广义最小二乘法GLS

2022年5月7日 11:16

- 方差偏差误差残差：
- 定义
  - 泛化误差可分解成偏差、方差、残差之和
  - **模型偏差bias：**偏差是指预测结果与真实值（观测的结果）之间的差异，排除噪声的影响，偏差更多的是针对某个模型输出的样本误差，偏差是模型无法准确表达数据关系导致，比如模型过于简单，非线性的数据关系采用线性模型建模，偏差较大的模型是错的模型。
    - 偏差又称为表观误差，是指个别测定值与测定的平均值之差，它可以用来衡量测定结果的精密度高低。在统计学中常用来判定测量值是否为坏值。精密度是指一样品多次平行测定结果之间的符合程度，用偏差表示。偏差越小，说明测定结果精密度越高。
  - **模型方差variance：**模型方差不是针对某一个模型输出样本进行判定，而是指多个(次)模型输出的结果之间的离散差异，反映的是模型每一次输出结果与模型输出期望（即均值） 之间的误差，即模型的稳定性。注意这里写的是多个模型或者多次模型，即不同模型或同一模型不同时间的输出结果方差较大，方差是由训练集的数据不够导致，一方面量 (数据量) 不够，有限的数据集过度训练导致模型复杂，另一方面质(样本质量)不行，测试集中的数据分布未在训练集中，导致每次抽样训练模型时，每次模型参数不同，输出的结果都无法准确的预测出正确结果；概率论中方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。统计中的方差（样本方差）是每个样本值与全体样本值的平均数之差的平方值的平均数。
    - **均方误差MSE：**mean squared error，作为机器学习中常常用于损失函数的方法，是通过计算每个预测值和实际值之间的差值的平方和再求平均，机器学习中它经常被用于表示预测值和实际值相差的程度。
    - **和方差SSE：**也就是误差平方和，the sum of squares due to error
    - **均方根误差RMSE：**ROOT mean square error是观测值与真值偏差的平方和与观测次数m比值的平方根。是用来衡量观测值同真值之间的偏差。
    - **平均绝对误差MAE：**Mean Absolute Error，是绝对误差的平均值，能更好地反映预测值误差的实际情况。
    - **标准差SD：**Standard Deviation，是方差的算数平方根，是用来衡量一组数自身的离散程度。标准差是表示个体间变异大小的指标,反映了整个样本对样本平均数的离散程度,是数据精密度的衡量指标。样本标准差公式如下：

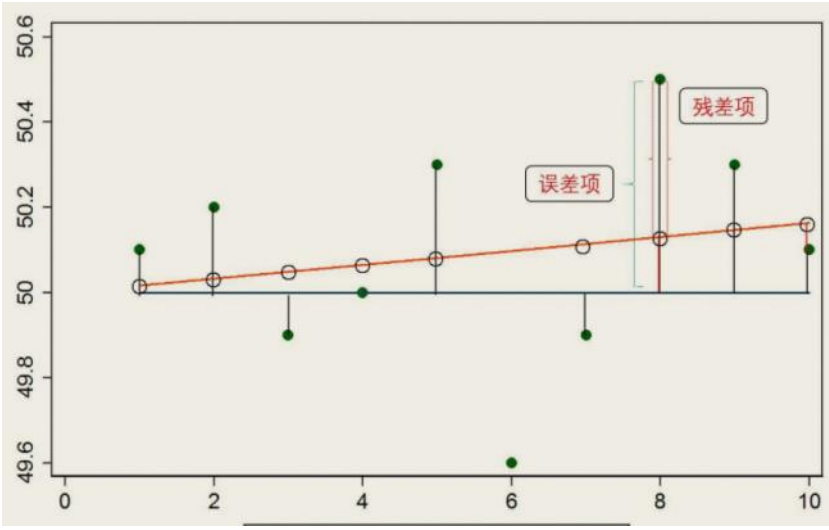
      $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

