Homework 8 of STAT 540
Section 001, Fall 2024
Due: Wednesday Nov 20, 2024 (before class)
Total Points: 96

Please hand in a hard copy of your homework in class (SAS code and output) and email your SAS code to Kaniz Fatema (KFATEMA@email.sc.edu). Work on the homework independently.

Problem 1. Download the data file Pima.txton your computer. The description of the data set can be found in the file readmePima.txt.

a. Read the data file **Pima.txt** into SAS and generate a SAS data set named **pimaData**. Use the following names for the variables in order in the text file: **npreg**, **glu**, **bp**, **skin**, **insulin**, **bmi**, **ped**, **age**, and **type**. Print out the first 10 observations to make sure that you have read the data correctly. Report the total number of observations in this data set. (5 points)

b1. Use **proc means** to find the following summary statistics: min, Q1, mean, median, Q3, max for variable **glu** at different levels of variable **type** and add title "Descriptive statistics of glu by type". Only these descriptive statistics and no other statistics are allowed for this procedure. Present the output. (5 points)

b2. Plot a side-by-side boxplot of **glu** by variable **type** with a title "Side-by-side boxplot of glu by type". (5 points)

c1. Use **proc means** to find the following summary statistics: min, Q1, mean, median, Q3, max for variable **bp** at different levels of variable **type** and add title "Descriptive statistics of bp by type". Only these descriptive statistics and no other statistics are allowed for this procedure. Present the output. (5 points)

c2. Plot a side-by-side boxplot of **bp** by variable **type** with a title "Side-by-side boxplot of bp by type". (5 points)

d. Create a SAS macro named **inference** with an argument **var1**. This macro basically does the following jobs as in the above questions: (1) read the text file **Pima.txt** into SAS, (2) obtain the same set of descriptive statistics of **var1** at different levels of **type**, and (3) create a side-by-side boxplot of var1 by variable type. Pay attention to using the right titles in the output. Run **%inference(var1=glu)** and **%inference(var1=bp)** to make sure you have written this macro correctly. Implement **%inference(var1=bmi)** and present the output. (15 points)

e. Use **proc ttest** to compare the population means of **bmi** for all potential pima people with **type=1** and people with **type=0**. State your conclusion clearly about your hypothesis test at 0.05 level of significance. (10 points)

Problem 2. The file **tlc.dat** contains the measurement of blood lead levels of children who were in a study. Those children were living in an environment exposed with paints containing lead in their life before the study. This data set contains 100 children with high lead levels who were randomized into two groups: A (treatment) and P (placebo). Four measurements were taken for blood lead level at week 0, week 1, week 4, and week 6, respectively. The columns in this file in order are: ID, group, Y_1, Y_2, Y_3, and Y_4, respectively. The data structure is called longitudinal data, in which the response variable is measured multiple times at different time points.

(a) Read the data into SAS with variable names: ID, group, Y_1, Y_2, Y_3, and Y_4 in a SAS data set named **tlcWide**. Print out the first 6 and last 6 observations in **tlcWide**. (5 points)

(b) Use **proc corr** to generate the scatter plot matrix of Y_1, Y_2, Y_3, and Y_4. Report the covariance matrix and correlation matrix of these four variables. (5 points)

(c) Now create a new data set **tlcLong** based on **tlcWide** with a new variable Y taking values of Y_1 through Y_4 and a new variable named time taking values of 0, 1, 4, and 6 for each subject. Discard variables Y_1, Y_2, Y_3, and Y_4 from **tlcLong**. Print out the first 20 observations in **tlcLong**. The next few questions are all based on **tlcLong**. (5 points)

(d) Sort **tlcLong** by group and time. Plot a series or trace plot of Y versus time for subject with ID=1 using **proc sgplot**. Create a separate figure to plot series plots for all subjects together. (7 points)

(e) Use **proc sgpanel** to create a panel plot of Y versus time for subjects in the placebo group and treatment group side-by-side. (7 points)

(f) Use **proc means** to obtain the sample mean and variance of Y for each group and each value of time. Save these statistics in a new SAS data set named statbyGroupTime. Print out all data in statbyGroupTime. (5 points)

(g) Based on the SAS data set **statbyGroupTime**, Create series plots **ymeanclass** versus **time** for the two different groups on the same figure. Add title “mean trend of blood lead level over time for different groups”. (5 points)

(h) Based on the SAS data set **statbyGroupTime**, create series plots **yvarclass** versus time for the two different groups on the same figure. Add title “variance plot of blood lead level over time for different groups”. (7 points)