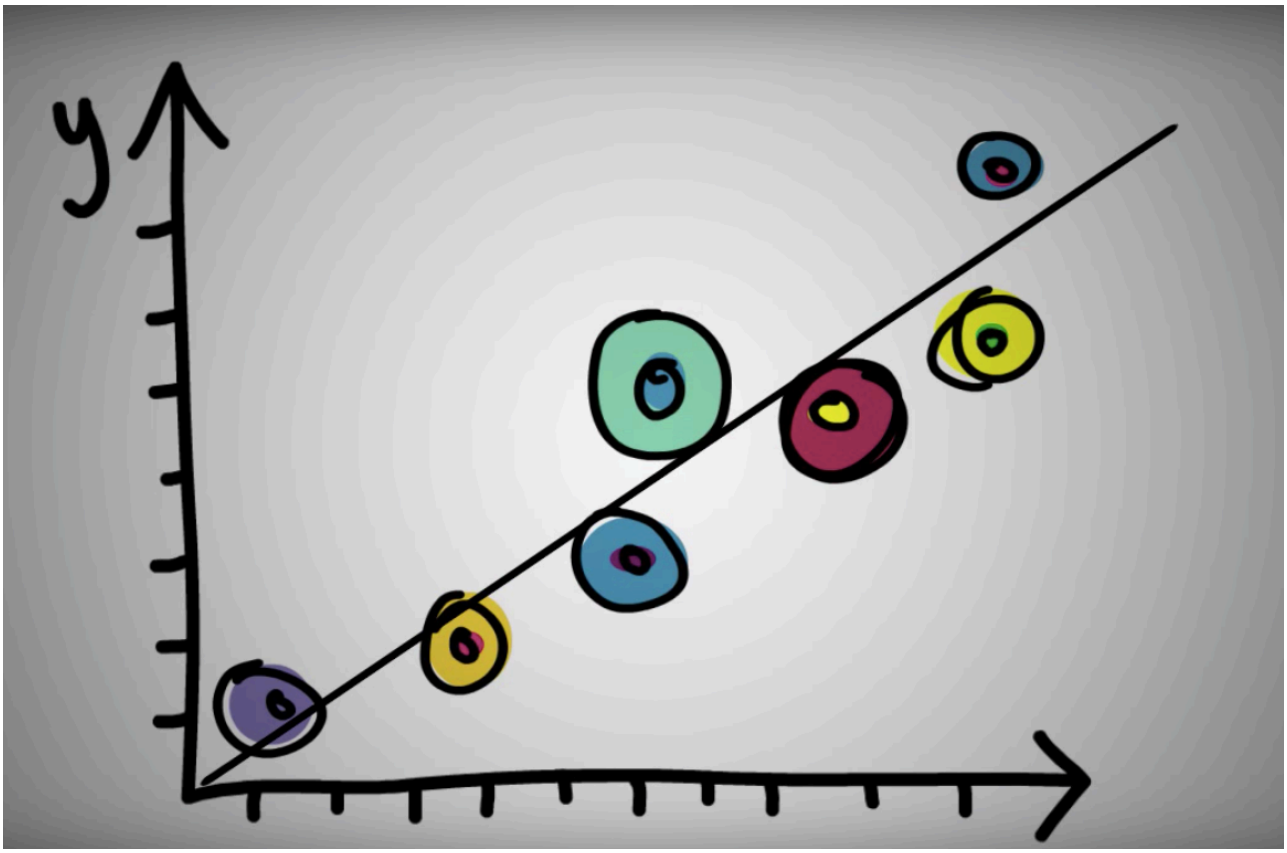
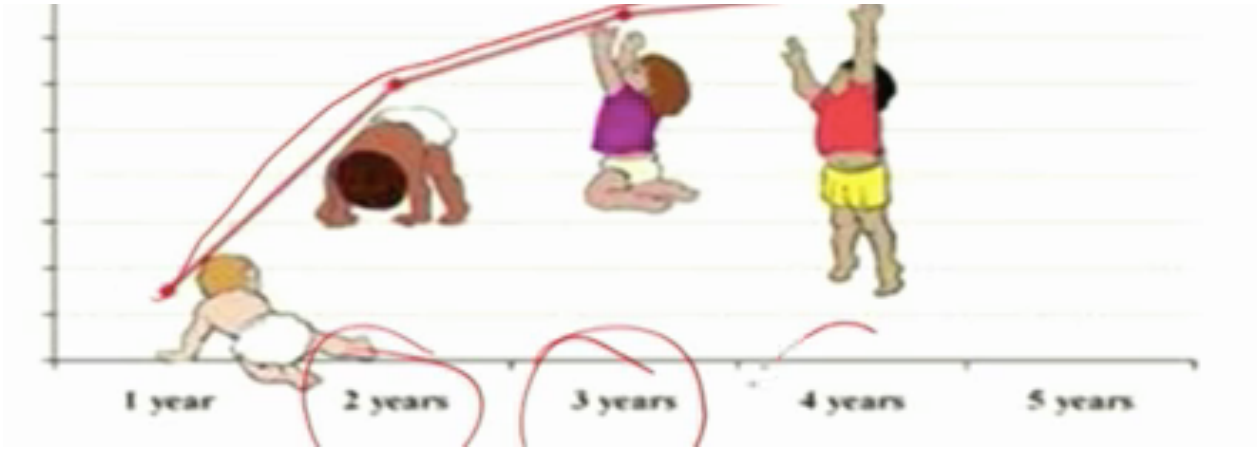
