STT 843 Final Project Spring 2018

Due Date: Monday before 5:00pm EST, April 23, in C418 Wells Hall. **Late papers will not be accepted.**

Instructions: You must work independently. You may not discuss this project with anyone except the course instructor.

The project aims to analyze a real data set using the multivariate analysis methods introduced in our course. You are free to perform any analyses that you consider to be relevant and informative. You may use any software to perform calculations and computations, but R is preferable. Do not submit pages of raw computer output, but you need include important tables or attach graphs to your report. Remember that there is no absolutely best answer and I expect to receive many different answers. You will be graded with respect to how well your analyses reveal interesting aspects of the data set, the interpretation of your results, how well you justify your methods of analysis, and the overall quality of your writing.

Your report should not exceed five typed pages, excluding attached graphs or tables, with the font size and spacing used in these instructions. You may cut out computer output and simply paste it in appropriate places in your report. You do not have to type your report. You can submit a handwritten report. The limit on the number of handwritten pages depends on how large you write, but is limited to what could have fit on five typed pages. Your report should contain the following:

1. Provide a one paragraph summary of your major findings. This should not contain any formulas or mathematical symbols. It should be well written so that it could be easily understood by anybody else who is not a statistician.

2. Provide a detailed description of your model and analysis. Discuss and interpret any important features of your model and state you conclusions in the context of the problem.

3. You may submit one more paragraph outlining additional analyses that you would have done if you had more time. You will earn points for good suggestions and lose points for suggestions with little potential value. (This is optional.)

Data set description:

The data sets were obtained from a study conducted by Rosenwald et al. (2002). A total of 240 patients were selected retrospectively on the basis of the availability of tumor-biopsy specimens. All the patients had received anthracycline-based chemotherapy. The microarray gene expression were obtained from the biopsy samples of the 240 diffuse large-B-cell lymphoma. The gene expression data set can be downloaded from the class website under the name "Microarray Data (NEJM)". Together with the gene expression data, some clinical features, gene expression signatures, follow-up time and status during the follow-up are included in another file named “Patient Data (NEJM)”. Both data sets can be found at `http://www.stt.msu.edu/users/pszhong/STT-843-2018.html`. More detailed information about these two data sets can be found in the reference given at the end.

To combine these two data sets together, you can match the “DLBCL sample (LYM number)” in the “patient data” with the LYM number in column names of the “microarray data”. The following R code might be useful for combining two data sets and impute missing values in microarray data set. Note that the columns of resulting data set “newLYMexpr” are matched with the rows in the patient data (named “newresponse”). This means that the k-th column of the data “newLYMexpr” and the k-th row of the data “patient data” are from the same patient.

```
expr=read.csv(file="microarray_data_NEJM.csv",header=TRUE)
LYMexpr=expr[,1:276]
response=read.csv(file="Patient_data_NEJM.csv",header=TRUE)
newresponse=response[order(response[,1]),]

idvec=NULL
for (j in 3:(dim(LYMexpr)[2]))
{
idvec[j]=substr(unlist(strsplit(names(LYMexpr)[j],"[_]"))[2],4,6)
}
idvec[1:2]=names(LYMexpr)[1:2]
colnames(LYMexpr)=idvec
indicator=c(TRUE,TRUE,(as.numeric(idvec[3:276]))%in%response[,1])
LYMexpr4use=LYMexpr[,indicator]
newLYMexpr=LYMexpr4use[,c(1,2,order(names(LYMexpr4use)[3:242])+2)]
imp_newLYMexpr=knnImputation(newLYMexpr,k=2)
```

References:

Rosenwald, A., Wright, G., Wing, C., Connors, J., Campo, E. et al. (2002). The Use of Molecular Profiling to Predict Survival After Chemotherapy for Diffuse Large-B-Cell Lymphoma. The New England Journal of Medicine, 346(25), 1937-1947.