The third report

Zheng Yuan

2019/5/18

```r
library(simstudy)
```

```
## Loading required package: data.table
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

Generate Dataset

```r
def <- defData(varname = "mu1", dist = "nonrandom", formula = 5, id = "idnum")
def <- defData(def,varname = "mu2", dist = "nonrandom", formula = 2, id = "idnum")
def <- defData(def,varname = "mu3", dist = "nonrandom", formula = 3, id = "idnum")
def <- defData(def,varname = "mu4", dist = "nonrandom", formula = 4, id = "idnum")
def <- defData(def,varname = "x0", dist = "nonrandom", formula = 1, id = "idnum")
def <- defData(def,varname = "x1", formula = "mu1", variance=1)
def <- defData(def,varname = "x2", formula = "mu2", variance=1)
def <- defData(def,varname = "x3", formula = "mu3", variance=1)
def <- defData(def,varname = "x4", formula = "mu4", variance=1)
def <- defData(def, varname = "y1", formula = "2*x0+4*x3", variance = 1)
def <- defData(def, varname = "y2", formula = "2*x0+4*x3+8*x4", variance = 1)
def <- defData(def, varname = "y3", formula = "2*x0+9*x1+4*x3+8*x4", variance = 1)
def <- defData(def, varname = "y4", formula = "2*x0+9*x1+6*x2+4*x3+8*x4", variance = 1)
dt <- genData(40, def)
dt<-dt%>%select(y1,y2,y3,y4,x0,x1,x2,x3,x4)
```

This simulation shows that leave-one-out cross validation at least works really well on model selection in this case, in fact, it always chooses the true model (100 out of 100 times) no matter what the true model is. Maybe the reason is that our dataset has only 40 observations. But it kind of doesn't agree with the simlulation result in the paper.

leave-one-out cross

####sample size=1500

```
loocv=function(fit){
  h=lm.influence(fit)$h
  mean((residuals(fit)/(1-h))^2)
}
```

```
md<-c()
for(i in 1:100){
dt <- genData(1500, def)
fit1 <- lm(y1 ~ x3, data = dt)

fit2 <- lm(y1 ~ x1+x3, data = dt)

fit3 <- lm(y1 ~ x2+x3, data = dt)

fit4 <- lm(y1 ~ x3+x4, data = dt)

fit5 <- lm(y1 ~ x1+x2+x3, data = dt)

fit6 <- lm(y1 ~ x1+x3+x4, data = dt)

fit7 <- lm(y1 ~ x2+x3+x4, data = dt)

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

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4),loocv(fit5),loocv(fit6),loocv(fit7),loocv(fit8))

md<-c(md,which.min(a))
}
table(md)
```

```
## md
##  1  2  3  4  5  6  7
## 75  8  6  4  1  2  4
```

```
md<-c()
for(i in 1:100){
  dt <- genData(1500, def)
fit1 <- lm(y2 ~ x3+x4, data = dt )

fit2 <- lm(y2 ~ x1+x3+x4, data = dt)

fit3 <- lm(y2 ~ x2+x3+x4, data = dt)

fit4 <- lm(y2 ~ x1+x2+x3+x4, data = dt)

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4))

md<-c(md,which.min(a))
}
table(md)
```

```
## md
##   1  2  3  4
## 74 11 14  1
```

```r
md<-c()
for(i in 1:100){
  dt <- genData(1500, def)
fit1 <- lm(y3 ~ x3+x4, data = dt )

fit2 <- lm(y3 ~ x1+x3+x4, data = dt)

fit3 <- lm(y3 ~ x2+x3+x4, data = dt)

fit4 <- lm(y3 ~ x1+x2+x3+x4, data = dt)

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4))

md<-c(md,which.min(a))
}
table(md)
```

```
## md
##  2  4
## 83 17
```

```r
md<-c()
for(i in 1:100){
  dt <- genData(1500, def)
fit1 <- lm(y4 ~ x1+x2+x4, data = dt )

fit2 <- lm(y4 ~ x1+x3+x4, data = dt)

fit3 <- lm(y4 ~ x2+x3+x4, data = dt)

fit4 <- lm(y4 ~ x1+x2+x3+x4, data = dt)

