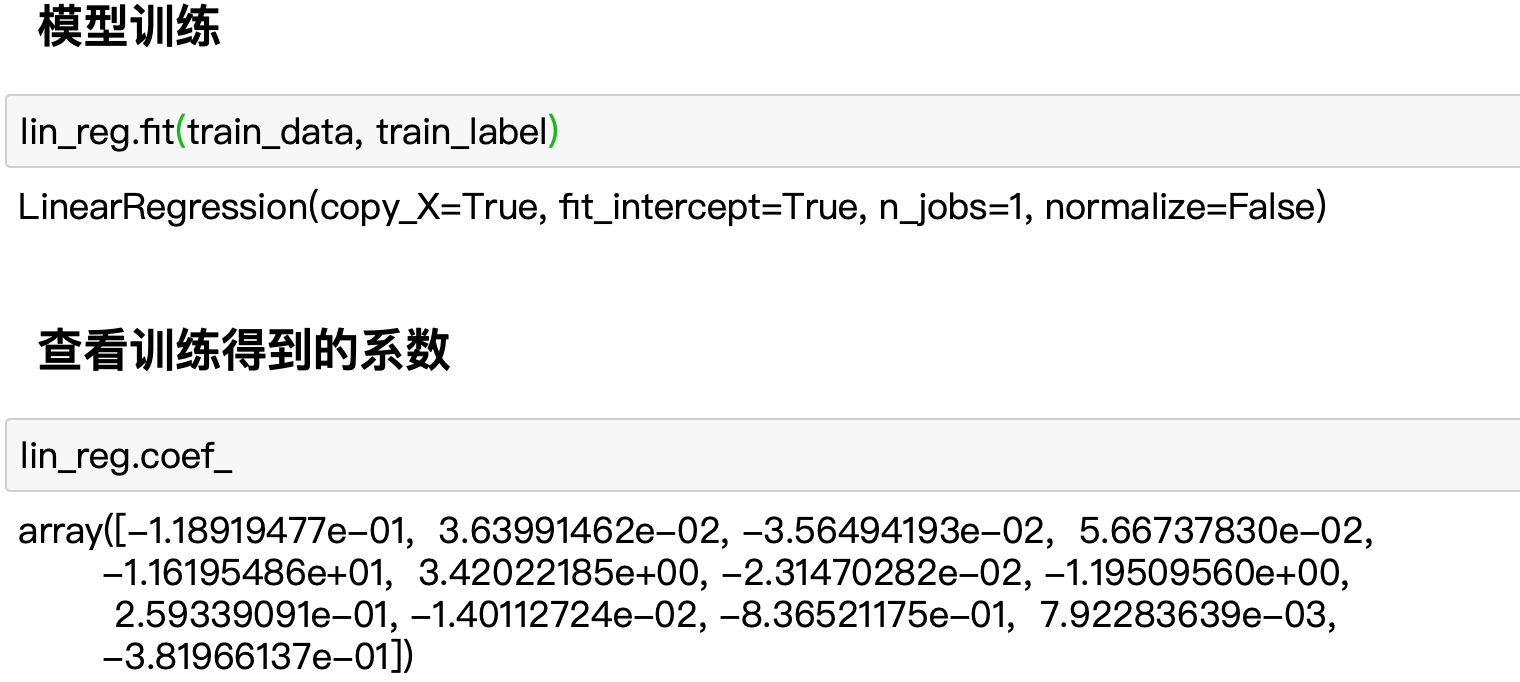
****

****

****

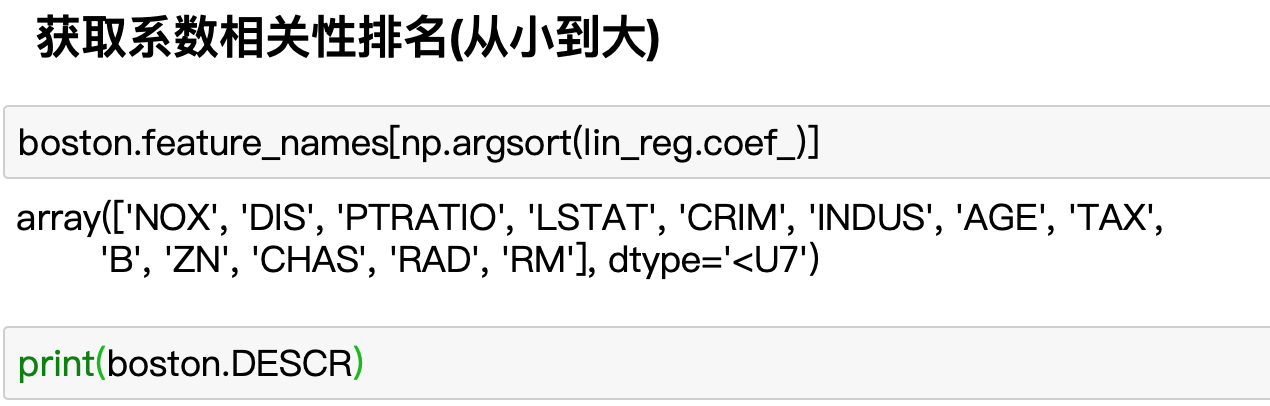
* sklearn线性回归实现

****

****

****

* 数据解释说明

****