Model Misspecification

Zheng Yuan

2019/10/21

LOOBIC

```
loocv = function(fit) {

  h = lm.influence(fit)$h
  mean((residuals(fit)/(1-h))^2)

}

cvbic = function(fit) {

  dim(fit$model)[1]*loocv(fit)+(fit$rank)*log(dim(fit$model)[1])

}
```

Traditional BIC

```
mse <- function(object) {

  mean(residuals(object)^2)
}

bic = function(fit) {
  dim(fit$model)[1]*mse(fit)+(fit$rank)*log(dim(fit$model)[1])
}
```

Prediction Error

```
prederror<-function (Object) {
  mean((dt1$y1-unlist(predict(Object,dt1)))^2)
}
```

Define data generator

The true model is non linear: $y = (2x_0 + 9x_1^2 + 4x_3)/\sqrt{n}$, where $n = 1000$, that is, the coefficients ar e sclaed by the square root of sample size.

```
def <- defData(varname = "x0", dist = "nonrandom", formula = 1)%>%

defData(,varname = "x1", dist="uniform",formula = "10;20")%>%

defData(,varname = "x2", dist="uniform",formula = "0;3")%>%
```

```
defData(,varname = "x3", dist="uniform",formula = "0;5")%>%

defData(,varname = "x4", dist="uniform",formula = "5;10")%>%

defData(, varname = "y1", formula = "(2*x0+9*x1^2+4*x3)/sqrt(1000)", variance = 1)
```

Test dataset with 1000 samples

```
dt1 <- genData(1000, def)
```

Options For The Models

```
dt <- genData(1000, def)

fit1 <- lm(y1 ~ x3, data = dt)
fit2 <- lm(y1 ~ x2+x3, data = dt)
fit3 <- lm(y1 ~ x1+x3, data = dt)
fit4 <- lm(y1 ~ x4+x3, data = dt)
fit5 <- lm(y1 ~ x1+x3+x4, data = dt)
fit6 <- lm(y1 ~ x2+x3+x4, data = dt)
fit7 <- lm(y1 ~ x1+x2+x3, data = dt)
fit8 <- lm(y1 ~ x1+x2+x3+x4, data = dt)

models = list(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8)
```

LOOBIC VS BIC

```
md1<-c()

md2<-c()

bestmodel<-c()

bestpe<-c()

md<-c()

for(i in 1:1000){

dt <- genData(1000, def)
dt1 <- genData(1000, def)## training and test set

fit1 <- lm(y1 ~ x3, data = dt)
fit2 <- lm(y1 ~ x2+x3, data = dt)
fit3 <- lm(y1 ~ x1+x3, data = dt)
fit4 <- lm(y1 ~ x4+x3, data = dt)
fit5 <- lm(y1 ~ x1+x3+x4, data = dt)
fit6 <- lm(y1 ~ x2+x3+x4, data = dt)
fit7 <- lm(y1 ~ x1+x2+x3, data = dt)
```

```r
fit8 <- lm(y1 ~ x1+x2+x3+x4, data = dt)

models = list (fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8)## model pools

pe <-c(prederror(fit1),prederror(fit2),prederror(fit3),prederror(fit4),prederror(fit5),prederror(fit6),

bestpe<-c(bestpe,pe[which.min(pe)])## prediction error of best model

bestmodel<-c(bestmodel,which.min(pe))

md1<-c(md1,which.min(c(cvbic(fit1),cvbic(fit2),cvbic(fit3),cvbic(fit4),cvbic(fit5),cvbic(fit6),cvbic(fi

md2<-c(md2,which.min(c(bic(fit1),bic(fit2),bic(fit3),bic(fit4),bic(fit5),bic(fit6),bic(fit7),bic(fit8))

md<-c(md,pe[md1[i]]-pe[md2[i]]) ## difference between the prediction error select
}

table(bestmodel)/1000
```

```
## bestmodel
##     3     5     7     8
## 0.410 0.219 0.251 0.120
```

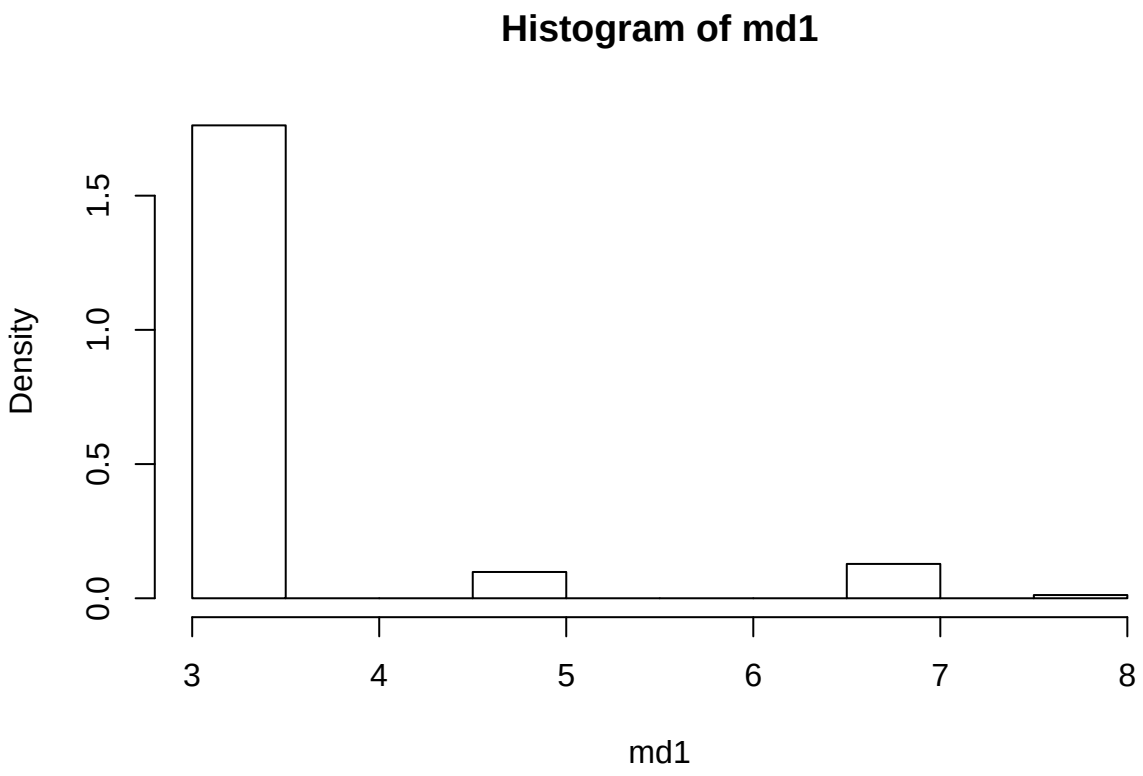
```r
table(md1)/1000## empirical probability for each model
```

```
## md1
##     3     5     7     8
## 0.881 0.049 0.064 0.006
```