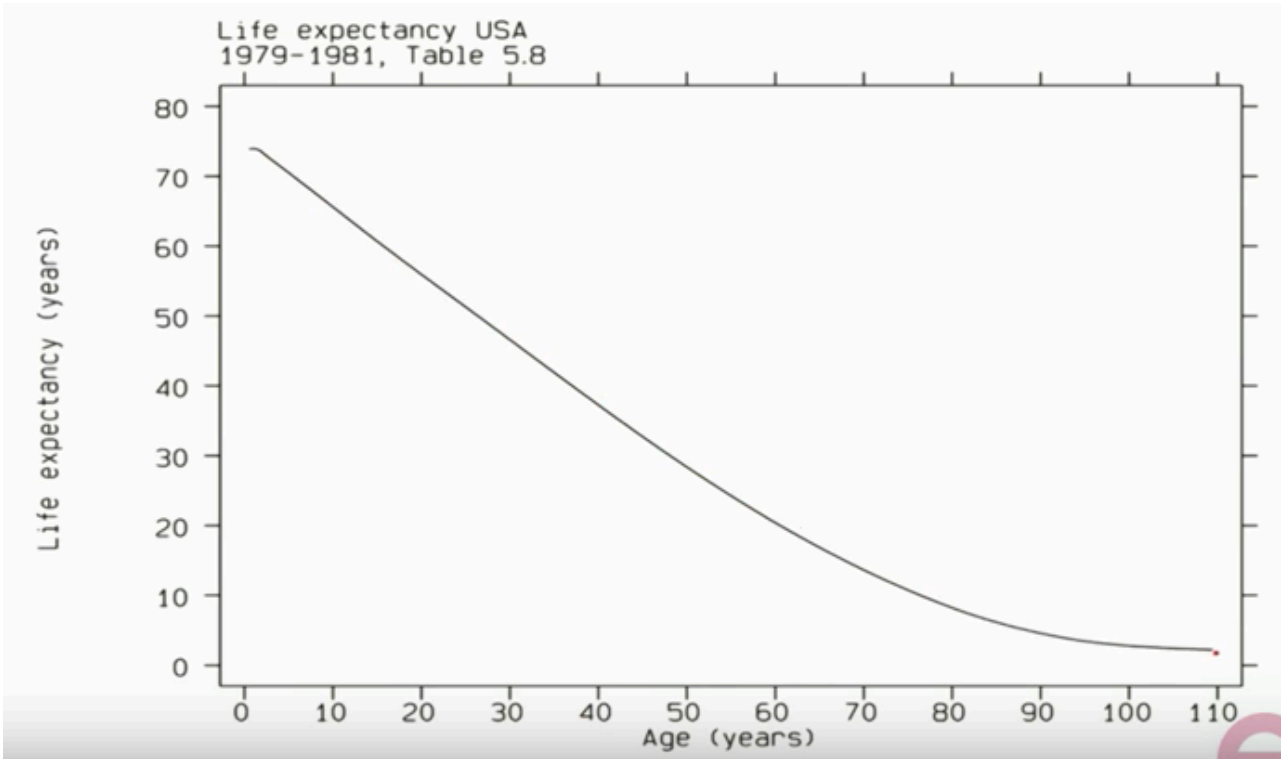
线性回归

