Bootstrap算法

许王莉

中国人民大学 统计学院

主要内容

- √ Bootstarp方法
- √ 自助法估计标准差
- √ 自助法估计偏差
- √ 刀切法Jackknife
- √ 偏差的Jackknife估计
- √ 标准差的Jackknife估计
- √ 估计的方差(Jackknife-after-Bootstrap)
- √ 自助置信区间
- √ 交叉验证

Bootstarp方法

(1). Efron 在1979,1981和1982年的工作中引入和进一步发展了Bootstarp方法, 此后发表了大量的关于此方法的研究.

(2). Bootstrap方法是一类非参数Monte Carlo方法, 其通过再抽样对总体分布进行估计. 再抽样方法将观测到的样本视为一个有限总体, 从中进行随机(再)抽样来估计总体的特征以及对抽样总体作出统计推断. 当目标总体分布没有指定时, Bootstrap方法经常被使用, 此时, 样本是唯一已有的信息.

Bootstrap方法

- √ Bootstrap 一词可以指非参数Bootstrap, 也可以指参数Bootstrap.参数Bootstrap是指总体分布完全已知, 利用Monte Carlo方法从此总体中抽样进行统计推断; 而非参数Bootstrap是指总体分布完全未知, 利用再抽样方法从样本中再抽样进行统计推断.
- √ 可以视样本所表示的有限总体的分布为一个"伪"总体, 其具有和真实总体类似的特征. 通过从此"伪"总体中重复(再) 抽样, 可以据此估计统计量的抽样分布. 统计量的一些性质, 如偏差, 标准差等也可以通过再抽样来估计.

Bootstrap方法

一个抽样分布的Bootstrap估计类似于密度估计的想法. 我们通过一个样本的直方图来估计密度函数的形状. 直方图不是密度, 但是在非参数问题中, 可以被视为是密度的一个合理估计. 我们有很多方法从已知的密度中产生随机样本, **Bootstrap则从经验分布中产生随机样本.**

假设$x = (x_1, \cdots, x_n)$为一个从总体分布$F(x)$中观测到得样本, X^*为从x中随机选择的一个样本, 则$P(X^* = x_i) = \frac{1}{n}, \quad i = 1, \cdots, n$. 从$x$中有放回的再抽样得到随机样本$X_1^*, \cdots, X_n^*$. 显然随机变量$X_1^*, \cdots, X_n^*$ 为$i.i.d$的随机变量, 服从$\{x_1, \cdots, x_n\}$上的均匀分布.

Bootstrap方法

经验分布函数$F_n(x)$是$F(x)$的估计，可以证明，$F_n(x)$是$F(x)$的充分统计量. 而且另一方面，$F_n(x)$ 本身是$\{x_1, \cdots, x_n\}$上的均匀分布随机变量X^*的分布函数.

因此在Bootstrap中有这个逼近. F_n逼近到F, Bootstrap重复下的经验分布函数F_n^*是F_n的逼近. 从x中再抽样,等价于从F_n中产生随机样本. 这两种逼近可以表示为

$$F \to X \to F_n$$

$$F_n \to X^* \to F_n^*$$

从x中产生一个Bootstrap随机样本可以这样实现，先从$\{1, 2, \cdots, n\}$中有放回的选取n次得到$\{i_1, \cdots, i_n\}$, 然后得到Bootstrap样本$x^* = (x_{i_1}, \cdots, x_{i_n})$.

假设θ是我们感兴趣的参数(向量), $\hat{\theta}$为θ的估计. 则$\hat{\theta}$分布的Bootstrap 估计可以通过如下方法得到

(1). 对Bootstrap重复的第b次$(b = 1, \ldots, B)$,

- 通过有放回的从$x_1, \cdots, x_n$中抽样得到再抽样样本$x^{*(b)} = x_1^*, \ldots, x_n^*$.
- 根据$x^{*(b)}$计算$\hat{\theta}^{(b)}$.

(2). $F_{\hat{\theta}}(\cdot)$的Bootstrap估计为$\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$的经验分布函数.

例

例：F_n与*Bootstrap*抽样

假设我们观察到样本$x=\{2,2,1,1,5,4,4,3,1,2\}$从$x$中再抽样依照选择1, 2, 3, 4, 5的概率分别为0.3, 0.3, 0.1, 0.2, 0.1进行. 从而从x中随机选择的一个样本X^*, 其分布函数就是经验分布函数, 即

$$F_{X^*}(x)=F_n(x)=\begin{cases}0, & x<1;\\ 0.3, & 1\leq x<2;\\ 0.6, & 2\leq x<3;\\ 0.7, & 3\leq x<4;\\ 0.9, & 4\leq x<5;\\ 1, & x\geq 5.\end{cases}$$

例

例(续)：注意如果F_n没有靠近F_X, 则重复抽样下的分布也不会靠近F_X. 上例中的样本x实际上是从*Poisson*(2)中随机产生的, 从x中大量重复抽样可以很好的估计F_n, 但是不能很好的估计F_X, 因为无论重复多少次再抽样, 得到的*Bootstrap* 样本都没有0.