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4))

md<-c(md,which.min(a))
}
table(md)
```

```
## md
##   4
## 100
```

####sample size=1000

```r
loocv=function(fit){
  h=lm.influence(fit)$h
  mean((residuals(fit)/(1-h))^2)
}
```

```r
md<-c()
for(i in 1:100){
dt <- genData(1000, def)
fit1 <- lm(y1 ~ x3, data = dt)
fit2 <- lm(y1 ~ x1+x3, data = dt)

fit3 <- lm(y1 ~ x2+x3, data = dt)

fit4 <- lm(y1 ~ x3+x4, data = dt)

fit5 <- lm(y1 ~ x1+x2+x3, data = dt)

fit6 <- lm(y1 ~ x1+x3+x4, data = dt)

fit7 <- lm(y1 ~ x2+x3+x4, data = dt)

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

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4),loocv(fit5),loocv(fit6),loocv(fit7),loocv(fit8))

md<-c(md,which.min(a))
}
table(md)
```

```
## md
##  1  2  3  4  5  7
## 59  7 14 14  3  3
```

```
md<-c()
for(i in 1:100){
  dt <- genData(1000, def)
fit1 <- lm(y2 ~ x3+x4, data = dt )

fit2 <- lm(y2 ~ x1+x3+x4, data = dt)

fit3 <- lm(y2 ~ x2+x3+x4, data = dt)

fit4 <- lm(y2 ~ x1+x2+x3+x4, data = dt)

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4))

md<-c(md,which.min(a))
}
table(md)
```

```
## md
##  1  2  3  4
## 69 17 11  3
```

```
md<-c()
for(i in 1:100){
  dt <- genData(1000, def)
fit1 <- lm(y3 ~ x3+x4, data = dt )

fit2 <- lm(y3 ~ x1+x3+x4, data = dt)

fit3 <- lm(y3 ~ x2+x3+x4, data = dt)

fit4 <- lm(y3 ~ x1+x2+x3+x4, data = dt)
a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4))

md<-c(md,which.min(a))
}
table(md)
```

```
## md
##  2  4
## 84 16
```

```
md<-c()
for(i in 1:100){
  dt <- genData(1500, def)
fit1 <- lm(y4 ~ x1+x2+x4, data = dt )

fit2 <- lm(y4 ~ x1+x3+x4, data = dt)

fit3 <- lm(y4 ~ x2+x3+x4, data = dt)

fit4 <- lm(y4 ~ x1+x2+x3+x4, data = dt)

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4))

md<-c(md,which.min(a))
}
table(md)
```

```
## md
##   4
## 100
```

sample size=40 1000 simulations

```
loocv=function(fit){
  h=lm.influence(fit)$h
  mean((residuals(fit)/(1-h))^2)
}
```

```
md1<-c()
for(i in 1:1000){
dt <- genData(40, def)
fit1 <- lm(y1 ~ x3, data = dt)

fit2 <- lm(y1 ~ x1+x3, data = dt)

fit3 <- lm(y1 ~ x2+x3, data = dt)

fit4 <- lm(y1 ~ x3+x4, data = dt)

fit5 <- lm(y1 ~ x1+x2+x3, data = dt)

fit6 <- lm(y1 ~ x1+x3+x4, data = dt)

fit7 <- lm(y1 ~ x2+x3+x4, data = dt)
fit8 <- lm(y1 ~ x1+x2+x3+x4, data = dt)

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4),loocv(fit5),loocv(fit6),loocv(fit7),loocv(fit8))


md1<-c(md1,which.min(a))
}
table(md1)
```

```
## md1
##   1   2   3   4   5   6   7   8
## 589 111 114 110  17  22  31   6
```

```
md2<-c()
for(i in 1:1000){
  dt <- genData(40, def)
fit1 <- lm(y2 ~ x3+x4, data = dt )

fit2 <- lm(y2 ~ x1+x3+x4, data = dt)

fit3 <- lm(y2 ~ x2+x3+x4, data = dt)

fit4 <- lm(y2 ~ x1+x2+x3+x4, data = dt)

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4))

md2<-c(md2,which.min(a))
}
table(md2)
```

```
## md2
##   1   2   3   4
## 691 142 143  24
```

```
md3<-c()
for(i in 1:1000){
  dt <- genData(40, def)
fit1 <- lm(y3 ~ x3+x4, data = dt )

fit2 <- lm(y3 ~ x1+x3+x4, data = dt)

fit3 <- lm(y3 ~ x2+x3+x4, data = dt)

fit4 <- lm(y3 ~ x1+x2+x3+x4, data = dt)

