CJK 模板文件

蒋冰

October 15, 2013

1 问题介绍

大多数传统的统计方法是在对研究的总体的分布形式和其他特征做出一些假设的情况下的，而实际总体不满足假设条件，或观察数据中包含不能代表总体特性的异常点（Outliers）时，传统的统计方法可能就会出现较大误差。由于仪器故障、操作失误等原因，异常点是很容易出现的。因此，偏离假设是常见的现象。

2 稳健回归分析

2.1 L估计

L估计（Linear Combination of Order Statistics）是用顺序统计量的现行组合来估计位置的一类估计量。

如果将总体的一组观测值按从小到大的次序排列起来，即$x_1 \leqq x_2 \leqq x_3 \leqq ... \leqq x_n$，它们就成为了一组顺序统计量。

算术平均值可以看作是系数为$\frac{1}{n}$的顺序统计量的线性组合，因而可以看作是L估计量的一种。但算术平均数很不稳定，所以又有学者提出了其他的各种统计量，给不同的顺序统计量以不同的系数，从而打到消弱异常点影响的目的。

2.1.1 α 修剪后均值(Linera Combination of Order Statistics)

α 修剪后均值($0 \leqq \alpha < 0.5$)是这样一种线性组合：

$$T_n = \frac{1}{(1-2\alpha)n} \sum_{i=1}^{n} \alpha_i x_i$$

其中

$$\alpha_i = \begin{cases} 1 & i \\ p & u \\ 0 & i \end{cases}$$

上市中的[α n]表示α n的整数部分，$p = 1 + [\alpha n] - \alpha n$。

α 修剪后均值均值实际上是蒋数据的最大和最小的[α n]个数去掉后剩余部分的加权平均值。当$[\alpha n] \neq \alpha n$时，剩余部分的最小和最大的各一个数的权为p，其余的权为1。当$[\alpha n] = \alpha n$时，剩余的所有输的权均为1.

由α_i的公式可知，当 α $= 0$时，T_n是算术平均值。当 α 趋近于0.5时，T_n是中位数。用中位数来估计位置比算术平均值文件的多。但另一方面，用中位数会损失数据缩提供的信息。因此，可以取 α 为0与0.5之间的适当数值，以在保证一定的稳健性的同时尽可能利用数据所提供的信息。

2.1.2 跳跃估计(Skipped Estimates)

按下式定义两点p_1和p_2:

$$p_1 = h_1 - c(h_2 - h_1)$$

$$p_2 = h_2 + c(h_2 - h_1)$$

式中$h_1 < h_2$为样本的两个四分位数，而c是控制稳健性的参数，通常取值1.0到2.0之间，单重跳跃估计法是先将小于p_1并大于p_2的数据“跳跃过去”，然后对剩余部分求稳健估计值。

2.2 M估计(Maximum Likelihood Estimates)

位置的最小二乘估计是通过对误差平方和$\sum_{i=1}^{n} = \sum_{i=1}^{n}(x_i - T_n)^2$取极小而求得。它是方程$\sum_{i=1}^{n} r_i = 0$的解。

M估计可以看作是对最小二乘估计的一种改善。它是通过对$\sum_{i=1}^{n} \rho(r_i)$取极小而得到的。$\rho(r_i)$是r_i的函数，它随r_i值的增大而增大，通常取$\rho(r_i) = -lnf(r_i)$，其中$f(r_i)$是r_i的分布密度函数。$\rho(r_i)$对取极小就相当于对r_i的似然函数的对数$\sum_{i=1}^{n} lnf(r_i)$取极大。故M统计值被成为最大似然估计。

为了消除测量单位对估计的影响，类似于

2.3 R估计(Rank Estimate)

R估计值及秩估计值，室友双样本秩检验引申来的。首先用算术平均数或中位数等给出T_n的初始值，再由下式构成第二样本

$$Y = (y_1, \cdots, y_n) = (2T_n - x_1, \cdots, 2T_n - x_n)$$

可见Y是X的以T_n为中心的镜像。当X是对称分布且T_n是位置的最佳估计时，X与Y应当完全重合。将两样本混合后记$x_1, \cdots, x_n$在混合样本中的序数为$R_1, \cdots, R_n$，定义秩统计量为

$$S_{n,n} = \sum_{i=1}^{n} a(R_i)$$

$a(R_i)$被称为R_i的得分(Score)，其选择应当使当T_n是位置的最佳估计时，$S_{n,n}$的值为0。例如$a(R_i) = \frac{R_i}{2n+1} - \frac{1}{2}$或$a(R_i) = \frac{R_i - \frac{1}{2}}{2n} - \frac{1}{2}$均可满足这一要求。实际计算时，可以逐渐调节$T_n$的值，使$S_{n,n}$所对应的$T_n$即所求的R估计值。

对明显不对称的分布形式，R估计量的效果并不好。