Bootstrap Estimation of Standard Error

估计量$\hat{\theta}$的标准差的Bootstrap估计，是Bootstrap重复$\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$ 的样本标准差：

$$\hat{se}_B(\hat{\theta}^*) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \overline{\hat{\theta}^*})^2}.$$

其中$\overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$. 根据Efron和Tibshirini(1993), 要得到标准差一个好的估计, 重复的次数B并非需要非常大. $B = 50$**常常已经足够了, $B > 200$是很少见的(置信区间除外).**

Bootstrap Estimation of Standard Error

例

例: (标准差的Bootstrap估计) 根据如下15所法律院校入学考试的平均成绩(LSAT)和GPA(乘了100).

	1	2	3	4	5	6	7	8
LSAT	576	635	558	578	666	580	555	661
GPA	339	330	281	303	344	307	300	343

	9	10	11	12	13	14	15
LSAT	651	605	653	575	545	572	594
GPA	336	313	312	274	276	288	296

估计LSAT和GPA之间的相关系数, 并求样本相关系数的标准差的Bootstrap估计.

Bootstrap Estimation of Standard Error

- 在本例中，数据是成对的$(x_i, y_i), i = 1, \ldots, 15$.
- 可以通过样本相关系数估计相关系数

$$\hat{\tau} = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\sqrt{n\sum_i x_i^2 - (\sum_i x_i)^2}\sqrt{n\sum_i y_i^2 - (\sum_i y_i)^2}}.$$

- 对这些数据对Bootstrap再抽样.

Bootstrap Estimation of Standard Error

因此, 算法如下

- 对Bootstrap重复的第b次($b = 1, \ldots, B$),
  - 通过有放回的从$x_1, \cdots, x_n$中抽样得到再抽样样本$x^{*(b)} = x_1^*, \ldots, x_n^*$. 这里$x_i$或者$x_i^*$为一个向量.
  - 根据$x^{*(b)}$计算$\hat{\tau}^{(b)}$.
- $F_{\hat{\tau}}(\cdot)$的Bootstrap估计为$\hat{\tau}^{(1)}, \ldots, \hat{\tau}^{(B)}$的经验分布函数.

Bootstrap Estimation of Standard Error

使用Bootstrap估计标准差的程序如下:

```
%for the data
LSAT=[576,635,558,578,666,580,555,661,651,605,653,575,545,572,594];
GPA=[339,330,281,303,344,307,300,343,336,313,312,274,276,288,296];
min(min(corrcoef(LSAT,GPA)))
%set up the bootstrap
B=200;
n=length(LSAT);
%bootstrap estimate of standard error of R
for b=1:B
    %randomly select the indices
    i=unidrnd(n,n,1);
    SLSAT=LSAT(i);
    SGPA=GPA(i);
    R(b)=min(min(corrcoef(SLSAT,SGPA)));
end
%output
std(R)
```

Bootstrap Estimation of Standard Error

问题：验证自助法计算的$\overline{x}$方差的估计是否为$VAR(\overline{x})$的无偏估计?

条件如下：$x_1, x_2, \cdots, x_n$ 服从正态分布$N\left(\mu, \sigma^2\right)$, 且$n = 20$

$$\widehat{VAR}(\widehat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\theta}_b - \hat{\theta}\right)^2$$

其中，$\widehat{\hat{\theta}} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\theta}_b$.

Bootstrap Estimation of Standard Error

我们可以计算出如下等式$\widehat{\theta_b} = \frac{\sum_{i=1}^{n} x_i'}{n}$, 故

$$VAR\left(\widehat{\theta_b}|x_1, x_2 \ldots x_n\right) = \frac{\text{var}\left(x'\right)}{n} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n^2}$$

$$E\left(\widehat{\theta}^b|x_1, x_2 \ldots x_n\right) = \overline{x}.$$

其中， $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Bootstrap Estimation of Standard Error

