STAT-COMP 2020 Final Exam

12/14/2020

The following code simulates the data that you will use to answer all the questions of this exam.

Be sure to simulate data using the seed specified in the script.

If the code is run properly, the sample-mean of y should be the same as the one printed below.

```
set.seed(12345)
p=2^5
n=150

X=matrix(nrow=n,ncol=p)
tmp=0.8
X[,1]=rnorm(n)
for(i in 2:p){ X[,i]=X[,i-1]*tmp+rnorm(n,sd=sqrt(1-tmp^2)) }

# True effects
 beta=rep(0,p)
 beta[c(5,15,14,25)]=1

 signal=X%*%beta

 vE=var(signal)/2
 error= rnorm(n=n,sd=sqrt(vE))
 y=123+error+signal

 mean(y)
```

```
## [1] 123.0843
```

1) Split the data into a training and a testing set

- Use observations in rows 1 to 100 for the training set and
- Observations in rows 101 to 150 for the testing set.

```
XTRN=X[1:100,]
yTRN=y[1:100]

XTST=X[101:150,]
yTST=y[101:150]
```

Fill into Table 1 of the word document the man value of y in the training and testing sets

```
kable(c('mean-y training'=round(mean(yTRN),3),'mean-y testing'=round(mean(yTST),3)),caption='Table 1')
```

Table 1: Table 1

	x
mean-y training	122.955
mean-y testing	123.344

For questions 2-4: Use only the training set for model fitting.

In question 4 use the testing set to select lambda.

2) Marginal association test

- Conduct marginal association analysis by regressing the outcome (y), on each of the predictors, one predictor at a time, store the p-values.
- To determine significance, consider three criteria:
  - Raw p-values < 0.05,
  - Bonferroni-adjusted p-values < 0.05 , and
  - FDR-adjusted p-values <0.05.

```
SMR=matrix(nrow=ncol(X),ncol=4)

for(i in 1:ncol(X)){
   SMR[i,]=summary(lm(yTRN~XTRN[,i]))$coef[2,]
}
```

```
# Without adjusting
 sig_raw<- SMR[,4] < 0.05

# Bonferroni -adjusted p-values
 sig_bonf<- p.adjust(SMR[,4],method='bonferroni')<0.05

# FDR-adjusted
 sig_fdr<-p.adjust(SMR[,4],method='fdr')<0.05
```

Complete Table 2 in the word document

- Start by completing the total number of discoveries (first column)
- Then separate those into true and false discoveries (last two columns)

```
 TMP=rbind(
     'Raw'=c('# of discoveries'=sum(sig_raw),
             'True discoveries' =sum((sig_raw)&(beta!=0)),
             'False discoveries'=sum((sig_raw)&(beta==0))
             ),

     'Bonferroni'=c('# of discoveries'=sum(sig_bonf),
                    'True discoveries' =sum((sig_bonf)&(beta!=0)),
                    'False discoveries'=sum((sig_bonf)&(beta==0))
               ),

     'FDR'=c('# of discoveries'=sum(sig_fdr),
             'True discoveries' =sum((sig_fdr)&(beta!=0)),
             'False discoveries'=sum((sig_fdr)&(beta==0))
             )
     )
```

```r
kable(TMP,caption='Table 2')
```

