SOC364/L - Social Statistics @ CSUN w/ Godard

Homework #8: Correlation and Regression

Using data from the 1984 General Social Survey (gss84.sys), do the following:

1. Make an argument about the possible relationship between the education level of respondents and that of their fathers, using father's education as the independent variable. Using all criteria discussed, assess whether this could be a *causal* relationship.
2. Identify (by name and label) the two variables in the dataset which measure these concepts, and identify their operationalization and level of measurement.
3. Resolve all missing values for each variable, and indicate how you've done so. (If there are more than three categories, you'll need recode all missing values to one, and use *one value* in the missing values command – *although,* if there aren't any cases *with* a missing value, you don't have to worry about – so check frequencies!)
4. Briefly analyze the univariate distributions of the two variables, including both center and spread, in both graphical and statistical terms (that is, tails and shapes *as well as* statistics and their interpretation) and compare them.
5. Make a scatter diagram using the plot command and discuss whether the plot implies the presence of a relationship. If so, what type of relationship appears – positive or negative? strong, moderate, or weak? Linear or something else? See any outliers?
6. Run a regression with respondents' education as the dependent variable and father's education as the independent variable. Then answer the following questions regarding the relationship between these two variables. (Yes, you also need to answer the questions asked above. That's why I asked them.)
   a) What is the slope? What is the y-intercept? Provide an interpretation of these two statistics.
   b) State the regressed prediction line in the form of an equation Y^ = a + bx
   c) Calculate predicted values of the dependent variable for four values (0, 8, 12, and 16) of the independent variable. Mark and label these (x,y) points on the scatterplot, and sketch the prediction line.
   d) For this data, what is the average education predicted for respondents whose fathers had no years of education? How much is each year of father's education "worth" on average in terms of respondents' years of education?
   e) If the null hypothesis is that B = 0 (i.e. that the regression line has a horizontal slope, and that there is a "flat" relationship, such that value of the dependent variable do not vary with values of the independent variable; i.e. there is "no" relationship), test that null hypothesis at the .05 level. (Yes, do all the steps of a hypothesis test: Interpret all values, and make firm conclusions, about both the hypotheses and the prose relationship.)
   f) What is the R-squared value? Interpret it.
7. Run a correlation procedure on the two variables and describe the correlation between the two variables. (Be sure to interpret the size, direction, and significance of the correlation, and to clearly dilineate these.) How is this statistic (the correlation coefficient) similar or different from the standardized regression coefficient computed for the same two variables?

- You may **not** work on this assignment with each other. You may always, of course, ask me anything you wish. Email is usually best – egodard@csun.edu will work.
- You may type answers into your output file (preferred) or a separate file, but *you must* submit your output.
- Be sure to write answers in sentence form. Be explicit and complete in explaining your answers.
- Show all work done, including any calculations, & explain all of the steps involved.