此处应注意，此处的x_1到x_n在再抽样中已经作为定值，不再为随机变量，bootstrip 即为从概率均为$1/n$的分布列$x_1, \cdots, x_n$ 中抽取样本，不能再以正态分布进行计算。当然，我们还可以计算出其条件期望与条件方差的期望。即去除x_1到x_n的随机性

$$E\left(VAR\left(\widehat{\theta_b}\right)\right) = E\left(VAR\left(\widehat{\theta_b}|x_1, x_2 \ldots \ldots x_n\right)\right) = \frac{(n-1)\sigma^2}{n^2}.$$

$$E\left(\widehat{\theta^b}\right) = E\left(E\left(\widehat{\theta^b}|x_1, x_2 \ldots x_n\right)\right) = \mu.$$

Bootstrap Estimation of Standard Error

在之后的计算中，我们需将x_1到x_n的随机性与bootstrap 抽样的随机性全部去除，即每一步都进行两次期望，进一步可以算出

$$E\left(\widehat{\theta_b^2}\right) = E\left(E\left(\hat{\theta}_b^2 | x_1, x_2 \ldots x_n\right)\right) = VAR\left(\widehat{\theta_b}\right) + E^2\left(\widehat{\theta_b}\right) = \mu^2 + \frac{(n-1)\sigma^2}{n^2}$$

$$E(\widehat{\widehat{\theta}}) = E\left(\frac{1}{B}\sum_{b=1}^{B}\widehat{\theta_b}\right) = \mu$$

$$VAR(\widehat{\hat{\theta}}) = VAR\left(\frac{1}{B}\sum_{b=1}^{B}\widehat{\theta_b}\right)$$

$$= \frac{VAR\left(\widehat{\theta_b}\right)}{B} = \frac{(n-1)\sigma^2}{n^2 B}$$

Bootstrap Estimation of Standard Error

于是

$$
\begin{aligned}
E(\widehat{VAR}(\hat{\theta})) &= E\left(\widehat{VAR}\left(\widehat{\theta}|x_1, x_2 \ldots x_n\right)\right) \\
&= = \frac{1}{B-1} E\left(\sum_{b=1}^{B} \widehat{\theta}_b^2 - B\bar{\theta}^2\right) \\
&= \frac{1}{B-1}\left(\sum_{b=1}^{B} E\left(\widehat{\theta}_b^2\right) - BE\left(\hat{\bar{\theta}}^2\right)\right) \\
&= \frac{1}{B-1}\left(B\left(\mu^2 + \frac{(n-1)\sigma^2}{n^2}\right) - B\left(E^2(\widehat{\theta}) + VAR(\widehat{\theta})\right)\right) \\
&= \frac{1}{B-1}\frac{(n-1)(B-1)\sigma^2}{n^2} = \frac{(n-1)\sigma^2}{n^2}
\end{aligned}
$$

由此我们可以发现，自助法估计出的方差并不是无偏估计，其估计值偏小。

Bootstrap Estimation of Bias

- √ θ的一个估计量$\hat{\theta}$的偏差定义为$bias(\hat{\theta}) = E\hat{\theta} - \theta$.
- √ 当$\hat{\theta}$的分布未知或者形式很复杂使得期望的计算不可能(从此分布中抽样变得很困难, Monte Carlo方法不可行), 以及在现实中, 我们也不知道θ的真值时(需要估计), 这种情况下偏差是未知的.
- √ 但是我们已经有了样本, $\hat{\theta}$是θ的估计, 而期望$E\hat{\theta}$可以通过Bootstrap方法进行估计. 从而可以得到偏差的估计:

$$\widehat{bias}_B(\hat{\theta}) = E^*\hat{\theta}^* - \hat{\theta}.$$

E^*表示Bootstrap经验分布.

Bootstrap Estimation of Bias

因此一个估计量的偏差的Bootstrap估计，是通过使用当前样本下的估计量$\hat{\theta}$来估计θ，而使用$\hat{\theta}$的Bootstrap重复来估计$E\hat{\theta}$.

对一个有限样本$x = (x_1, \cdots, x_n)$，有$\hat{\theta}(x)$的B个i.i.d估计量$\hat{\theta}^{(b)}$. 则$\{\hat{\theta}^{(b)}\}$的均值是期望值$E\hat{\theta}^*$的无偏估计，因此偏差的Bootstrap估计为

$$\widehat{bias}_B(\hat{\theta}) = \overline{\hat{\theta}^*} - \hat{\theta}.$$

这里$\overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$.

正的偏差意味着$\hat{\theta}$平均来看过高估计了θ；而负的偏差意味着$\hat{\theta}$平均来看过低估计了θ. 因此，一个经过偏差修正(Bias-correction)的估计量为$\tilde{\theta} = \hat{\theta} - \widehat{bias}_B(\hat{\theta})$.