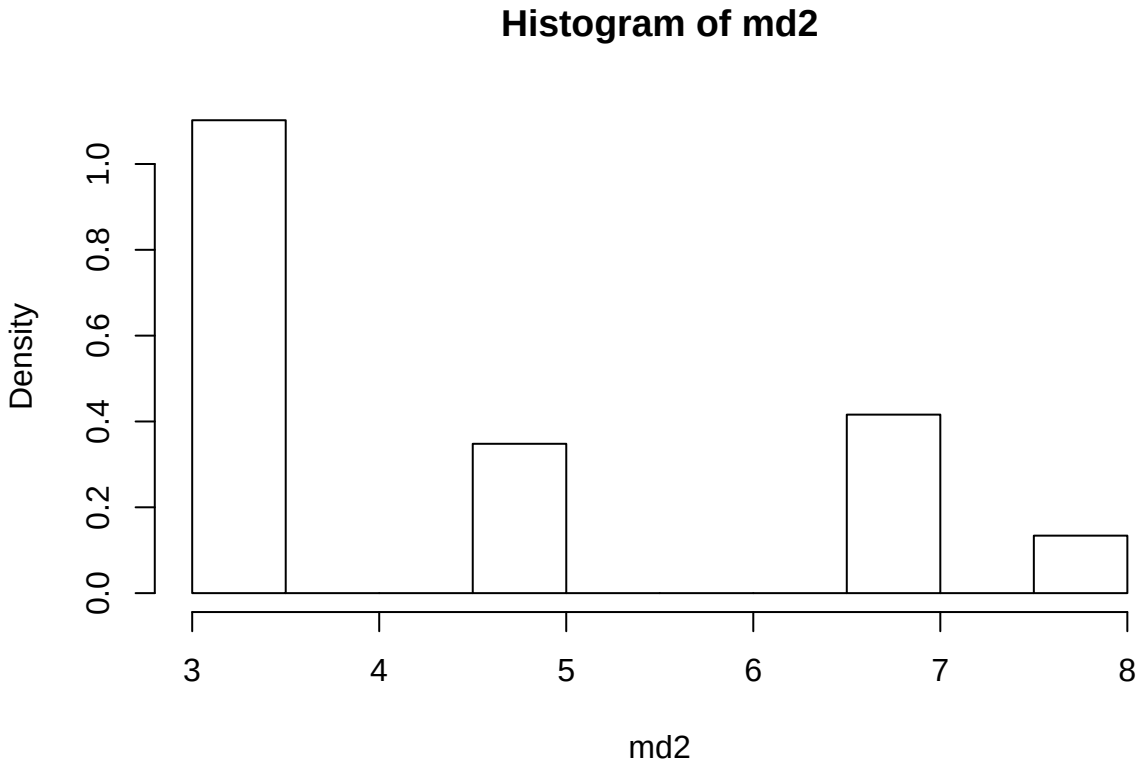
```r
table(md2)/1000
```

```
## md2
##     3     5     7     8
## 0.551 0.174 0.208 0.067
```

```r
hist(md1, probability=TRUE)
```




```
hist(md2, probability=TRUE)
```



It turns out that LOOBIC tends to select the best model in prediction while traditional BIC performs not stably compares with LOOBIC.

In each time of simulation, check if LOOBIC and BIC chooses the best model or not

```
modeltable = data.frame(bestmodel,md1,md2,bestpe,md)%>%mutate(loobic = ifelse(bestmodel==md1,1,0))%>%mu

modeltable[1:10,]
```

```
##    bestmodel md1 md2   bestpe          md loobic bic
## 1          3   3   5 5.761663 -0.02522537      1   0
## 2          7   3   3 5.872231  0.00000000      0   0
## 3          7   3   3 5.582637  0.00000000      0   0
## 4          7   5   5 6.011964  0.00000000      0   0
## 5          3   3   3 5.463034  0.00000000      1   1
## 6          8   3   3 5.210667  0.00000000      0   0
## 7          3   3   8 5.463783 -0.03622317      1   0
## 8          3   3   3 5.561057  0.00000000      1   1
## 9          3   3   7 5.538290 -0.02490965      1   0
## 10         3   3   3 5.776175  0.00000000      1   1
```

```
mean(modeltable$loobic)
```

```
## [1] 0.361
```

```
mean(modeltable$bic)
```

```
## [1] 0.269
```

Interpretation: in this case, LOOBIC selects best model 33.7% out of 1000 times, while BIC selects best model 24.3% out of 1000 times, besides, the difference between the prediction errors of the models selected by LOOBIC and BIC is actually small.