POLI 300 PROBLEM SET #13 due 12/8/10

REGRESSION AND CORRELATION

Note. Some of the following problems are taken or adapted from David S. Moore, *Statistics: Concepts and Controversies*, a text book previously used in this course.

1. Below are three small data sets (a), (b), and (c). In each, X is the independent variable and Y is the dependent variable. For each data set, (i) draw a scattergram, (ii) calculate the correlation coefficient, (iii) calculate the regression coefficient and the constant/intercept, and (iv) draw the regression line in the scattergram.

(a)

X	Y
4	−4
4	4
−4	4
−4	−4

(b)

X	Y
4	−4
3	−2
0	0
−3	2
−4	4

(c)

X	Y
4	4
2	−2
−2	2
−4	−4

2. Below are three even smaller data sets (d), (e), and (f). As before, X is the independent variable and Y is the dependent variable. For each data set, (i) draw a scattergram, (ii) calculate the correlation coefficient, (iii) calculate the regression coefficient and the constant/intercept, and (iv) draw the regression line in the scattergram.

(d)

X	Y
1	4
2	4
3	4

(e)

X	Y
4	1
4	2
4	3

(f)

X	Y
6	10
10	18

Comment on the results of your calculations. (For each data set, your calculations should produce "strange" results.)

3. The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this association is very regular, as shown by the following data on speed (MPH) and mileage (MPG):

MPH:	20	30	40	50	60
MPG:	24	28	30	28	24

Make a scattergram of MPG by MPH. Show that the correlation between the variables is $r = 0$. Explain why the correlation is zero even though there is clear association between mileage and speed.

4. Your data consist of observations on the age of several subjects (measured in years) and the reaction times of these subjects (measured in seconds). In what units (if any) are each of the following descriptive statistics measured?

   (a) The mean age of the subjects.
   (b) The standard deviation of the subjects' reaction times.
   (c) The correlation between age and reaction time.
   (d) The median age of the subjects.
   (e) The variance of the subjects ages.
   (f) The regression coefficient with age independent and reaction time dependent.

5. A college newspaper interviews a psychologist about student ratings of the teaching of faculty members. The psychologist says, "Empirical research indicates that the correlation between the research productivity and teaching rating of faculty members is close to zero." The paper reports this as "Professor McDaniel said that good researchers tend to be poor teachers, and vice versa." Explain why the paper's report is wrong. Write a statement in plain language (don't use the word "correlation") to explain the psychologist's meaning.

6. Each of the following statements contains a blunder. Explain in each case what is wrong.

   (a) "There is a high positive correlation between the gender of American workers and their income."
   (b) "We found a high correlation (r = +1.09) between students' ratings of faculty teaching and ratings made by other faculty members."
   (c) "The correlation between age and income was found to be r = +0.53 years."

7. For each of the following pairs of variables, would you expect a substantial negative correlation, a substantial positive correlation, or little or no correlation?
   (a) The age of second-hand cars and their prices.
   (b) The weight of new cars and their gas mileages in MPG.
   (c) The heights of men and their adult sons.
   (d) The heights and weights of adult men.
   (e) The heights and IQ scores of adult men.

8. Now make a reasonable guess at the value of the *regression coefficient* for each pair of variables (a)-(e) above. Remember that the numerical answer to the regression question (unlike the correlation question) will depend (1) on which variable you consider to be *independent* and which *dependent* and also (2) on the *units* in which each variable is measured (e.g., is FATHER'S HEIGHT in feet, inches, centimeters, etc.?) Thus you must specify in each answer which variable you consider independent and which dependent and also the units used to measure each variable. Successfully answering such questions requires you to (i) understand what question it is that a regression coefficient answers (which *is* important for POLI 300) and (ii) have some appropriate contextual knowledge, e.g., pertaining to the prices, weights, and MPG of different kinds of cars (which *is not* important for POLI 300). With respect to (i), remember that the regression coefficient answers this question: ***how much, on average, does the dependent variable increase (in dependent variable units) for each one unit increase (in independent variable units) of the independent variable***. For example, for (a) it is natural to consider (i) the age of cars to be the independent variable and to measure it in years and (ii) the value of cars to be the dependent variable and measure it in dollars and, so the regression coefficient answers this question: ***how much, on average, does the value of a car increase in dollars for each additional year of age?***

9. Refer to your scattergrams (and/or those provided in the Answers and Discussion) for Questions 1.1 (RE-ELECTION PERFORMANCE by APPROVAL RATINGS), 1.3 (RE-ELECTION PERFORMANCE by GOODNESS OF TIMES, and 2 (ELECTORAL VOTES by POPULAR VOTES) in Problem Set #11.

   (a) *Visually estimate* the magnitude of the *regression coefficient b* in each data set. This requires you to (i) draw the regression line on the scattergram on an "eyeball" basis (which you should be able to do fairly accurately because each pair of variables exhibits a strong correlation) and then (ii) to determine the slope of this line by examining an appropriate "rise over run" triangle. In the manner of Question 8 just above, state in words what each regression coefficient means (i.e., what substantive question is this number the answer to?).

   (b) Also estimate the *intercept a* in each data set. State in words what this number means? You should find that the intercept in value for ELECTORAL VOTES by POPULAR VOTES gives a nonsensical answer to the question that the intercept normally answers. Can you explain why this is so?

   (c) Suppose the U.S. had a system of *proportional representation*, so that each party won electoral votes proportional to its share of the popular vote. Can you say what the regression coefficient then would be? The intercept? The correlation coefficient? (*Note*: there are 538 electoral votes.)

10. This question refers to the scattergram on the back of this page. Read the description of the scattergram that appears below it. Then answer the following questions.