Bootstrap Estimation of Bias

例：Bootstrap偏差估计: 估计上例中样本相关系数的偏差

```
LSAT=[576,635,558,578,666,580,555,661,651,605,653,575,545,572,594];
GPA=[339,330,281,303,344,307,300,343,336,313,312,274,276,288,296];
theta.hat=min(min(corrcoef(LSAT,GPA)));
%bootstrap estimate of bias
B=2000;
n=length(LSAT);
for b=1:B
    i=unidrnd(n,n,1);
    SLSAT=LSAT(i);
    SGPA=GPA(i);
    theta.b(b)=min(min(corrcoef(SLSAT,SGPA)));
end
bias=mean(theta.b-theta.hat)
```

这个值和例3中的函数返回的结果非常相近.

Bootstrap Estimation of Bias

例5: Bootstrap偏差估计: 假设$x = (x_1, \cdots, x_{10}) \sim N(\mu, \sigma^2)$, 求$\sigma^2$的估计量$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$的偏差

```
n=10;
x=normrnd(0,10,1,n);
sigma2.hat=(n-1)*var(x)/n;
%bootstrap estimate of bias
B=2000;%larger for estimating bias
for b=1:B
    i=unidrnd(n,n,1);
    sigma2.b(b)=(n-1)*var(x(i))/n;
end
bias=mean(sigma2.b-sigma2.hat)
```

在这种情形下, $\hat{\sigma}^2$过低的估计了参数σ^2.

Bootstrap Estimation of Bias

例：比值参数估计的偏差的Bootstrap估计.

以包bootstrap里的patch数据为例. 该数据是测量了8个人使用3种不同的药物后血液中某种荷尔蒙的含量. 这三种药物分别是安慰剂, 旧药品(经过FDA审批的), 新药品(某个新工厂相同的工艺下生产的, 按FDA规定, 新工厂生产的药品也要审批). 研究的目的是比较新药和旧药的等价性. 如果可以证明新药和旧药之间的等价性, 则对新药就不需要完全重新向FDA申请审批了. 等价性的标准是对比值参数

$$\theta = \frac{E(new) - E(old)}{E(old) - E(placebo)}.$$

若$|\theta| \le 0.20$, 则新药和旧药就等价. 估计θ的统计量为$\bar{Y}/\bar{Z}$. 这两个变量在patch数据中给出. 我们的目标是计算此估计偏差的Bootstrap估计.

Bootstrap Estimation of Bias

```
y=[-1200,2601,-2705,1982,-1290,351,-638,-2719];
z=[8406,2342,8187,8459,4795,3516,4796,10238];
n=length(y);
B=2000;
theta.hat=mean(y)/mean(z);
%bootstrap
for b=1:B
    i=unidrnd(n,n,1);
    yy=y(i);
    zz=z(i);
    theta.b(b)=mean(yy)/mean(zz);
end
est=theta.hat
bias=mean(theta.b)-theta.hat
se=std(theta.b)
cv=bias/se
```

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, $\overline{\hat{\theta}_{(\cdot)}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$. 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:

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

所以, 在Jackknife偏差估计的定义中有系数$n-1$.

偏差的Jackknife估计

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.

```
y=[-1200,2601,-2705,1982,-1290,351,-638,-2719];
z=[8406,2342,8187,8459,4795,3516,4796,10238];
n=length(y);
theta.hat=mean(y)/mean(z)
%compute the jackknife replicates,leave-one-out estimates
for i=1:n
    yy=y;
    yy(i)=[];
    zz=z;
    zz(i)=[];
    theta.jack(i)=mean(yy)/mean(zz);
end
bias=(n-1)*(mean(theta.jack)-theta.hat)
```

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θ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$\hat{se}_{jack} = \sqrt{Var(\hat{\theta})}$.

标准差的Jackknife估计

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))
```

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.

标准差的Jackknife估计

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.

