Degree of Freedom

> reference: 统计学“自由度”详解 - 知乎 (zhihu.com)reference: (9条消息) 【线性回归】线性代数角度解释最小二乘法_Jesszen的博客-CSDN博客_最小二乘法 线性代数

Definition

第四种定义：自由度是一个随机向量的自由维度数，也就是一个向量能被完整描述所需的最少单位向量数。

Fisher 给“Student（t分布的发现者）”解释自由度的时候是这么来解释的：将n个样本随机变量构造成一个随机向量，那么这个向量可以看成是n维空间的一个点，每有一个约束条件，则向量的自由维度减1。比如n个样本在求样本方差的时候要先计算样本均值，所以最后一个变量就和前面n-1个相关，这样随机向量只能有n-1个元素可以在n-1维空间自由取值。在Fisher指出老皮尔逊的卡方检验方法自由度计算错误的时候他是用“约束”这个词来解释的，这个“约束”有点像上面的定义二。

如果所研究的问题能抽象为模型，使用第四种定义计算自由度会容易很多。n个随机样本看成n个随机变量，这n个样本随机变量可以表示为
$$
Iy\\
y \text{是n个样本的随机向量} ,I\text{是n维单位矩阵}
$$

，因为 I 的列空间是n维，Iy=y，所以y一定在 I 的列空间中，y的n个元素可以在n维空间自由取值，推出y的自由度是n。

下面计算线性回归拟合值（回归方程部分）的自由度。
$$
y=\hat{y}+\epsilon, \hat{y}=X\hat{\beta},\\
X\text{是设计矩阵,}\hat{\beta}\text{是估计出来的回归系数向量}
$$

很显然，拟合值向量$\hat{y}$一定在设计矩阵X的列空间中，X列空间维度是回归系数个数，假设有p个预测变量，加上截距则回归系数个数是p+1，所以拟合值向量的自由维度就是p+1。回归平方和
$$
SSR=\sum^n_{i=1}(y_i-\hat{y})^2
$$

是样本因变量的平均值，需要估计出来，失去一个自由度，所以SSR的自由度是拟合值的自由度减1，即p。因变量y的自由度是n，拟合值的自由度是p+1，那么残差向量的自由度是n-(p+1)

y

=

$\hat{y}$

+

ϵ

残差向量

$$\varepsilon$$

垂直于“设计矩阵列空间”(后面有说明)，也就是它在设计矩阵列空间的垂直补空间中，y是n维，设计矩阵列空间是 p+1维，p+1<=n，则残差向量维度是 n-(p+1)，也就是残差向量的自由度为n-(p+1)，接着可以推出残差平方和

$$
SSE=\sum^n_{i=1}(\epsilon_i)^2
$$

的自由度是 n-(p+1)。

Vertical between ϵ and y

1. 我们知道Xβ=y‘，拟合的y’是m维向量。

   观测值y同样是m维向量。

2. 观测值y和拟合值y‘这两个向量，因为必然存在的误差致使Xβ=y【观测值】无解【y观测值不在列空间】。

   那么我们拟合的y’，只能尽可能接近y【观测值】。

3. 我们假设y和y‘不再一个平面，我们知道y’，是由COL(x)线性组合表示的，假设图中的超平面是列空间col(x),那么y‘，必定落在这个平面。【多维，想象超平面】

   perpendicular.png

4. 那么Xβ=y‘和y观测值，怎么才能最接近？换个角度，就是距离最短？

在这个空间我们用欧氏距离度量，我们知道欧氏距离的涉及**到平方和的根号**，所以‘最小二乘法’，中的‘二乘’就是这个概念。

那么最小二乘法，最小又该怎么理解？联想到距离的概念

5. **向量e=y - y’ =y - Xβ**

   |e|自然就是距离【其实这个对应到RSS也就是残差平方和】，距离最小，必然就是正交投影。

   e向量自然属于R m维的子空间，称为也属于左零空间【A转置的零空间**】，左零空间垂直于列空间COL(X)。**