# MA409: Statistical Data Analysis (SAS)

## Assignment 3

**Note:** Please work on problems 1, 2, 4, 5 by SAS procedures. The SAS datasets are under *~/my_shared_file_links/u44964992/Assignments/* on SAS OnDemand for Academics. Please provide the SAS code in a separate ***.sas file*** and the outputs from SAS (using screenshots) together with problem 3 in a ***PDF file***.

1. Generate a random sample of size $n = 20$ from the standard normal distribution $N(0, 1)$, use the **one-sample t-test** to test the hypothesis $H_0: \mu = 0$ vs. $H_1: \mu \neq 0$ ($\mu$ is the population mean of the sample). Repeat the process 10000 times where each time you would obtain a p-value. Provide the histogram of the 10000 p-values and state your finding. Note: **use seed 12321** to make sure that everyone gets the same histogram. Hint: to put the 10000 p-values in a dataset, you may need to perform the t-test by raw coding instead of using PROC TTEST. (10 point)

2. This is a game to convince you that multiple comparisons (also called multiple testing) would result in Type I error inflation. Suppose that a researcher would like to study the association between 6 common diseases (i.e., arthritis, asthma, coronary heart disease (CHD), diabetes, liver cirrhosis, lung cancer) and 8 candidate "risk factors" (i.e., Leo Zodiac sign (狮子座), born between 6:00am and 8:00am, Kobe Bryant's fan, purple as favorite color, red hair, first-name begins with letter "C", ID number ends with 1, both parents born in summer) based on a large cohort ($n = 10{,}000$). All variables (6 for the diseases and 8 for the risk factors) are binary (0/1 meaning without/with the disease or risk factor). The researcher decides to use the Pearson's chi-square tests to test independence between all $6 \times 8 = 48$ pairs of variables simultaneously and to summarize the p-values in a table of the form

| | arthritis | asthma | CHD | diabetes | liver cirrhosis | lung cancer |
|---|---|---|---|---|---|---|
| Leo | | | | | | |
| 6:00-8:00am | | | | | | |
| Kobe fan | | | | | | |
| love purple | | | | | | |
| red hair | | | | | | |
| first name C | | | | | | |
| ID number 1 | | | | | | |
| summer | | | | | | |

(1) Simulate a cohort with $n = 10{,}000$ individuals and 14 binary variables. Assume for

simplicity that each disease has an incidence of 10% and each risk factor has a frequency of 5%. Most importantly, simulate the 14 variables independently so that there is no underlying association between the diseases and the risk factors. Note: **use seed 12321** to make sure that everyone gets the same dataset and results. (5 points)

(2) Before performing the Pearson's chi-square tests, compute (by hand calculation) and interpret the FWER of the $6 \times 8 = 48$ simultaneous tests without any adjustment for multiple comparisons. Suppose $\alpha = 0.05$ is used for each test. (5 points)

(3) Perform each of the $6 \times 8 = 48$ chi-square tests and fill the table above with the corresponding p-values (fill the table by hand). Highlight the corresponding cell(s) of the table if the p-value is significant and briefly interpret your conclusion. Hint: using the macro facility may help. (10 points)

3. Show that the weighted least squares estimate defined by Eq. (5.17) in the lecture notes is the best linear unbiased estimate (BLUE) of $\boldsymbol{\beta}$. (10 points)

4. The dataset "*exercise.sas7bdat*" consists of data describing the amount of weight loss achieved by 900 participants in a year-long study of 3 different exercise programs, a jogging program, a swimming program, and a reading program which serves as a control activity. Researchers were interested in how the weekly number of hours subjects chose to exercise predicted weight loss. For simplicity, focus on three variables in this problem:

   - *loss*: response variable describing the average weekly weight loss for participants (positive scores denote weight loss, negative scores denote weight gain).
   - *hours*: predictor variable describing the average number of weekly hours of exercise.
   - *effort*: predictor variable describing the average weekly subject-reported effort scores (range from 0 to 50, 0 denotes minimal physical effort and 50 denotes maximum effort).