```
n=10;
x=[29,79,41,86,91,5,50,83,51,42]; %x=unidrnd(100,n,1);
%jackknife estimate of se
for i=1:n   %leave one out
    y=x;
    y(i)=[];
    M(i)=median(y);
end
Mbar=mean(M); sqrt((n-1)/n*sum((M-Mbar).^2))
%bootstrap estimate of se
for b=1:1000
    i=unidrnd(n,n,1);
    y=x(i);
    Mb(i)=median(y);
end
std(Mb)
```

标准差的Jackknife估计

问题：对于刀切法对方差的估计，即$\theta=\sigma$，求该估计（下式）与$\hat{\theta}$的关系

$$\hat{E}\hat{\theta}=\frac{n-1}{n}\sum_{i=1}\left(\hat{\theta}_{(i)}-\hat{\theta}\right)+\hat{\theta}$$

仅需讨论$\hat{E}\hat{\theta}-\hat{\theta}$即可，即

$$bias_{jack}=\frac{n-1}{n}\sum_{i=1}\left(\hat{\theta}_{(i)}-\hat{\theta}\right)$$

其中

$$\hat{\theta}_{(i)}=\frac{\sum_{j\neq i}(x_j-\overline{x}_j)^2}{n-1}$$

$$=\frac{\sum_{j\neq i}x_j^2-(n-1)\left(\frac{\sum_{j\neq i}x_j}{n-1}\right)^2}{n-1}$$

标准差的Jackknife估计

故可得

$$\sum_{i=1} \hat{\theta}_{(i)} = \frac{1}{n-1}\left((n-1)\sum_{i=1} x_i^2 - \frac{1}{n-1}\sum_{j=1}\left(\sum_{i=1} x_i - x_j\right)\right)$$

$$= \frac{1}{n-1}\left((n-1)\sum_{i=1} x_i^2 - \frac{\sum_{i=1} x_i^2}{n-1} - \frac{n-2}{n-1}\left(\sum_{i=1} x_i\right)^2\right)$$

又因为$\hat{\theta} = \frac{\sum_{i=1}(x_i - \hat{x})^2}{n} = \frac{1}{n}\left(\sum_{i=1} x_i^2 - \frac{(\sum_{i=1} x_i)^2}{n}\right)$

因此

$$
\begin{aligned}
&\text{bias}_{\text{ jack}} \\
&= \frac{n-1}{n}\sum_{i=1}\left(\hat{\theta}_{(i)}-\hat{\theta}\right) \\
&= \frac{n-1}{n}\left(\sum_{i=1}x_i^2-\frac{\sum_{i=1}x_i^2}{(n-1)^2}-\frac{n-2}{(n-1)^2}\left(\sum_{i=1}x_i\right)^2-\left(\sum_{i=1}x_i^2-\frac{(\sum_{i=1}x_i)^2}{n}\right)\right) \\
&= \frac{n-1}{n}\left(\frac{1}{n(n-1)^2}\left(\left(\sum_{i=1}x_i\right)^2-n\sum_{i=1}x_i^2\right)\right) \\
&= -\frac{1}{n(n-1)}\sum_{i=1}(x_i-\overline{x})^2=-\frac{1}{n-1}\hat{\theta}
\end{aligned}
$$

所以有$E\hat{\theta}=\frac{n-1}{n}\sum_{i=1}\left(\hat{\theta}_{(i)}-\hat{\theta}\right)+\hat{\theta}=\frac{n-2}{n-1}\hat{\theta}$

Jackknife-after-Bootstrap

前面我们介绍了使用一个估计量的偏差和标准差的Bootstrap估计. 这些估计本身又是估计量, 那么这些估计量的方差该如何估计呢? 一种方法就是使用Jackknife方法来估计这些估计量的方差.

注意到$\hat{se}(\hat{\theta})$是B次$\hat{\theta}$的Bootstrap重复统计量的样本标准差, 那么如果我们丢掉第i个样本, 则Jackknife算法就是对每个i, 从剩余的$n-1$个样本值中再抽样B次, 来计算$\hat{se}(\hat{\theta}_{(i)})$ (Bootstrap过程), 即一个Jackknife重复. 最后我们得到

$$\hat{se}_{jack}(\hat{se}_B(\hat{\theta})) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{se}_{B(i)}(\hat{\theta}) - \overline{\hat{se}_{B(\cdot)}}(\hat{\theta}))^2}$$

其中$\overline{\hat{se}_{B(\cdot)}(\hat{\theta})} = \frac{1}{n}\sum_{i=1}^{n}\hat{se}_{B(i)}(\hat{\theta})$. 即对每个$i$, 我们将重复Bootstrap本身. 这当然是效率低下的, 庆幸的是有方法可以避

Jackknife-after-Bootstrap

”Jackknife-after-Bootstrap” 方法是对每个"leave-one-out"的Bootstrap样本计算一个估计. 具体如下:

记$x_i^* = (x_1^*, \ldots, x_n^*)$为一次Bootstrap抽样, $x_1^*, \cdots, x_B^*$表示样本大小为B的Bootstrap样本. 令$J(i)$表示Bootstrap样本中不含x_i的那些样本指标; $B(i)$表示不含x_i的Bootstrap样本个数, 因此我们可以使用丢掉$B - B(i)$个含有x_i的样本后其余的样本来计算一个Jackknife重复. 故标准差估计量的Jackknife估计为$\hat{se}_{jab}(\hat{se}_B(\hat{\theta})) = \sqrt{\frac{n-1}{n}\sum_{i=1}^n(\hat{se}_{B(i)} - \overline{\hat{se}_{B(\cdot)}})^2}$, 其中

$$\hat{se}_{B(i)} = \sqrt{\frac{1}{B(i)}\sum_{j\in J(i)}[\hat{\theta}_{(j)} - \overline{\hat{\theta}_{(J(i))}}]^2}, \qquad \overline{\hat{\theta}_{(J(i))}} = \frac{1}{B(i)}\sum_{j\in J(i)}\hat{\theta}_{(j)}.$$

例10: (Jackknife-after-Bootstrap), 对例6中标准差的Bootstrap 估计$\hat{se}_B(\hat{\theta})$, 使用Jackknife-after-Bootstrap 方法估计其标准差.

