Stat 602 (2017 Spring) Final Project Guidelines

- About the data
  - The data set is a medical clinic data with the following characteristics
    - Coded Patient IDs (in the first row)
    - 12042 Genes (in the first column in one sheet)
    - Yearstobirth
    - Vitalstatus (1 – death, 0 – censored)
    - Daystodeath
    - Daystolastfollowup
  - The data have been formatted to fit the need of the class.
  - The main response variable will be daystolastfollowup. If the value of the response variable is NA for a particular patient, the value of daystodeath is instead used. Total number of patients is 568. Another response variable is TP53.
  - Those 12042 genes are pre-selected into 16 subsets using a particularly designed sampling scheme. Each team will work on 2 subsets selected from Doodle poll. Each subset contains about 180 genes.
- About the models
  - Linear regression models
    - Try whatever models and methods you learned from Stat 602 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
    - Choose 5 functions with one being linear such that
      $Y=g(x1,x2,\ldots,xp)+e$
      Maximize $var(g(x))/(var(g(x))+var(e))$-lambda1 $|cov(g(x),e)|$-lambda2(Lasso)
      For each response variable.
    - Using provided R code to maximize
      $GMC(Y|g(X))$-lambda (lasso)
  - 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 a paper copy to TA Yuqing Xu at 10:05am on May 11, 2017. Submit an electronic copy to Professor Zhengjun Zhang by 10:05am on May 11, 2017.
  - 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.