Wednesday, January 31, 2018 4:00 PM



https://onlinecourses.science.psu.edu/stat501/node/255
https://en.wikipedia.org/wiki/Linear_regression





1. 简单的线性回归(simple linear regression)
   a. 很多决定是根据两个或者多个变量之间的关系而作出
   b. 回归分析(regression analysis)用来建立方程模拟两个或者多个变量之间的关系

      广告投放量与销售额之间的关系；

c. 被预测的是因变量(dependent variable) , Y;被用来进行预测的是自变量(independent variable), X
d. 用一条直线来表示两个变量间的关系
e. 如果自变量多于一个，则是多元线性回归(multiple regression)

2. 简单线性回归模型
   a. 用来描述Y与 X 以及偏差 error
   b. 回归模型：

$$Y = \beta_0 + \beta_1 X + \epsilon$$

参数，偏差

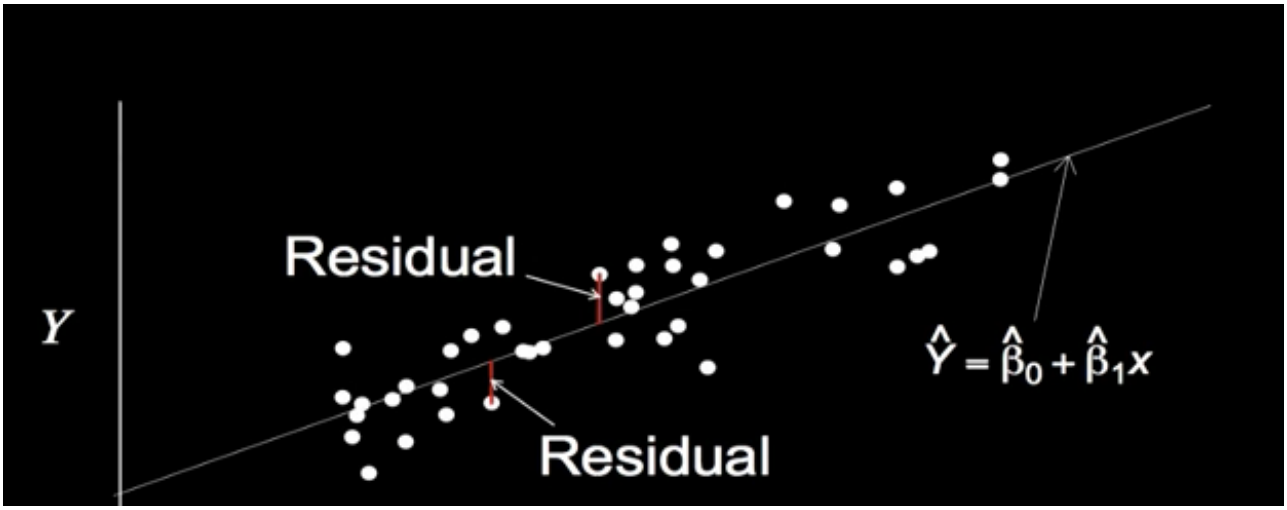
   c. 该模型对应一条直线
   d. 截距，斜率

3. 线性回归分析：用样本数据来估计参数，根据得来的参数和自变量(X), 来预测未知的因变量 (Y).