```
%initialize
y=[-1200,2601,-2705,1982,-1290,351,-638,-2719];
z=[8406,2342,8187,8459,4795,3516,4796,10238];
n=length(y);
B=2000;
%set up storage for the sample indices
indices=zeros(B,n);
%jackknife-after-boostrap step 1: run the bootstrap
for b=1:B
    i=unidrnd(n,n,1);
    yy=y(i);
    zz=z(i);
    theta.b(b)=mean(yy)/mean(zz);
    %save the indices for the jackknife
    indices(b,:)=i;
end
%jackknife-after-bootstrap to est. se(se)
for i=i:n
    %in i-th replicate omit all samples with x(i)
    [a,b]=find(indices==i);
    se.jack(i)=std(theta.b(a));
end
std(theta.b); sqrt((n-1)*mean((se.jack-mean(se.jack)).^2))
```

自助置信区间Bootstrap Confidence Intervals

本节中我们介绍几种在Bootstrap中构造目标参数的渐近置信区间的方法, 其中包括标准正态Bootstrap置信区间, 基本的Bootstrap置信区间, Bootstrap百分位数(percentile)置信区间和Bootstrap t 置信区间.

The Standard Normal Bootstrap Confidence Interval

标准正态Bootstrap置信区间是一种比较简单的方法. 假设$\hat{\theta}$是参数θ的估计量, 以及估计量的标准差为$se(\hat{\theta})$. 若$\hat{\theta}$渐近到正态分布, 即$Z = \frac{\hat{\theta} - E\hat{\theta}}{se(\hat{\theta})}$ 渐近服从标准正态分布. 则若$\hat{\theta}$为θ的无偏估计, 那么θ的一个渐近的$100(1-\alpha)\%$ 标准正态Bootstrap 置信区间为$\hat{\theta} \pm z_{\alpha/2}\hat{se}_B(\hat{\theta})$, 其中$z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. 此区间容易计算, 但是有正态性假设或者CLT需成立. 以及$\hat{\theta}$为θ的无偏估计.

The Percentile Bootstrap Confidence Interval

由形式$P(L \le \hat{\theta} \le U) = 1 - \alpha$ 知, 可以使用Bootstrap重复的样本百分位数来估计L和U. 而$\hat{\theta}$为θ 的估计, 因此就取θ的$1-\alpha$置信区间上下界分别为Bootstrap重复的样本$1-\alpha/2$ 百分位数$\hat{\theta}^*_{[(B+1)(1-\alpha/2)]}$ 和$\alpha/2$百分位数$\hat{\theta}^*_{[(B+1)\alpha/2]}$.

Efron & Tibshirani 证明了百分位数区间相比于标准正态区间有着更好的理论覆盖率. 下面我们还会介绍"bias-corrected and accelerated"(BCa) 百分位数区间, 它是百分位数区间的一个修正版本, 有着更好的理论性质以及在应用中有着更好的覆盖率.

The Basic Bootstrap Confidence Interval

由$P(L < \hat{\theta} - \theta < U) = 1 - \alpha$ 在$\hat{\theta} - \theta$的分布未知时, 由于Bootstrap重复$\hat{\theta}^*$的样本分位数$\hat{\theta}^*_{[(B+1)\alpha/2]}$和$\hat{\theta}^*_{[(B+1)(1-\alpha/2)]}$满足$P(\hat{\theta}^*_{[(B+1)\alpha/2]} - \hat{\theta} \leq \theta^* - \hat{\theta} \leq \hat{\theta}^*_{[(B+1)(1-\alpha/2)]} - \hat{\theta}) \approx 1 - \alpha$. 因此$100(1-\alpha)\%$置信区间为$(2\hat{\theta} - \hat{\theta}^*_{[(B+1)(1-\alpha/2)]}, 2\hat{\theta} - \hat{\theta}^*_{[(B+1)\alpha/2]})$.

