Stat 601 (2018 Fall) Final Project Guidelines

- About the data
  - The data set is for **normal hematopoiesis**
    - Novershtern N, Subramanian A, Lawton LN, Mak RH et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 2011 Jan 21;144(2):296-309. PMID: 21241896
  - The data has been formatted to fit the need of the class.
    - The data has 8968 genes (rows) and 185 cells and progenitors (columns)
  - The main response variables are not defined. We shall treat columns as samples in this final project. This project defines the following two response variables.
    - Using all genes in Figure 5, regress one gene on the rest of 36 genes. Report your best model, i.e., which one is the response variable.
    - Find PCA1 of all genes in Figure 5, treat PCA1 as a response variable.
  - Those 8968 genes are approximately evenly partitioned into 15 subsets. When PCA1 is used as a response variable, all genes in Figure 5 should be removed from your data set.
- About the models
  - Linear regression models
    - Try whatever models and methods you learned from Stat 601 to the data fitting. The final reported models shouldn't be more than three models for each response variable.
    - Carefully state your variable selection procedures and rules.
  - GMC variable selections
    - Consider the model
      Y=g(x1,x2,…,xp)+e
      With g(x1,x2,…,xp)=poly(b_1x_1+...+b_px_p, k) for k=2:5
      Maximize var(g(x))/(var(g(x))+var(e))-lambda1 |corr(g(x),e)|-lambda2(Lasso)
      For each response variable. Please note that making g(x) values large will give a large negative e value. As a result, |corr(g(x),e)| will be close to 1, while the first term is close to ½. Consider lambda1 to be in [1/2, 1], and lambda2>0.01.
    - Using provided R code to maximize
      GMC(Y|g(X))-lambda (lasso) with lambda>0.01
  - From the linear regression models, using the idea taught in class, you covert the response variables into dichotomized observations, i.e., 0 and 1, then fit three logistic regression models and compare your fitted parameter values with the fitted parameter values in your linear regression models.
- About the project report
  - The report must be a typed report. Submit an electronic copy to both TA Yuqing Xu and Professor Zhengjun Zhang on Dec. 19, 2018 (Wednesday) 07:05 PM.
  - The total length of the report should be within 15 pages, and the fonts should be no smaller than 11 points.

- The total length of main text body should be within the first 5 pages. Figures and tables can be placed on pages 6-15.
- You don't have to describe the biological issues related to the data.
- What are needed in the report:
  - Main findings: one paragraph or more
  - Sections of your analyses of the data sets, details are needed.
  - Limitations and remedies of analysis.
  - Future work

❖ About grading

Overall presentation will be graded up to 15 points.

Each data set will be analyzed by two different teams. For each data set, the best performance team gets 5 points, and the other team's score will be proportion to 5 points. The proportion will be subjected to how the results are reported.

Set	Rows	Set	Rows	Set	Rows
a	1 - 597	f	2991 - 3588	m	5980 - 6577
b	598 - 1196	g	3589 - 4186	n	6578 - 7175
c	1197 - 1794	h	4187 - 4784	o	7176 - 7773
d	1795 - 2392	i	4785 - 5381	p	7774 - 8371
e	2393 - 2990	k	5382 - 5979	s	8372 - 8969