4. $$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

关于偏差的假设：随机变量，均值为0, 独立的，满足正态分布

5. 估计参数：找到最合适的那一条直线对应的参数



$$X$$

如何找到呢？

$$\sum |e_i| = \sum |Y_i - \hat{Y}_i|$$

$$\begin{aligned}\sum e_i^2 &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2\end{aligned}$$

最小化偏差平方和，推导过程涉及极值求导，这里不讲。

最终的公式是

最终的公式是：

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

$$R^2 = 1 - \frac{SSE}{SST}$$

$$SSE = \sum(Y_i - \hat{Y})$$

$$SST = \sum(Y_i - \bar{Y})$$

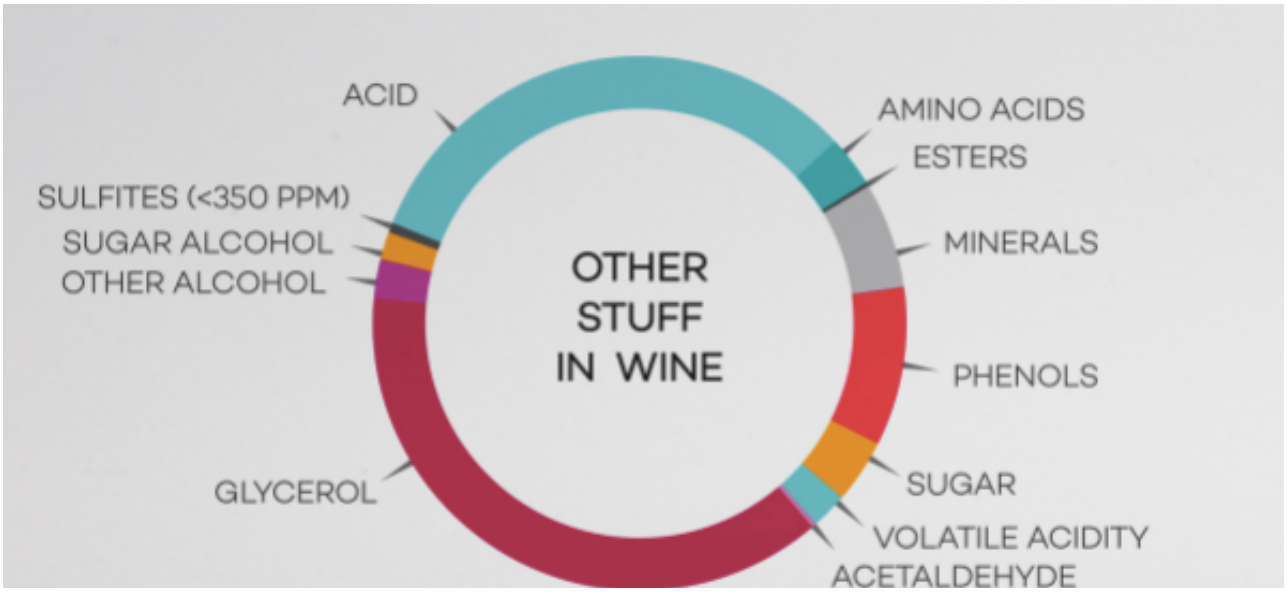
R平方是来测量X对Y的解释力的，值越大（0-1），说明X对Y的解释力越大（拟合度越高）。

SSE是回归方程的方差（未被解释的离差）

SST是Y的总方差。

用1减去不能解释的部分，那么剩下的就是解释的部分，也就是说自变量解释了因变量变动的百分比的多少，那么r方的值肯定是越大越好，意味着该模型把y的变动解释得好，R方的范围显然是0到1，在预测实践中，人们往往采纳R方最高的模型。

葡萄酒品质预测的故事：


ACID
AMINO ACIDS
ESTERS
SULFITES (<350 PPM)
SUGAR ALCOHOL
OTHER ALCOHOL
MINERALS
OTHER
STUFF
IN WINE
PHENOLS
GLYCEROL
SUGAR
VOLATILE ACIDITY
ACETALDEHYDE