The Bootstrap t interval

即使当$\hat{\theta}$的分布是正态分布, 且$\hat{\theta}$为θ的无偏估计, $Z = (\hat{\theta} - \theta)/se(\hat{\theta})$ 的分布也不会一定是正态的, 这是因为我们估计了$se(\hat{\theta})$. 我们也不能说Z的分布是t分布, 因为Bootstrap估计$\hat{se}(\hat{\theta})$的分布未知. Bootstrap t区间并没有使用t分布作为推断分布. 而是使用再抽样方法得到一个"t类型"的统计量的分布. 假设$x = (x_1, \ldots, x_n)$为观测到得样本, 则$100(1-\alpha)\%$ Bootstrap t 置信区间为$(\hat{\theta} - t^*_{1-\alpha/2}\hat{se}(\hat{\theta}), \hat{\theta} - t^*_{\alpha/2}\hat{se}(\hat{\theta}))$ 其中$\hat{se}(\hat{\theta}), t^*_{\alpha/2}$ 和$t^*_{1-\alpha/2}$由下面的方法计算:

The Bootstrap t 区间

- 计算观测到得$\hat{\theta}$.
- 对每个重复, $b = 1, \cdots, B$:
  - 从x中有放回的抽样得到第b个样本$x^{(b)} = (x_1^{(b)}, \ldots, x_n^{(b)})$.
  - 由第b个再抽样样本计算$\hat{\theta}^{(b)}$.
  - 计算标准差估计$\hat{se}(\hat{\theta}^{(b)})$.(对每个Bootstrap样本$x^{(b)}$, 再单独进行一个Bootstrap估计).
  - 计算第b个重复下的"t"统计量: $t^{(b)} = (\hat{\theta}^{(b)} - \hat{\theta})/\hat{se}(\hat{\theta}^{(b)})$.
- 重复样本$t^{(1)}, \ldots, t^{(B)}$的分布作为推断分布. 找出样本分位数$t^*_{\alpha/2}$和$t^*_{1-\alpha/2}$.
- 计算$\hat{se}(\hat{\theta})$, 即Bootstrap重复$\{\hat{\theta}^{(b)}\}$的样本标准差.
- 计算置信界$(\hat{\theta} - t^*_{1-\alpha/2}\hat{se}(\hat{\theta}), \hat{\theta} - t^*_{\alpha/2}\hat{se}(\hat{\theta}))$.

The Bootstrap t 区间

Bootstrap t区间的一个缺点是要再次使用Bootstrap方法得到标准差的估计$\hat{se}(\hat{\theta}^{(b)})$. 这是在Bootstrap 里面嵌套Bootstrap. 若$B = 1000$, 则Bootstrap t 区间方法需要比别的方法1000倍的时间.

Better Bootstrap Confidence Intervals

对百分位数区间进行修正可以得到更好的Bootstrap置信区间，其具有更好的理论性质和更好的实际覆盖率.

对$100(1-\alpha)\%$置信区间，使用两个因子来调整常用的$\alpha/2$和$1-\alpha/2$分位数：一个偏差(bias)的修正，一个偏度(skewness)的修正. 偏差的修正记为z_0，偏度或者"加速"修正记为a. 更优的Bootstrap置信区间也常称为BCa.

$100(1-\alpha)\%$ BCa 置信区间为：先计算

$$\alpha_1 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1-\hat{a}(\hat{z}_0 + z_{\alpha/2})}),$$
$$\alpha_2 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1-\hat{a}(\hat{z}_0 + z_{1-\alpha/2})})$$

其中$z_\alpha = \Phi^{-1}(\alpha)$. $\hat{z}_0, \hat{a}$由下面的式子计算. 则BCa 区间为$(\hat{\theta}^*_{\alpha_1}, \hat{\theta}^*_{\alpha_2})$. BCa 区间的下界和上界是Bootstrap重复的经验的α_1和α_2分位数.

Better Bootstrap Confidence Intervals

