ST512 – Lab 01

These are not exercises to be handed in for a grade. These are just practice problems that should help with completing the homework assignment covering the same material; there may also be relevant conceptual extras not covered in lecture. You're welcome to discuss these problems with your classmates or with the instructor/TA in office hours.

Part 0: Data Description.

(Taken from Casella & Berger, Example 11.3.1.) During the month of July, grape vines produce clusters of berries, and the number of such clusters can be used to predict the ultimate crop yield at harvest time. The data consists of two variables—number of clusters and grape crop yield (units = tons per acre)—for a twelve-year period. The goal is to analyze these data, in particular, to determine the relationship between the number of clusters and crop yield, for the purpose of prediction.

Part 1: Loading Data in SAS.

The SAS data step is the crucial step where the to-be-analyzed data are loaded. There are basically two ways to complete the data step: type the data in manually (or copy–paste) or read the data from an external file, such as an Excel spreadsheet. Here is a brief description of each approach; try them out.

- *Load data manually.* The basic syntax is as follows:

  ```
  DATA name;
  INPUT var1 var2 ...;
  DATALINES;
  ## ## ...
  ## ## ...
  ;
  ```

  This results in a SAS data set named `name` that contains variables `var1`, `var2`, etc. The SAS program on the course website contains the relevant code for reading in the data set `grape` that contains the two variables `nclust` and `yield`.
- *Read data from file.* Go to File → Import Data and follow the steps. You need to save the data file somewhere on your local machine and then point the import wizard to that directory and file. Also, it may help to save the code that is generated by the wizard, which is one of the options; you can use a modified version of this code to read in other data sets without using the wizard.

Part 2: Graphical and Numerical Summaries in SAS.

After loading the data into SAS, we are ready to do some analysis. Before we get to the simple linear regression modeling for the grape yield data, we should go over some basic graphical and numerical summaries in SAS. The basic SAS procedures for this are PROC MEANS and PROC UNIVARIATE, and these (with the relevant options) can give you virtually any graphical or numerical summaries imaginable. Use the SAS code on the course website to generate some summaries for the grape yield data and answer the following questions.

1. Find the means and standard deviations for `yield` and `nclust`.
2. Find the five-number summaries[1] for `yield` and `nclust`.
3. Find a 95% confidence interval for the mean of `yield`.
4. Draw histograms to visualize the marginal distributions of `yield` and `nclust`. Do either of these distributions look normal?

In ST512 you will have to submit write-ups of your homework solutions and there you will need to insert *relevant* SAS output; think of these as "practice research reports." Unfortunately, what you see in the SAS output window cannot easily be copied and pasted into a written document, so you need to *export* the output to a separate file from which you can easily copy–paste. This is done by inserting the `ods` command (with options) at the beginning and end of the chunk of code whose output you want to save; `ods` stands for "Output Delivery System." The provided SAS code shows how to do this for the grape data, but some tailoring may be necessary for your specific machine.

Part 3: Simple Linear Regression in SAS.

There are a number of ways that linear regression can be carried out in SAS, the one we will consider here is PROC REG. This procedure will give us relevant output from the least-squares fit, as well as some plots to help us assess whether the model assumptions are satisfied. Use the SAS code provided on the course website to fit the simple linear regression model to the `grape` data set—with `yield` as the response variable and `nclust` as the predictor variable—and answer the following questions.

1. Draw a scatterplot of the data with `nclust` on the horizontal axis and `yield` on the vertical axis. Does the relationship between these two variables seem to be linear, or at least approximately linear?
2. Find the estimates of the slope and intercept parameters and the corresponding standard errors.
3. Is the slope is significantly different from zero? Formally state your hypotheses, carry out the test, and carefully state your conclusions.
4. Suppose that the number of clusters this year is `nclust` = 102. Estimate the mean grape crop yield for this value of `nclust` and give a 95% confidence interval.
5. Draw a plot of the residuals versus the predictor variable `nclust`. What does this plot say concerning the assumptions of the simple linear regression model?

[1] The five-number-summary consists of the minimum, 25th percentile, median, 75th percentile, and maximum.