   (a) Though he also lost, 1988 Democratic presidential candidate Michael Dukakis did considerably better than 1984 Democratic presidential candidate Walter Mondale. (See below in (d)). How does this fact show up in the scattergram?

   (b) Were there any states in which Dukakis in 1988 did no better or even worse than Mondale did in 1984? Explain how you can determine this from the scattergram.

   (c) Taking DV88 to be the dependent variable, the regression coefficient *b* is about +0.84 and the intercept *a* is about 12.3. Given this information, draw the regression line directly on the scattergram as accurately as you can. Explain how you did this.

   (d) Nationwide, Mondale received about 41% of the two-party popular vote in 1984, and Dukakis received about 46% in 1988. That is, nationwide there was a swing from 1984 to 1988 of about 5% of the popular vote toward the Democrats. Suppose hypothetically that this national 5% swing from 1984 to 1988 in favor of the Democrats had been *uniform* across states — that is, suppose that in every state Dukakis's 1988 percent of the vote were equal to Mondale's 1984 percent plus 6 percentage points.

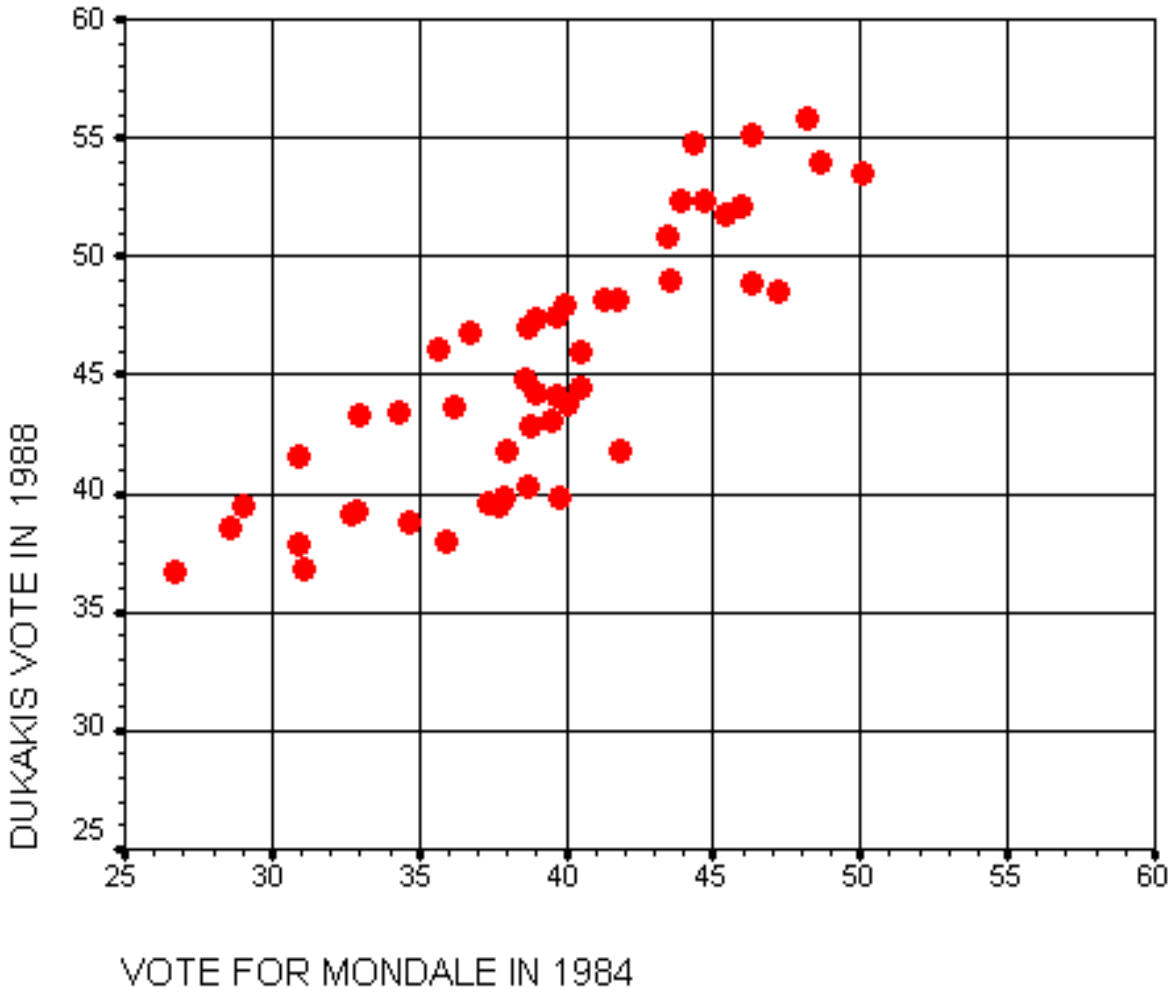
      (1) What would the scattergram look like under this hypothetical circumstance?

      (2) What would the regression (b) and correlation (r) coefficients and the intercept (a) be under this hypothetical circumstance?

      (3) What characteristic of the actual scattergram demonstrates that the swing was *not* a uniform 6 percentage points in every state?

OVER =>

SCATTERGRAM: DUKAKIS VOTE BY MONDALE VOTE



The *cases* plotted in this scattergram are the 50 American states. The *horizontal variable* is the percent of the state's popular vote received by the Democratic Presidential candidate (Mondale) in 1984. The *vertical variable* is the percent of the state's popular vote received by the Democratic Presidential candidate (Dukakis) in 1988. Both percentages are based on the total vote for the two major parties only, so that D% + R% = 100% with minor candidates/parties excluded from the calculation.