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4))

md3<-c(md3,which.min(a))
}
table(md3)
```

```
## md3
##   2   4
## 803 197
```

```
md4<-c()
for(i in 1:1000){
  dt <- genData(40, def)
fit1 <- lm(y4 ~ x1+x2+x4, data = dt )

fit2 <- lm(y4 ~ x1+x3+x4, data = dt)

fit3 <- lm(y4 ~ x2+x3+x4, data = dt)

fit4 <- lm(y4 ~ x1+x2+x3+x4, data = dt)

a<-c(loocv(fit1),loocv(fit2),loocv(fit3),loocv(fit4))

md4<-c(md4,which.min(a))
}
table(md4)
```

```
## md4
##    4
## 1000
```

```
table(md1)
```

```
## md1
##   1   2   3   4   5   6   7   8
## 589 111 114 110  17  22  31   6
```

```
table(md2)
```

```
## md2
##   1   2   3   4
## 691 142 143  24
```

```
table(md3)
```

```
## md3
##   2   4
## 803 197
```

```
table(md4)
```

```
## md4
##    4
## 1000
```

####MCCV

```
fitControl <-
  trainControl(
  method = "LGOCV",
  p = 0.375
)
```

```
md1<-c()
for(i in 1:100){
dt <- genData(40, def)
fit1 <- train(y1 ~ x3, data = dt,
              method="lm",
```

```r
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit2 <- train(y1 ~ x1+x3, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit3 <- train(y1 ~ x2+x3, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit4 <- train(y1 ~ x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit5 <- train(y1 ~ x1+x2+x3, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit6 <- train(y1 ~ x1+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit7 <- train(y1 ~ x2+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit8 <- train(y1 ~ x1+x2+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

a<-c(fit1$results$RMSE,fit2$results$RMSE,fit3$results$RMSE,
     fit4$results$RMSE,fit5$results$RMSE,fit6$results$RMSE,
     fit7$results$RMSE,fit8$results$RMSE)


md1<-c(md1,which.min(a))
}
table(md1)
```

```
## md1
##  1  2  3  4  7
## 66 12 11  9  2
```

```r
md2<-c()
for(i in 1:100){
dt <- genData(40, def)
fit1 <- train(y2 ~ x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit2 <- train(y2 ~ x1+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))
fit3 <- train(y2 ~ x2+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit4 <- train(y2 ~ x1+x2+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

a<-c(fit1$results$RMSE,fit2$results$RMSE,fit3$results$RMSE,
     fit4$results$RMSE)


md2<-c(md2,which.min(a))
}
table(md2)
```

```
## md2
##  1  2  3  4
## 70 13 13  4
```

```
md3<-c()
for(i in 1:100){
dt <- genData(40, def)
fit1 <- train(y3 ~ x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit2 <- train(y3 ~ x1+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit3 <- train(y3 ~ x2+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit4 <- train(y3 ~ x1+x2+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

a<-c(fit1$results$RMSE,fit2$results$RMSE,fit3$results$RMSE,
     fit4$results$RMSE)


md3<-c(md3,which.min(a))
}
table(md3)
```

```
## md3
##  2  4
## 87 13
```

```
md4<-c()
for(i in 1:100){
dt <- genData(40, def)
fit1 <- train(y4 ~ x1+x2+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit2 <- train(y4 ~ x1+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit3 <- train(y4 ~ x2+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

fit4 <- train(y4 ~ x1+x2+x3+x4, data = dt,
              method="lm",
              trControl = trainControl(method = "LGOCV",p = 0.375))

a<-c(fit1$results$RMSE,fit2$results$RMSE,fit3$results$RMSE,
     fit4$results$RMSE)


md4<-c(md4,which.min(a))
}
table(md4)
```

```
## md4
##   4
## 100
```

```
table(md1)
```

```
## md1
##  1  2  3  4  7
## 66 12 11  9  2
```

```
table(md2)
```

```
## md2
##  1  2  3  4
## 70 13 13  4
```

```
table(md3)
```

```
## md3
##  2  4
## 87 13
```

```
table(md4)
```

```
## md4
##   4
## 100
```