Course Project:
Relationship between Plasma Homocysteine and Cognitive Status
Due December 7, 2022

You are to work *alone* on this project and direct any questions about the project to Professor Leatherman and not to Samantha or Priya.

In cross-sectional studies, elevated plasma homocysteine levels have been associated with poor cognition and dementia. A prospective study used data from the Framingham Study to show that increased plasma homocysteine is an independent risk factor for the development of Alzheimer's disease (NEJM 2002).

We will use a subset of the data used in that study to perform various analyses. The data are in three saved SAS data sets that you can download from Blackboard Learn. Combine the data into a single SAS data set, create other variables as necessary, and perform the analyses as requested. You are to use only one program for the project and attach the program, log, and output as appendices to your written report.

Variables:

DEMOG_BS805_F22.SAS7BDAT

DEMOGID	Subject ID in Demographics data set
AGE	age in years
MALE	(=1 if male, =0 if female)
EDUCG	(=1 if education < 8 years, =2 if education >=8 years but no HS degree, =3 if HS degree but no college, = 4 if at least some college)
PKYRS	pack years of cigarette smoking

LABS_BS805_F22.SAS7BDAT

LABSID	Subject ID in Labs data set
HCY	plasma homocysteine level (µmol/L)
FOLATE	plasma folate (nmol/mL)
VITB12	plasma vitamin B12 (pmol/L)
VITB6	plasma vitamin B6 (nmol/L)

NEURO_BS805_F22.SAS7BDAT

NEUROID	Subject ID in Neuro data set
MMSE	Mini-Mental State Examination (a measure of cognitive function with range 0-30)
ADIN7YRS	(=0 if no AD in 7 years of follow-up, =1 if AD in 7 years of follow-up

Please note:

a) You should supply answers to the questions below in a written report. Your report should include introductory and summary paragraphs. For each question, you should write a summary paragraph supplemented by in-text tables to display all relevant data. Please include your SAS program, log, and output in an appendix to your main report. Also, using SAS "title" statements, please clearly indicate on your output which question each particular section of your output is addressing. Comments written on your output do not substitute for including information on your report, so please make sure that **all** information that you want to present and discuss is in the report.

b) In the output that you place in the appendix, please make sure that **all** variables have labels and are formatted where appropriate.

c) **You must work alone on this project. Collaboration with other students is not allowed. Your submission of the project assumes that you concur that the work that you have submitted is solely your own. Please consult with Prof. Leatherman if you have questions.**

d) Below, we ask questions for which formal statistical analyses are required. It will be up to you to determine and implement the appropriate procedure (e.g., test of means, test of medians, test of proportions, etc.). **Please perform all statistical tests using a significance level of 0.05.** If you identify any outliers or influence points note them in your report but *do not remove any data values*.

e) **Do not submit any output from PROC PRINT.**

Notes on grading:

We will be looking for the following:

- Correct use of statistical tests
- Clarity of presentation and interpretation
- Clarity of programming
- Use of higher level programming skills learned in this course

1) As noted above, first combine the three data sets into a single, temporary SAS data set.

2) Using the temporary SAS set in 1), create a permanent SAS data set that includes the following new variables:
   - Plasma homocysteine is very skewed. Create a new variable LHCY, which is the natural log of HCY. We will use this as the continuous form of plasma homocysteine.
   - Create a categorical homocysteine variable, HCYGE14, which is 1 for those whose homocysteine (HCY) is at least 14 and 0 for those whose homocysteine is less than 14.
   - Create a 4 level AGEGRP variable (65-74, 75-79, 80-84, and 85-89).
   - Create a binary variable called HSDEG that is based on EDUCG: High school degree or higher vs. Less than high school degree.
   - Create a variable EXCLUDE, which is 1 for those whose ADIN7YRS variable is missing (these subjects will be excluded later), and 0 for those with non-missing ADIN7YRS.
   - Create a variable MMSEF, which flags subjects with cognitive deficits according to the MMSE. This variable is based on education and MMSE, and is formed as follows.