Table 2: Table 2

	# of discoveries	True discoveries	False discoveries
Raw	23	4	19
Bonferroni	14	3	11
FDR	21	4	17

3) Conditional association test using Ordinary least squares

- Obtain p-values by regressing the response on all the predictors together.
- To determine significance, consider same criteria used in Question 2, that is:
  - Raw p-values < 0.05,
  - Bonferroni-adjusted p-values < 0.05 , and
  - FDR-adjusted p-values <0.05.

```r
fmOLS=lm(yTRN~XTRN)
summary(fmOLS) # this was not requested, I am just printing it to check my resutls in table 3
```

```
##
## Call:
## lm(formula = yTRN ~ XTRN)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5785 -1.1396  0.0133  1.2882  3.2653
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 123.05620    0.23691 519.426  < 2e-16 ***
## XTRN1        -0.36031    0.45311  -0.795  0.42931
## XTRN2         0.75674    0.52903   1.430  0.15724
## XTRN3        -0.73496    0.52482  -1.400  0.16601
## XTRN4         0.35808    0.46069   0.777  0.43973
## XTRN5         0.92814    0.51434   1.805  0.07564 .
## XTRN6        -0.11213    0.49733  -0.225  0.82231
## XTRN7         0.63568    0.51735   1.229  0.22348
## XTRN8        -0.66878    0.50295  -1.330  0.18812
## XTRN9         0.78220    0.51090   1.531  0.13047
## XTRN10       -0.83884    0.46887  -1.789  0.07812 .
## XTRN11        0.69300    0.47593   1.456  0.15004
## XTRN12       -0.02399    0.52930  -0.045  0.96398
## XTRN13       -0.03377    0.50888  -0.066  0.94729
## XTRN14        1.48231    0.48156   3.078  0.00302 **
## XTRN15        0.95626    0.47478   2.014  0.04801 *
## XTRN16       -0.25109    0.51351  -0.489  0.62646
## XTRN17       -0.58264    0.53017  -1.099  0.27571
## XTRN18        0.06698    0.55520   0.121  0.90433
## XTRN19       -0.15435    0.47361  -0.326  0.74551
## XTRN20        0.37429    0.48774   0.767  0.44554
## XTRN21       -0.40212    0.47793  -0.841  0.40313
```

```
## XTRN22       -0.10851    0.53786  -0.202  0.84073
## XTRN23        0.76885    0.64796   1.187  0.23959
## XTRN24        0.01098    0.47210   0.023  0.98152
## XTRN25        0.61383    0.51023   1.203  0.23320
## XTRN26        0.32245    0.57368   0.562  0.57595
## XTRN27       -0.12344    0.47813  -0.258  0.79707
## XTRN28       -0.34038    0.54207  -0.628  0.53218
## XTRN29       -0.14308    0.53778  -0.266  0.79101
## XTRN30       -0.23539    0.42095  -0.559  0.57790
## XTRN31       -0.54776    0.46139  -1.187  0.23935
## XTRN32        0.81571    0.45810   1.781  0.07951 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.972 on 67 degrees of freedom
## Multiple R-squared:  0.7526, Adjusted R-squared:  0.6344
## F-statistic: 6.368 on 32 and 67 DF,  p-value: 1.033e-10
```

Complete Table 3 in the word document

- Start by completing the total number of discoveries (first column)
- Then separate those into true and false discoveries (last two columns)

```
pVals=summary(fmOLS)$coef[-1,4] # remove resutls for interecept because we are not testing H0: mu=0

# Without adjusting
sig_raw<- pVals < 0.05

# Bonferroni -adjusted p-values
sig_bonf<- p.adjust(pVals,method='bonferroni')<0.05

# FDR-adjusted
sig_fdr<-p.adjust(pVals,method='fdr')<0.05


TMP=rbind(
    'Raw'=c('# of discoveries'=sum(sig_raw),
            'True discoveries' =sum((sig_raw)&(beta!=0)),
            'False discoveries'=sum((sig_raw)&(beta==0))
            ),

    'Bonferroni'=c( '# of discoveries'=sum(sig_bonf),
                    'True discoveries' =sum((sig_bonf)&(beta!=0)),
                    'False discoveries'=sum((sig_bonf)&(beta==0))
                ),

    'FDR'=c('# of discoveries'=sum(sig_fdr),
            'True discoveries' =sum((sig_fdr)&(beta!=0)),
            'False discoveries'=sum((sig_fdr)&(beta==0))
            )
    )

kable(TMP,caption='Table 3')
```

Table 3: Table 3

	# of discoveries	True discoveries	False discoveries
Raw	2	2	0
Bonferroni	0	0	0
FDR	0	0	0

4) Lasso

- Fit a Lasso regression to the training data.
- **Produce a plot with correlations between predictions and observations in the testing set, versus log-lambda**

