Midterm Exam 1

Final code and report are due on **3/28** at **11pm** via the **Midterm 1** 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> Midterm1.sas*' and '*<Your-First-Name> <Your-Last-Name> Midterm1.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** statements for **proc freq**, 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, the text book *A Handbook of Statistical Analyses using SAS*, and the recommended text book *Common Statistical Methods for Clinical Research with SAS Examples.*

Data Set 1:

For exercises 1-5, use the **blood** data set as defined in **Midterm1Data.sas** file from the course space. This data is based on a simulated clinical trial project that was done here at UIUC by Serena Chan and Ruixuan Zhou. Systolic blood pressure measures the severity of a patient's hypertension. There are four treatment groups (coded in variable **RTRTN**, which stands for randomized treatment groups). The main objective here is to investigate whether the new investigational drug, ABC123, is more effective at improving patient's hypertension compared to a reference drug. The variables in **blood** data are:

- **USUBJID:** patient's ID number
- **AGE**
- **SEX**
- **RACE**
- **RTRTN:** 1. Reference 2. ABC123 20mg 3. ABC123 40mg 4. ABC123 80mg
- **SITE**: study's participating sites
- **BASE**: baseline systolic blood pressure
- **VALUE**: systolic blood pressure measured at the visit
- **Responder**: whether a patient responds to the treatment or not: 1. Yes, 0. No. The definition of responder is whether blood pressure value<=120 or not.

Exercise 1

For systolic blood pressure value (variable **value**), provide basic descriptive statistics and histograms and comment on general features (eg, mean, median, spread, skewness, and range). Visually and quantitatively test whether normality is a reasonable assumption for systolic blood pressure values.

Do the same analysis for blood pressure **value** by treatment groups (RTRTN) and comment on normality assumption for different treatment groups. Also compare the means and standard deviations among different treatment groups.

Exercise 2

Focus only on two populations: reference (RTRTN=1) and ABC123 80mg (RTRTN=4). Perform a two-population means comparison. Test whether ABC123 80mg significantly reduces the blood pressure compared to the reference drug. State your conclusion.

Hints: Use the results from exercise 1 to determine which hypothesis tests should be used. To answer this question, perform one-sided test to see if the mean blood pressure in the reference group is significantly higher than the mean blood pressure in ABC123 80mg group.

Exercise 3

Perform frequency analysis for association between treatment groups (**RTRTN**) and responder categories **(responder)**. Consider patients in all four treatment groups. Comment on apparent associations, perform appropriate tests for association and state your conclusions about the association (if there is association, state the magnitude of association). Is it appropriate to use the Mantel-Haenszel chi-square test in this case, why or why not?

Consider patients only in the reference and ABC123 80mg groups. In addition to commenting on apparent associations, testing for significant association, and magnitude of association, test whether there are more responders in the ABC123 80mg group compared to the reference group.

Exercise 4

There are four categorical predictors in the data set (**sex, race, rtrtn, site**). Obtain your best ANOVA model for blood pressure value (**value**) as a function of these categorical variables. First, test out the main effects using PROC GLMSELECT and stepwise selection, retain only the statistically significant main effects. Then investigate the interactions between the significant main effects using type III SS. State how you obtained your final model, comment on the proportion of variation in blood pressure value explained by the model.

Compare the LS means, and differences between LS means among different treatment groups. Rank the treatment groups based on efficacy at reducing blood pressure, state the differences between LS means, 95% CI of the differences, and p-values. Which treatment groups are statistically significantly different in efficacy? Perform model diagnostics to validate your model assumptions. Perform formal normality tests on the residuals.

Exercise 5

Now consider an ANCOVA model, with blood pressure value (**value**) as the response variable. Use all the categorical predictors (**sex, race, rtrtn, site**) and covariate baseline score (**base**) as your predictors, obtain your best ANCOVA model using PROC GLMSELECT and stepwise selection. Retain only the statistically significant main effects, you do not need to test out interaction terms. State which main effects are kept in your final ANCOVA model. Explain how one-unit increase in the baseline score impacts the final blood pressure value. Compare the predictors chosen and total variation explained between the ANOVA model from Exercise 4 and this ANCOVA model. Check model diagnostics for highly influential observations (use Cook's cut-off 1). Specify your finding even there is no highly influential observation. Perform formal normality tests on the residuals. You do not need to check the equal slope assumption (i.e. you don't need to check the interaction terms).

Data Set 2:

For exercise 6, use the dataset **housing** defined in **Midterm1Data.sas** in the course space. 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). Our goal is to predict median housing price using housing characteristics. The following variables are included in **housing.**

1. **CRIM**: per capita crime rate by town
2. **ZN**: proportion of residential land zoned for lots over 25,000 sq.ft.
3. **INDUS**: proportion of non-retail business acres per town
4. **CHAS**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. **NOX**: nitric oxides concentration (parts per 10 million)
6. **RM**: average number of rooms per dwelling
7. **AGE**: proportion of owner-occupied units built prior to 1940
8. **DIS**: weighted distances to five Boston employment centres
9. **RAD**: index of accessibility to radial highways
10. **TAX**: full-value property-tax rate per $10,000
11. **PTRATIO**: pupil-teacher ratio by town
12. **BB**: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. **LSTAT**: % lower status of the population
14. **MEDV**: Median value of owner-occupied homes in $1000's

Exercise 6

Consider a linear regression model, predicting median housing price (**MEDV**) as a function of all 11 continuous predictors (use all predictors except **CHAS** and **RAD**). First, use stepwise selection method to choose your best linear regression model for **MEDV** as a function of the statistically significant predictors. After you obtain the best model, check if there's any multicollinearity problem among the remaining predictors. Remove highly correlated predictors if appropriate. Remove any points you deem highly influential and refit the model if necessary. If you need a rule of thumb for influence, leave no points in the data that have a Cook's distance greater than 1. State your final model. Comment on the amount of variation explained by the model. Interpret the relationship between median housing price and predictors chosen (eg, explain how one-unit increase in predictor X impacts the outcome). Identify any issues that remain in the diagnostics. Perform formal normality checks on the residuals.

Hints: you need to specify your findings in VIF and Cook's distance even if you end up not removing any predictors or data points.

[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.