

## STA 402/502 Homework 3

Due: September 21th (Friday), before class

Please read the homework guidelines before working on the homework. Please submit a well-formatted printed copy of the homework. Please include graphs/tables that you think are important and most relevant to your answers. Do not include all the output in the homework. You are to complete this assignment on your own. Remember to include an intro comment block on all programs written. Each problem should be attempted as its own program.

1. Create simulated data for a possible regression model, so that the predictor values  $x$  range from 1 to 20 in steps of 0.1 and the response value  $Y$  (at each  $x$ ) is normally distributed with mean  $\mu = 1.2 - 3x + 0.2x^2 + \sin(\frac{1}{2}\pi x)$  and standard deviation  $\sigma = 4$ . Use PROC SGPLOT to create a scatter plot of the data with a superimposed regression line. (Hint: you may use `constant("pi")` in SAS to represent  $\pi$ )
2. The file "BenAndJerrys.dat" contains various ice cream flavors and their nutritional information. The variables are flavor name, portion size (g), calories, calories from fat, fat (g), saturated fat (g), trans fat (g), cholesterol (mg), sodium (mg), total carbohydrate (g), dietary fiber (g), sugars (g), protein (g), year introduced, year retired, content description, and notes.
  - (a) Read the dataset into SAS. Examine the raw data and make sure you specify appropriate variable type and length to read in everything correctly. Include the sas log related with reading in the dataset in your solution. How many observations are there in this dataset?
  - (b) Some icecream flavors cannot be purchased at the grocery store (that is the flavors that have retired, or the flavors which are scoop shop exclusives as described in the variable "notes"). Please create a new dataset, keeping only the flavors that can be purchased

at grocery store. Include the sas log related with creating this subset of data. How many observations are there in this new dataset?

- (c) (STA 502 only) If an icecream has calories per tablespoon less than 30, then it can be viewed as “low calories icecream”. Create a variable that calculates the calories in one tablespoon ( $TB$ ) of ice cream. Assume that  $1TB = 15g$ . Subset the data again. Keep only observations related with “low calories icecream” flavors and make sure the calories per tablespoon value is not missing. Also make sure the icecream flavors are able to be purchased at the grocery store. Print a table containing the flavor, calories, the content description and the calories in one TB for the final dataset.
  - (d) Sort the data you read from question (a), by calories (by descending order), cholesterol (by descending order) and sodium (by ascending order). Print the flavor, calories, cholesterol and sodium information for the observations with calories value equal to 280. (Hint: check how to sort data by several variables in PROC SORT)
3. Refer to the homework 2 question 2, “wls.sas7bdat” dataset and the introduction. Suppose the survey questions are measured on a scale of 0 (least) to 3 (most). Suppose all the questions in the second visit have been reverse coded. That is 3 (least) to 0 (most). Use array to convert the measure of those six visits to the correct scale. After that, please print the first three observations of the dataset. Only print ID, height, weight, BMI and the six questions in the second visit related with these observations.