Experiment3.1

Eliyahu Addess

2022-12-12

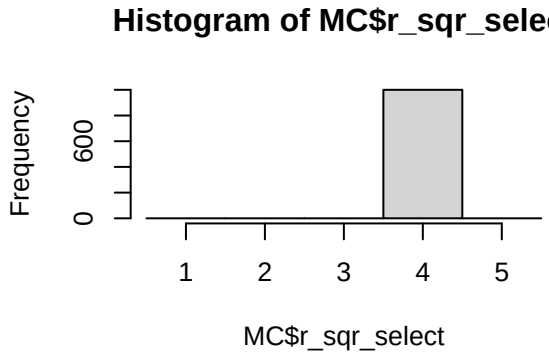
Now we'll try a a more precise but parsimonious model. Instead of first 6 variables, we'll use first 8.

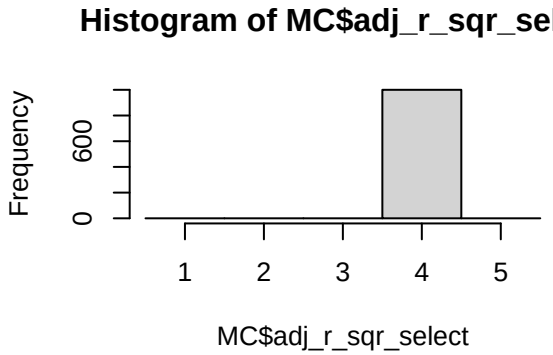
```
# load libraries
library(tidyverse)
library(future.apply) #parallel processing
library(tictoc) #timing code
library(knitr) #tables

#load functions
source("ProjectFunctions.R")

# set up parallel processing
plan(multisession)

# set seed for reproducibility
set.seed(42)

# load data
data <- read.csv("data_generic.csv")
```

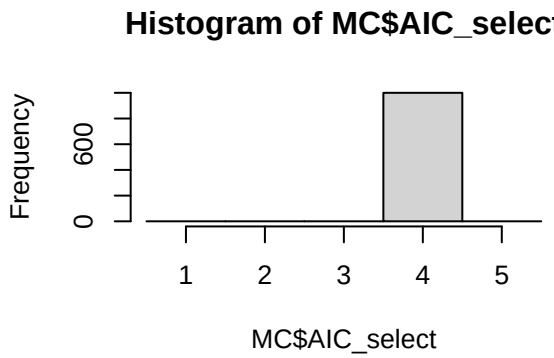
```
# fit base model
mod1 <- lm(y~x1+ x2 + x3 +x4 +x5 +x6 +x7 +x8 +x9, data = data)
coefs1 <- unname(mod1$coefficients)
```

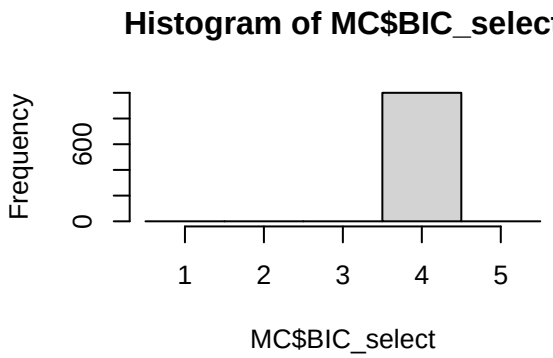
```
# simulation
MC3.1 <- as.data.frame(t(future_replicate(1000,
                                          experiment3(sd = 10000, data = data),
                                          future.seed = TRUE)))
```

```
#results
histo(MC3.1,title =  "MC3.1")
```









MC3.1

```
kable(props(MC3.1))
```

models	r_sqr_prop	adj_r_sqr_prop	AIC_prop	BIC_prop
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	1	1	1	1
5	0	0	0	0

We see that the results are the same as Experiment2.1. This is interesting because in light of Sharma et al.(2019) we would think that AIC and BIC would select the parsimonious model over the over-fit model.

My conjecture is that this parsimonious model is still an "under-fit" model in the sense that it leaves out an important predictor, namely x9. Examination of the correlation plot in the base case reveals a relatively strong correlation between x9 and y_sim. Perhaps even though the over-fit model included covariates that were not in the data generating process, since it included the important variable x9, it was selected as a better model than the parsimonious approximation.