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.

```r
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
```

Submit to D2L:

- Your script and outputs (Rmd, html, pdf, or just your script)
- A word document with your answers, template available here.
- In question 4 you need to report a plot. You can copy-and-paste it in the place for Question 4 in the word document, submit a sparate file, or include it in your Markdown file.

Question 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.
- **Fill Table 1 of the word document** with the man value of y in the training and testing sets

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

In question 4 use the testing set to select lambda.

Question 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.
- **Complete Table 2 in the word document**:
  - Start by completing the total number of discoveries (first column)
  - Then, discriminate those into true and false discoveries (last two columns)

Question 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.
- **Complete Table 3 in the word document**:
  - Start by completing the total number of discoveries (first column)
  - Then, discriminate those into true and false discoveries (last two columns)

Question 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** (copy-and-paste the plot in the Word document, or submit it as either a separate file or as part of your markdown file).

Complete Table 4 in the word document

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

Question 5: What method do you recommend to use?

Enter your answer (no more than on paragraph 5-6 lines) in the word document.