     - For subjects with < 8 years of education, MMSEF = 0 if MMSE >22 and MMSEF = 1 if MMSE is 22 or less
     - For subjects with >=8 years of education but no HS degree, MMSEF = 0 if MMSE >24 and MMSEF = 1 if MMSE is 24 or less
     - For subjects with a HS degree but no college, MMSEF = 0 if MMSE >25 and MMSEF = 1 if MMSE is 25 or less
     - For subject with at least some college, MMSEF = 0 if MMSE > 26 and MMSEF = 1 if MMSE is 26 or less.

3) Perform appropriate statistical hypothesis tests to compare those excluded to those not excluded with respect to age, sex, education (use HSDEG), cigarette smoking (use PKYRS), cognitive status (MMSE), and homocysteine (use both LHCY and HCYGE14).

4) Using the data set created in 2., create a new, temporary SAS data set excluding those subjects whose ADIN7YRS variable is missing. Make sure this data set does not contain the variable EXCLUDE.

PERFORM ALL SUBSEQUENT ANALYSES (5. – 12.) ON THE DATA SET CREATED ABOVE IN QUESTION 4.

5) Create vertical bar charts in the graphics window and generate descriptive statistics for each of the following variables and briefly comment on their distributions: pack years of cigarette smoking; homocysteine (both untransformed and log transformed), plasma folate, and MMSE.

   We will focus on LHCY as the primary independent variable and MMSE as the primary dependent variable in our analyses.

6) Both homocysteine and cognitive function may change with age.
   - Is LHCY (dependent) linearly associated with continuous age?
   - Is MMSE (dependent) linearly associated with continuous age?

7) Now consider the relationship between age group and log homocysteine. Test the null hypothesis that mean LHCY is the same in the four age groups. Do your results support a linear relationship between age and log homocysteine?

8) Now, fit a piecewise linear model to assess to association of log homocysteine (dependent variable) and age by constructing a model with slopes for each of the age intervals: 65 to < 75; 75 to <80; 80 to < 85; and 85-89. Based on the results of 7., 8., and 9., when we later examine the relationship of LHCY and MMSE adjusted for other factors, which form of age would you choose to model as a potential confounding variable or effect modifier, continuous age, piecewise continuous age, or categorized age? Why?

9) Perform a multiple linear regression with interaction with a dummy variable for gender to assess the relationship of LHCY to MMSE (dependent). Is there significant evidence of effect modification by gender on this relationship?

10) Fit a linear regression model with MMSE as the outcome and LHCY as the single predictor. Using any PROC in SAS that you prefer and is appropriate, determine whether LHCY has a linear relationship with MMSE.

11) Fit a full multiple linear regression model with LHCY, MALE, EDUCG (categorical), age (using your choice of variable determined above), and PKYRS as predictors of MMSE. Fully report on the results of this model. This should include estimated values for the slopes, 95% confidence intervals for the slopes, p-values and standardized regression coefficients for each predictor in the model. In addition, perform regression diagnostics on the final regression model. If any dummy variables are included in the model, how would we interpret them? Are there any outliers or influence points in the data that we should be concerned about? Identify and report the DEMOGID values, studentized residuals and Cook's distance for any such points **but do not remove them**. Also, is there multicollinearity among the predictors from the final model? Cite specific numeric evidence that supports your conclusion on this. Finally, is there evidence of joint confounding by factors in this model on the relationship of LHCY and MMSE? Present specific numeric evidence supporting your conclusion.

12) Using PROC GLMSELECT and LASSO with an AIC-based selection criterion, identify the best model using LHCY, MALE, EDUCG (categorical), age (using you choice of variable determined above), and PKYRS as predictors of MMSE. Summarize the results of this analysis. How does this model compare to the model examined in 11.? If they are different in terms of the independent variables selected, speculate on why the results were different. If they are same, also suggest why the results were the same.