Exercise 1.

Bias of estimators: analytical derivation, Monte Carlo simulation and bootstrap simulation

G. Bontempi

Question

Let us consider a r.v. $\mathbf{z}$ such that $E[\mathbf{z}] = \mu$ and $\text{Var}[\mathbf{z}] = \sigma^2$. Suppose we want to estimate from i.i.d. dataset D_N the parameter $\theta = \mu^2 = (E[\mathbf{z}])^2$. Let us consider three estimators:

$$\hat{\theta}_1 = \left(\frac{\sum_{i=1}^N z_i}{N} \right)^2$$

$$\hat{\theta}_2 = \frac{\sum_{i=1}^N z_i^2}{N}$$

$$\hat{\theta}_3 = \frac{(\sum_{i=1}^N z_i)^2}{N}$$

- Are they unbiased?
- Compute analytically the bias.
- Verify the result by Monte Carlo simulation for different values of N.
- Estimate the bias by bootstrap.

Analytical derivation of bias

1st estimator

Since $\text{Cov}[\mathbf{z}_i, \mathbf{z}_j] = 0$ and $E[\mathbf{z}^2] = \mu^2 + \sigma^2$,

$$E[\hat{\theta}_1] = \frac{1}{N^2} E\left[\left(\sum_{i=1}^N \mathbf{z}_i \right)^2 \right] = \frac{1}{N^2} E\left[\sum_{i=1}^N \mathbf{z}_i^2 + 2 \sum_{i<j}^N \mathbf{z}_i \mathbf{z}_j \right] = = \frac{1}{N^2} (N\mu^2 + N\sigma^2 + N(N-1)\mu^2) = \mu^2 + \frac{\sigma^2}{N}$$

then the bias of the first estimator is $B_1 = E[\hat{\theta}_1] - \mu^2 = \frac{\sigma^2}{N}$.

2nd estimator

$$E[\hat{\theta}_2] = \frac{1}{N} E\left[\left(\sum_{i=1}^N \mathbf{z}_i^2 \right) \right] = \frac{N\mu^2 + N\sigma^2}{N} = \mu^2 + \sigma^2$$

then the bias of the second estimator is $B_2 = E[\hat{\theta}_2] - \mu^2 = \sigma^2$.

3rd estimator

$$E[\hat{\theta}_3] = NE[\hat{\theta}_1] = N\mu^2 + \sigma^2$$

then the bias of the thirs estimator is $B_3 = E[\hat{\theta}_3] - \mu^2 = (N-1)\mu^2 + \sigma^2$.

The three estimators are biased.

Random variable distribution

```
rm(list=ls())

muz=2
sdz=1

N=100  ## number of samples

## Analytical results (see above)
anB1=sdz^2/N
anB2=(sdz^2)
anB3=(sdz^2)+(N-1)*muz^2
```

Monte Carlo simulation

We need to make an hypothesis about the $\mathbf{z}$ distribution if we want to simulate sample generation. We assume here the $\mathbf{z} \sim N(\mu, \sigma^2)$ is Normal.

```
S=10000 ## number of Monte Carlo trials

muhat2.1=NULL
muhat2.2=NULL
muhat2.3=NULL

for (s in 1:S){
  DN=rnorm(N,muz,sd=sdz)
  muhat2.1=c(muhat2.1,mean(DN)^2)
  muhat2.2=c(muhat2.2,sum(DN^2)/N)
  muhat2.3=c(muhat2.3,sum(DN)^2/N)
}

mcB1= mean(muhat2.1)-muz^2
mcB2= mean(muhat2.2)-muz^2
mcB3= mean(muhat2.3)-muz^2
```

Bootstrap estimation

Let us first note that only the first estimator is a plug-in estimator of $(E[\mathbf{z}])^2$. This is then the one that should be used to estimate the gap

$$\text{Bias}_{bs} = \frac{\sum_{b=1}^{B} \theta_{(b)}}{B} - \hat{\theta}_1$$

for all the three estimators.

```r
B=10000
muhat2.1=mean(DN)^2 ## plug-in estimator
muhat2.2=sum(DN^2)/N
muhat2.3=sum(DN)^2/N
muhatb=NULL
muhatb2=NULL
muhatb3=NULL
for (b in 1:B){
  Ib=sample(N,rep=TRUE)
  Db=DN[Ib]
  muhatb=c(muhatb,(mean(Db)^2))
  muhatb2=c(muhatb2,sum(Db^2)/N)
  muhatb3=c(muhatb3,sum(Db)^2/N)
}

bsB1=mean(muhatb)-muhat2.1
bsB2=mean(muhatb2)-muhat2.1
bsB3=mean(muhatb3)-muhat2.1
```

Final check

```r
cat("anB1=",anB1, "mcB1=", mcB1, "bsB1=", bsB1, "\n",
     "anB2=",anB2, "mcB2=", mcB2, "bsB2=", bsB2, "\n",
    "anB3=",anB3, "mcB3=", mcB3, "bsB3=", bsB3, "\n")
```

```
## anB1= 0.01 mcB1= 0.008180712 bsB1= 0.0004705264
##  anB2= 1 mcB2= 0.997601 bsB2= 0.9001592
##  anB3= 397 mcB3= 396.8181 bsB3= 407.4583
```

Try for different values of μ, σ^2, N, B and S.