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.0-2
```

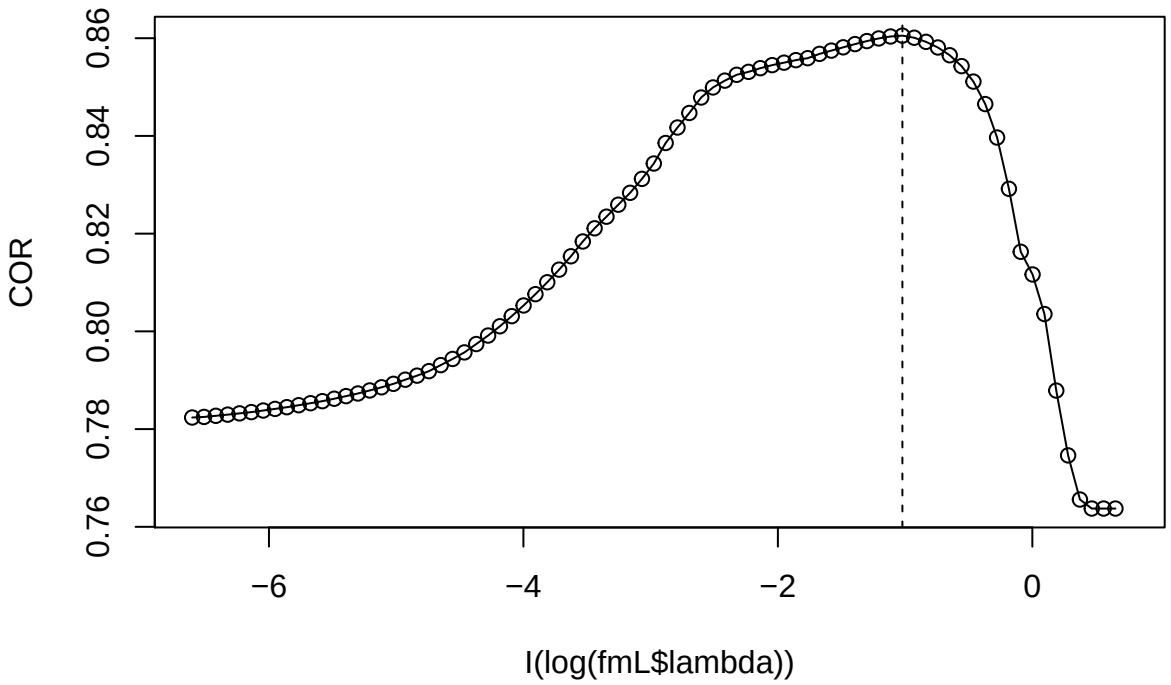
```
fmL=glmnet(y=yTRN,x=scale(XTRN))

yHatTST=XTST%*%fmL$beta

COR=rep(NA,ncol(yHatTST))
for(i in 1:ncol(yHatTST)){ COR[i]=cor(yTST,yHatTST[,i]) }
```

```
## Warning in cor(yTST, yHatTST[, i]): the standard deviation is zero
```

```
plot(COR~I(log(fmL$lambda)),type='o')
lambda=fmL$lambda[which.max(COR)]
abline(v=log(lambda),lty=2)
```



```
  print(lambda)
```

```
## [1] 0.3600872
```

Complete Table 4 in the word document

- Start by entering the value of lambda that led to the highest prediction correlation.
- Then for that lambda report in columns 2-4 of Table 4:
  - Total number of discoveries
  - Total number of true discoveries
  - Total number of false discoveries

```
tmp=which.max(COR)
TMP= c('lambda'=lambda,
        '# of discoveries'=sum(fmL$beta[,tmp]!=0),
        'True discoveies'=sum( (fmL$beta[,tmp]!=0)&(beta!=0)),
        'False discoveries'=sum( (fmL$beta[,tmp]!=0)&(beta==0))
      )
kable(TMP,caption='Table 4',digits=2)
```

Table 4: Table 4

	x
lambda	0.36
# of discoveries	8.00
True discoveies	4.00
False discoveries	4.00

5) What method do you recommend to use?

In the simulation there are 4 true non-zero effects (Ha's) and 28 true zero effects (H0). The marginal association tests detected all (raw-palues, and fdr-adjusted p-values) or almost all of the Ha's (Bonferroni), but the proportion of false discoveries (false discoveries/total discoveries) was very high. The mutliple regression model (Question 3) had a very low power. Finally, Lasso detected all the true effects with a false discovery proprotion of ~50%. In summary, in this case Lasso performed better than the other methods. But the performance of Lasso is still poor because 50% is a very high proportion of false discoveries.