统计计算 张楠 2019秋: The Jackknife

(返回 统计计算 张楠 2019秋)

Jackknife(刀切法)是由Quenouille(1949,1956)提出的再抽样方法. Jackknife 类似于“leave-one-out”的交叉验证方法. 令 $x = (x_1, \ldots, x_n)$ 为观测到的样本，定义第 i 个 Jackknife 样本为丢掉第 i 个样本后的剩余样本，即 $x_{(i)} = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$. 若 $\hat{\theta} = T_n(x)$, 则定义第 i 个 Jackknife 重复为 $\hat{\theta}_{(i)} = T_{n-1}(x_{(i)})$, $i = 1, \cdots, n$. 假设参数$\theta = t(F)$, 为分布F的函数. F_n为F的经验分布函数. 则θ的“plug-in”估计为$\hat{\theta} = t(F_n)$. 称一个“plug-in”估计$\hat{\theta}$是平滑的, 如果数据的小幅变化相应于$\hat{\theta}$ 的小幅变化.

偏差的Jackknife估计

如果$\hat{\theta}$为一个平滑的(plug-in)估计量，则$\hat{\theta}_{(i)} = t(F_{n-1}(x_{(i)}))$，以及 偏差的Jackknife估计(Quenouille)为 $\widehat{bias}_{jack} = (n-1)(\overline{\hat{\theta}_{(\cdot)}} - \hat{\theta})$, 其中$\overline{\hat{\theta}_{(\cdot)}} = \frac{1}{n}\sum_{i=1}^n \hat{\theta}_{(i)}$. 我们以$\theta$为总体方差为例来说明为什么偏差的Jackknife估计中系数是$n-1$. 由于方差的“plug-in”估计为 $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$. 估计量$\hat{\theta}$是$\sigma^2$的有偏估计，偏差为 $bias(\hat{\theta}) = E\hat{\theta} - \sigma^2 = -\frac{\sigma^2}{n}$. 每一个Jackknife估计是基于样本量$n-1$的样本构造, 因此Jackknife重复$\hat{\theta}_{(i)}$ 的偏差为 $-\frac{\sigma^2}{n-1}$. 所以：

$$E[\hat{\theta}_{(i)} - \hat{\theta}] = E[\hat{\theta}_{(i)} - \theta] - E[\hat{\theta} - \theta] = bias(\hat{\theta}_{(i)}) - bias(\hat{\theta}) = -\frac{\sigma^2}{n-1} - (-\frac{\sigma^2}{n}) = \frac{bias(\hat{\theta})}{n-1}.$$

所以, 在Jackknife偏差估计的定义中有系数$n-1$.

例7 偏差的Jackknife估计 计算patch数据中比值参数的估计偏差的Jackknife估计.

```
data(patch, package = "bootstrap")
n <- nrow(patch)
y <- patch$y
z <- patch$z
theta.hat <- mean(y) / mean(z)
print (theta.hat)
#compute the jackknife replicates, leave-one-out estimates
theta.jack <- numeric(n)
for (i in 1:n)
    theta.jack[i] <- mean(y[-i]) / mean(z[-i])
bias <- (n - 1) * (mean(theta.jack) - theta.hat)
print(bias)  #jackknife estimate of bias
```

标准差的Jackknife 估计

对平滑的统计量$\hat{\theta}$, 其标准差的Jackknife估计(Tukey)定义为 $\hat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^n (\hat{\theta}_{(i)} - \overline{\hat{\theta}_{(\cdot)}})^2}$. 比如当$\theta$为总体均值时, $\hat{\theta} = \bar{x}$, 其方差估计为 $Var(\hat{\theta}) = \frac{\hat{\sigma}^2}{n} = \frac{1}{n(n-1)}\sum_{i=1}^n (x_i - \bar{x})^2$ 记$\theta_{(i)} = \frac{n\bar{x} - x_i}{n-1}$, 则$\overline{\hat{\theta}_{(\cdot)}} = \frac{1}{n}\hat{\theta}_{(i)} = \hat{\theta}$, $\hat{\theta}_{(i)} - \overline{\hat{\theta}_{(\cdot)}} = \frac{\bar{x} - x_i}{n-1}$. 因此有 $\hat{se}_{jack} = \sqrt{Var(\hat{\theta})}$. **例8** 标准差的Jackknife估计 计算patch数据中比值参数的估计标准差的Jackknife估计.

```
se <- sqrt((n-1) *
    mean((theta.jack - mean(theta.jack))^2))
print(se)
```

Jackknife失效情形

若估计量$\hat{\theta}$不够平滑, Jackknife方法就可能会失效. 中位数就是一个不平滑统计量的例子.

例9 (Jackknife方法失效) 用Jackknife方法估计从1,2,...,100中随机抽取的10个数的中位数的标准差.

```
set.seed(123) #for the specific example given
#change the seed to see other examples
n <- 10
x <- sample(1:100, size = n)
#jackknife estimate of se
M <- numeric(n)
for (i in 1:n) {        #leave one out
    y <- x[-i]
    M[i] <- median(y)
}
Mbar <- mean(M)
print(sqrt((n-1)/n * sum((M - Mbar)^2)))
#bootstrap estimate of se
Mb <- replicate(1000, expr = {
        y <- sample(x, size = n, replace = TRUE)
        median(y) })
print(sd(Mb))
```

本例中Jackknife估计和Bootstrap估计相差很远, 显然存在问题. 事实上, 由于中位数不是平滑的, Jackknife方法失效了.