    - 标准误SE：standard error，标准误反映样本平均数对总体平均数的变异程度,从而反映抽样误差的大小 ,是量度结果精密度的指标。样本标准误如下：

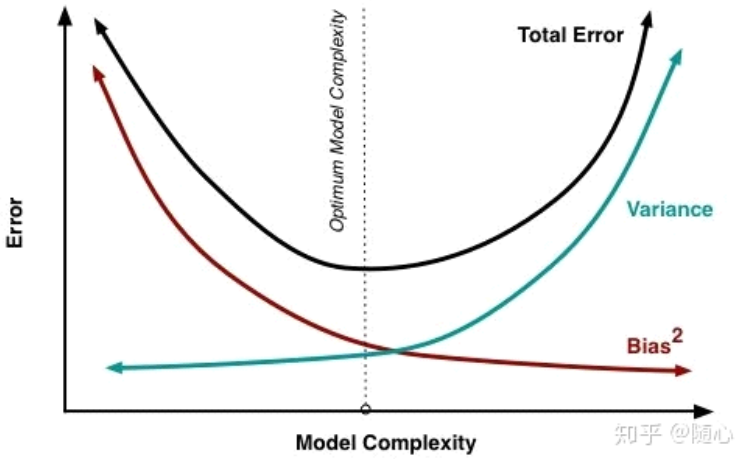
      $$s_{\bar{x}} = (1/\sqrt{n})\,s$$

  - **残差(噪声)：**残差是指预测结果（拟合值）与真实值（观察值）之间的差异，这么一看，和模型偏差的定义很接近，两者的区别是偏差模型拟合度不够导致，而残差是模型准确，但仍然与真实值有一定的差异，这里可以理解成噪声，噪声是随机的，意味着不可预测，而偏差不是随机产生的，可通过一定的特征工程进行预测；
    - 在回归分析中，测定值与按回归方程预测的值之差，以δ表示。残差δ遵从正态分布N(0，σ2)。（δ-残差的均值）/残差的标准差，称为标准化残差，以δ表示。δ遵从标准正态分布N(0，1)。实验点的标准化残差落在(-2，2)区间以外的概率≤0.05。若某一实验点的标准化残差落在(-2，2)区间以外，可在95%置信度将其判为异常实验点，不参与回归直线拟合。显然，有多少对数据，就有多少个残差。残差分析就是通过残差所提供的信息，分析出数据的可靠性、周期性或其它干扰。残差计算即是残差的平方和除以(残差个数-1)的平方根。

- 观测值的误差也被称为扰动，是观测值与总体量(不可观测)真实值的偏差。残差与误差的对比如下图，黑线是真实值，红线是拟合值，绿点是观测值：



- 对模型的影响：
  - 对模型起决定性影响的是偏差和方差，模型过于简单必然导致偏差过大，过于复杂必然导致方差过大，那该如何折中选择



  - 上图可以分为两个部分，以中间的虚线隔开，左边部分为欠拟合状态，右边部分为过拟合状态，针对欠拟合和过拟合的处理方式如下：
    - 欠拟合：偏差过大，做特征工程、减小(弱)正则化系数
    - 过拟合：方差过大，可增加样本、减少特征、增加(强)正则化系数

普通最小二乘OLS和广义最小二乘GLS比较

- 假设你有一把尺子，去测量一个物体的长度。你用同一把尺子测量n次，每次的测量误差就是这个尺子的误差（忽略其他因素），这就是我们所说的最小二乘里的同方差假定。现在你换一种方法，还是测量n次，但是你每次测量用的尺子精度不一样，有点大，有的小。这就是说所谓的“异方差”Heteroscedasticity，这个时候你用普通最小二乘，就会导致估计不一致，这个时候，你想到一个办法就是，对于估计量中的样本，除以相应样本的那把尺子的误差，这样处理之后，就又变成同方差了。
- 异方差：又稱分散不均一性，指的是一系列的隨機變數間的變異數不相同，相對於同質變異數（Homoscedasticity）。當我們利用普通最小平方法（Ordinary Least Squares）進行迴歸估計時，常常做一些基本的假設。其中之一就是誤差項（Error term）的變異數（方差）是不變的。異質變異數是違反這個假設的。如果普通最小平方法應用於異質變異數模型，會導致估計出的變異數值是真實變異數值的偏誤估計量（Biased standard error），但是估計值（estimator）是不偏的（unbiased）。

- 简单地说,用回归变量X来拟合响应变量Y，其中Y中的每个变量，存在内部方差(var)和外部协方差(cov),一起构成协

方差阵(vcv)。想象一下，当你用尺子量一样东西，把尺子所带来的误差想象成内部方差，自己的心情导致的误差为外部协方差。因为X一般当做固定的，所以Y的协方差阵其实也就是误差项的协方差阵

1.如果存在外部协方差,即协方差阵不是对角阵，就是广义最小二乘，当用不同的尺子来量时，除了尺子误差不固定，我每次量的时候我自己的心情也会对这个误差造成影响，所以外部协方差不是对角阵。

2.如果协方差阵是对角阵，且对角线各不相等，就是权重最小二乘，当我用不同的尺子来量时，尺子误差不固定，所以每次量的方差不一样，所以协方差中的对角线不一样。

3.如果协方差阵是对角阵，且对角线相同，就是普通最小二乘法：用一把尺子量时，因为这把尺子的误差是固定的，所以每次量的方差都是一样的，所以协方差阵是对角阵，且对角线都是一样的（同一个尺子的方差）。

*公式都是一样的 min RSS=误差项T * vcv-1 * 误差项（T是转置, -1是逆矩阵）