   (1) Compute the mean and standard deviation (sd) of *effort*. Then, create a categorical variable *effort_cat* based on *effort*: if *effort* is between -1 and 1 sd of its mean, set *effort_cat* to "medium"; if *effort* is below its mean minus 1 sd, set *effort_cat* to "low"; if *effort* is above its mean plus 1 sd, set *effort_cat* to "high". Plot the scatterplot of *loss* vs. *hours* with a regression line by *effort_cat* (i.e., the interaction plot) and describe your finding. (10 points)

   (2) Fit a linear regression model of *loss* on *hours*, *effort* and possibly the interaction effect of *hours* and *effort* depending on your finding in (1). Provide the estimated regression coefficients and briefly state your conclusion. (5 points)

   (3) For the model in (2), check the homoscedasticity assumption (with both the fitted vs. residual plot and the White test) and the normality assumption (with both the Q-Q plot and the Shapiro-Wilk test). Display your results and briefly state your conclusion. (5 points)

5. "*cancer_reg.csv*" provides data aggregated from several sources including cancer.gov and census.gov for 3,045 US regions averaged from year 2010 to 2016. The response variable is *deathRate* (*per capita* (100,000) cancer mortalities, i.e., number of death due to cancer per 100,000 residents) and 14 predictor variables are included:

| | |
|---|---|
| $X_1$: *per capita* (100,000) cancer diagnosis | $X_8$: Percent of residents (age 16 and over) unemployed |
| $X_2$: Median household income (in 1,000 dollars) | $X_9$: Percent of residents with private health coverage |
| $X_3$: Percent of populace in poverty | $X_{10}$: Percent of residents with public health coverage |
| $X_4$: Median age of residents | $X_{11}$: Percent of residents who identify as White |
| $X_5$: Median age of male residents | $X_{12}$: Percent of residents who identify as Black |
| $X_6$: Median age of female residents | $X_{13}$: Percent of residents who identify as Asian |
| $X_7$: Percent of residents who are married | $X_{14}$: Percent of residents (age 25 and over) with bachelor's degree |

(1) Check the pairwise Pearson correlation coefficients using PROC CORR. Is there any collinearity between pairs of predictor variables? Please explain. (5 points)

(2) Fit the regression model of *deathRate* on all 14 predictor variables. Does multicollinearity exist? (5 points)

(3) Use PROC REG to fit the ridge regressions of *deathRate* on all 14 predictor variables with 21 equally spaced values of $\lambda$ on $[0, 1]$. Output the parameter estimates under ridge regression models with different $\lambda$ to a SAS dataset, then plot the lines showing the parameter estimates of X1, X3, X4, X5, X6, X8, X14 against $\lambda$, and state your findings. Note: make sure to standardize the response and predictor variables before fitting the ridge regression models (standardization can be performed using PROC STANDARD). (10 points)

(4) Use PROC GLMSELECT to perform stepwise variable selection, specifically, apply the BIC criterion (i.e., SBC in PROC GLMSELECT) to determine the order in which variables enter or leave at each step, as well as to select the best model. Display the adjusted R-squared, Mallows's $C_p$, AIC and BIC values at each step and generate a plot of these criteria by step. State your final model and **explain your conclusions**. (10 points)

(5) Based on your final model in (4), compute the studentized residual $r_{\text{stu}}$ for each region and filter out those with $|r_{\text{stu}}| > 3$ as possible outliers. Print the filtered regions (with *Region*, *deathRate*, *X1,* other variables in the model, and *rstu*) sorted by *X1* (i.e., number of cancer diagnosis per 100,000 residents) in ascending order. Among these possible outliers, is there anything unusual so that you would like to remove some of them from your analysis? Please **state your justification** clearly. (10 points)