

ST 307 Project 3

For this activity you will create a SAS program and upload that program to Moodle.

- Everyone must submit one's own code.
- Be sure that your SAS file adheres to the SAS file submission guidelines (available on Moodle). You can save the program at any time. If you'd like to work on it more before submitting, you'll probably want to save it and email it to yourself (or use google drive, dropbox, etc.)

Dataset:

The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: <http://www.vinhoverde.pt/en/> or the reference [Cortez et al., 2009].

Variable	Description
quality	The quality of wine. (range from 1 to 10)
type	The type of wine. (1 for red wine and 2 for white wine)
sugar, freeSO2, totalSO2, density, pH, alcohol	6 different physical and chemical indicators that are possibly related to the quality of wine.

Total points: 26

Tasks:

Write code corresponding to each step below, that is, do not change the code for step 1 to do step 2 (you can copy and paste it so you don't have to retype it, but leave the answer to each step in your program):

1. Create a library called Project3, in which you want to put your data sets. (1 pt)
2. In a DATA step, read in the winequality.csv and save it as dataset "winequality" in library Project3. (2 pt) Create a new variable COLOR, which equals "Red" for the individuals with type = 1 and "White" for the others. Drop the TYPE variable. (1 pt)
3. Total sulfur dioxide (totalSO2) is a measure of both the free sulfur dioxide (freeSO2) and bound forms sulfur dioxide. Write a PROC step to test whether the mean of totalSO2 is greater than that of freeSO2 by 85 at significant level 0.05. (Hint: Use PAIRED statement) (3 pt) And answer the following questions in comments:
 - a. Check the normality assumption and answer in the comments. (1 pt)
 - b. State the p-value and a conclusion for the test conducted. (1 pt)

- c. State the confidence interval and does it agree with your conclusion in the previous part. (1 pt)
4. Write a PROC step to test whether the quality of red wine is different from that of white wine at significant level 0.05. (3 pt)
 - a. State the p-value and a conclusion for the test conducted. (1 pt)
 - b. Which one is greater, the quality of red wine or that of white wine? How do you conclude this based on SAS output result? (1 pt)
5. Perform a correlation analysis among 4 variables: SUGAR, PH, ALCOHOL and QUALITY. (1pts) Answer in comments: which variable has the highest correlation with QUALITY? (1pts)
6. Some experts believe that adding more sugars (SUGAR) can improve the subjective quality of wine (QUALITY). Run a simple linear regression to answer this question. (1pts) Answer in comments:
 - a. Does adding sugars have an effect on the quality and how do you know?(1pts)
If yes, in a positive way or negative? (1pts)
 - b. Write down the fitted linear model. (1pts)
7. Your friend wants to build a regression model to predict the quality of wines. Fit a multiple linear regression model with the QUALITY as the response and the other variables except for COLOR as the predictors. (2pts) Answer in comments:
 - a. Use the missing Y trick to obtain and report the prediction interval for a bottle of wine with the following variable values: sugar = 2.5, freeSO2 = 17, totalSO2 = 42, density = 0.998, PH = 3.4, alcohol = 9.6. (2pts)
 - b. Which variable(s) are not significant? (1pts)