22S:30/105, Statistical Methods and Computing
Instructor: Cowles
Spring 2013, Homework 3

Due: Fri., 02/15 in class
Please put your name at the top of your homework, and list the names of any classmates with whom you collaborated.

For problems requiring SAS, include:

- SAS data step for reading the data file (not including embedded data)

- SAS procedure to generate output required to answer each part of the question

- only those parts of the SAS output that directly relate to the question being asked

1. Refer to the dataset called "winehrt.dat" on the course web page.

   (a) Use SAS to produce a scatterplot that shows how national wine consumption may help to explain heart disease death rates.

   (b) Describe the form of the relationship. Is there a linear pattern? How strong is the relationship?

   (c) Is the direction of the association positive or negative. Explain in simple language what this says about wine and heart disease. Do you think these data give good evidence that drinking wine *causes* a reduction in heart disease deaths? Why?

   (d) Use SAS to compute the correlation coefficient r between alcohol from wine and heart disease death rate.

   (e) In the data, wine consumption is measured as yearly liters of alcohol from drinking wine, per person. If it had instead been measured in yearly *ounces* per person, how would the value of r have been affected?

2. Investors ask about the relationship between returns on investments in the U.S. and on investments overseas. The data file "stocks.dat" gives the total returns on U.S. and overseas common stocks over a 26-year period. (The total return is change in price plus any dividends paid, converted into U.S. dollars.) Both returns are averages over many individual stocks. Use SAS to get the numbers required for answering the following questions.

   (a) Make a scatterplot suitable for predicting overseas returns from U.S. returns. (Use SAS.)

   (b) Find the correlation coefficient r. Describe the relationship between U.S. and overseas returns in words, using r to make your description more precise.

   (c) Find the 5-number summaries for both U.S. and overseas returns. Make boxplots to compare the two distributions. (These will not be side-by-side on one plot).

   (d) Were returns generally higher in the U.S. or overseas during this period? Explain your answer.

   (e) Were returns more volatile (that is, more variable) in the U.S. or overseas during this period? Explain your answer.

   (f) Find the least-squares regression equation of overseas returns on U.S. returns.

   (g) In 1997, the return on U.S. stocks was 33.4%. Use the regression line to predict the return on overseas stocks. (You may either calculate this by hand or use SAS output.) The actual overseas return was 2.1%. Are you confident that predictions using the regression line will be quite accurate? Why?

   (h) In your plot, circle the point that has the largest residual (either positive or negative). What year is this?

   (i) Are there any points that seem likely to be very influential? If so, circle them in your scatterplot. (Just write the circles by hand on the printed plot from SAS.)

3. Textbook problems: 4.16, 4.22, 4.23, 4.34, 4.35