偏差修正因子实际上是测量$\hat{\theta}$的重复$\hat{\theta}^*$的中位数偏差. 其估计为$\hat{z}_0 = \Phi^{-1}(\frac{1}{B}\sum_{b=1}^{B} I(\hat{\theta}^{(b)} < \hat{\theta}))$. 加速因子是从Jackknife重复中估计: $\hat{a} = \frac{\sum_{i=1}^{n}(\overline{\theta_{(\cdot)}}-\theta_{(i)})^3}{6\sum_{i=1}^{n}((\overline{\theta_{(\cdot)}}-\theta_{(i)})^2)^{3/2}}$. $\hat{a}$之所以称为是加速因子, 是因为它估计的是相对于目标参数θ, $\hat{\theta}$的标准差的变化率. 我们在使用标准正态Bootstrap置信区间时, 是假设方差为一个常数, 与θ无关. 但是很多时候方差都可能和θ有关. 加速因子的目的就是要考虑到估计量的方差可能会与目标参数有关, 因此对置信界进行调整.

BCa方法的来源可以参看阅读材料.

"BCa的性质"

BCa Bootstrap置信区间有两个重要的理论性质:

一是不变性, 即若θ的置信区间为$(\hat{\theta}^*_{\alpha_1}, \hat{\theta}^*_{\alpha_2})$, $g(\cdot)$为一一变换函数, 则$g(\theta)$得置信区间为$(g(\hat{\theta}^*_{\alpha_1}), g(\hat{\theta}^*_{\alpha_2}))$.

另外一个性质是二阶精确性, 即误差以$1/n$的速度趋于0.

Bootstrap t 置信区间是二阶精确的, 但是不具有不变性. Bootstrap 百分位数区间有不变性, 但是不是二阶精确的; 标准正态置信区间既没有不变性, 也没有二阶精确性.

交叉验证Cross Validation

交叉验证(Cross Validation)是一种分割数据方法，其可以用来验证参数估计的稳健性，分类算法的准确度，模型的合理性等等. Jackknife 可以被视为是交叉验证的一种特例，其主要用来估计偏差和估计量的标准差.

最简单的交叉验证方法是所谓的"holdout" 方法. 其将数据随机分为训练集(training set)和测试集(testing set)两个子集. 然后仅使用训练集样本进行建模，然后通过测试集来对模型进行评估. 其优点是training/testing 比例不依赖于重复次数. 其缺点是依赖于数据的分割方式，某些点可能从不进入到测试集中，而某些点可能多次进入测试集. 这种方法呈现出"Mente Carlo"波动性，即随机分割不同，结果会波动.

"K-fold"交叉验证是对"holdout"方法的改进, 其将数据分割为K个子集, 然后重复"holdout"方法K次. 每次第i个子集被作为测试集来评估模型, 其余的$K-1$个子集被作为训练集进行建模. 最后计算K次的平均误差. 其优点是对数据的分割方式依赖性不是很强, 每个点仅有一次在测试集中, 而在训练集中有$K-1$次. 因此估计的方差会随着K的增加而减少. 缺点是计算的时间复杂度增加.

"leave-one-out" 交叉验证是"K-fold"交叉验证的一个特例($K=n$), 仅使用一个点作为测试集, 其余的点作为训练集. 其缺点是计算的时间复杂度可能会比较高.

例18: 模型选择问题包DAGG 里的ironslag 数据描述了两种方法(chemical, magnetic)测量含铁量的53次结果. 散点图显示chemical和magnetic变量是正相关的, 但是关系可能不是线性的. 从散点图上可以看出, 二次多项式, 或者可能一个指数的, 或对数模型能更好的拟合数据.

交叉验证Cross Validation

本例我们使用交叉验证来进行模型选择. 通过交叉验证来估计模型的预测误差. 候选的模型有

- 1. 线性模型: $Y = \beta_0 + \beta_1 X + e$.
- 2. 二次的: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$.
- 3. 指数: $log(Y) = log(\beta_0) + \beta_1 X + e$.
- 4. log-log: $log(Y) = \beta_0 + \beta_1 log(X) + e$.

交叉验证Cross Validation

四种模型的参数估计程序如下:

见Matlab程序

事实上, 用最小二乘法就可以求出上述模型的估计.

然后我们使用交叉验证来对每个模型的预测误差进行估计. 算法如下

1. 对$k = 1, \ldots, n$, 令(x_k, y_k)为检验样本, 使用其余样本对模型参数进行估计. 然后计算预测误差.

(a) 使用其余的样本对模型进行拟合.

(b) 计算预测值: $\hat{y}_k = \hat{\beta}_0 + \beta_1 x_k$.

(c) 计算预测误差: $e_k = y_k - \hat{y}_k$.

2. 计算均方预测误差: $\sigma_e^2 = \frac{1}{n} \sum_{i=1}^{n} e_k^2$.

计算程序

见Matlab程序

结果表明二次多项式回归的预测误差最小.