Midterm Exam 2

Final code and report are due on **May 3** at **11pm** via the **Midterm 2** assignment in the course space. Late submissions will not be graded.

Save your files with your name (e.g. '*<Your-First-Name> <Your-Last-Name> Midterm2.sas*' and '*<Your-First-Name> <Your-Last-Name> Midterm2.pdf*' for final code and report). You need to submit your report as a pdf file.

Any "collaboration" on this exam is a violation of the student code of academic integrity and will be dealt with accordingly.

By accepting this exam, you agree that you:

- **will do the exam by yourself**
- **will not discuss any portion of the exam with anyone other than the instructor**
- **will abide by all aspects of the campus code of academic integrity as noted in the syllabus**

Use **ods** statements to obtain only the outputs you need. You don't need to use **ods select** statements, but do turn off results you don't need like we have done in class and homework. In the final report, write your comments/explanations of results close to the results so they are easy for a reader to follow.

Use confidence levels of .95 and significance levels of .05 unless the instructions state otherwise. You may only use the SAS help, materials in the course space, and the text book *A Handbook of Statistical Analyses using SAS*.

Data Sets:

For Exercises 1-4, use the **Cancer** data set defined in Midterm2Data.sas. **Cancer** data set in the **Midterm2Data.sas** file is based on the **biomarkers** data that's simulated by the instructor. The raw data is contained in **biomarkers.csv** in the course space. The data contains ten attributes of protein biomarkers. Our goal is to identify the cancerous population from the healthy population using the ten biomarkers. The variables are:

- **P8**
- **P14**
- **P19**
- **P33**
- **P37**
- **P49**
- **P55**
- **P64**
- **P70**
- **P80**
- **grp:** classification variable, 1 for healthy, 2 for cancerous
- **id:** patient identification number

Exercise 1

In this exercise, we'd like to investigate whether biomarkers' attributes can distinguish a patient's health status.

a) Perform a principal components analysis on the attributes of biomarkers (all variables except **grp** and **id**), determine how many components you would keep to retain at least 50% of the total variation from the original variables. Also comment on how many components would be chosen based on the average eigenvalue and scree plot methods.

b) Based on the total variation rule from exercise a, explain what features the retained components pick out of the data.

c) Make a scatter plot for each group using the first two principal components. Comment on how the different groups separate visually. In particular, comment on the ranges of values for PCA1 and PCA2 of different groups, and how they are different.

Exercise 2

a) Fit a logistic regression model for **cancerous** status as a function of all predictors except **grp** and **id**. Use stepwise selection to choose the best model. Comment on which predictors are chosen based on the stepwise selection method.

b) Given the model chosen from part a), remove any influential points using a cutoff value of 1. Comment on the diagnostic plots (focus on residual plots and influence points). What does Hosmer-Lemeshow's test tell us about goodness of fit of the final model?
Hint: if there's no influential point above the cutoff value, you need to state so in the report.

c) Using the final model chosen from part b), comment on the results of the global test and significance of the parameter estimates. Interpret what the odds ratios tell us about the relationship of the predictors to cancerous status.

d) Predict the probabilities of cancerous status using the final model chosen from part b). Only print the first 10 observations for the predicted probabilities. Classify each observation into cancerous vs healthy using the predicted probabilities. Comment on the separation performance of the logistic regression model. What's the total number of misclassified observations?

Exercise 3

a) Perform a discriminant analysis for **grp** as a function of all the continuous variables (except **id**). Test whether LDA or QDA is more appropriate. Comment on what the MANOVA tests tell us about possibility to discriminate between class types based on these variables.

b) Repeat part a) using only the biomarkers chosen in a stepwise discrimination as predictors for the classification. Comment on which predictors are chosen based on the stepwise discrimination procedure. Test whether LDA or QDA is more appropriate. Comment on the cross-validation error results and how well the discrimination matches the class types.

c) Now split the cancer data into a training set (N=112) and a test set (N=50). Observations from the test data should be randomly chosen using **proc surveyselect** (with random seed=123456789). Perform a discriminant analysis for **grp** on the training set using the model chosen from part b). How does the discriminant analysis perform on the test data?

Exercise 4

Based on the results from exercises 2-3 (for exercise 3, compare the results from 3a and 3b), which method best classifies cancerous vs healthy population in this particular dataset? Justify your conclusion based on the total number of misclassified observations in each method.

Exercise 5

For Exercise 5, use the **housing** data set in **Midterm2Data.sas**. The data set is based on the **housing** data[i] from the UCI Machine Learning Repository[ii] and the raw data is contained in **housing.txt** in the course space. The data contains 13 housing characteristics (the predictors) and 1 outcome variable MEDV (median value of owner-occupied home).

a) Fit a gamma model with log link for **medv** as a function of all 13 predictors. Comment on which predictors seem to be significant and insignificant in this model based on type 3 analysis.
b) Start with all predictors in the model. Based on significance of the parameter estimates, AIC criteria, and Type 3 analysis, manually perform backward selection by removing statistically insignificant predictor one at a time until all remaining predictors reach 0.05 significance level. State which terms you would keep based on this selection.
c) With the final model chosen in b), check residual plots against predicted values for any indication of problems with model assumptions. Interpret what the model tells us about the relationship between the predictors and median home values.

Exercise 6

Use the **epi** data set in **Midterm2Data.sas** based on the Epilepsy Data on pages 266 and 267 of the text. This data set is described below, for this exercise we will only focus on the number of seizures in the fourth period (**Period4**).

- **ID:** subject identifier
- **Period1**: number of seizures in first period
- **Period2**: number of seizures in second period
- **Period3**: number of seizures in third period
- **Period4**: number of seizures in fourth period
- **Treat**: treatment, 0=placebo, 1=progabide
- **BL**: baseline seizure count
- **Age**: age in years

In many real life applications, after you finalize the statistical models in your statistical analysis plan (SAP), any changes made to your pre-specified models (such as removing any statistically insignificant predictors) may be a huge challenge and ordeal. For example, in double-blind, randomized, confirmatory clinical trials, after data base is locked and unblinded, any changes made to the SAP will need to go through company-wide and regulatory approvals. The statistician will need to justify these changes and will be closely scrutinized.

a) Fit a log-linear Poisson model of **Period4** as a function of **Treat**, **BL**, and **Age**. Use over-dispersion with deviance scale if needed. Comment on significance of the parameter estimates, and what the type 1 and type 3 analyses tell us about terms a statistician would want to remove from the model, but do not remove those terms from the model.

b) Continue with the model from part a) containing all three predictors. Check residual plots against predicted values for any indication of problems with model assumptions. Interpret what the model tells us about the relationship between the predictors (especially treatment) and seizure counts after four treatment periods.
   Hint: for residual plots, you need to zoom in to the area where most residuals are concentrated at.

[i] http://archive.ics.uci.edu/ml/datasets/housing

[ii] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.