POLI 300 **PROBLEM SET #11** 11/17/10

SCATTERGRAMS: ANSWERS AND DISCUSSION

General Comments

In the past, many students' work has demonstrated quite fundamental problems. Most generally and fundamentally, these evidently resulted from failing *to stop and think* whether what you were doing made sense relative to the sentence/hypothesis in question. More specifically by:

(1) simply using the *wrong variable*(s), e.g., using TURNOUT (or PRESIDENTIAL APPROVAL) data as a measure of INCUMBENT RE-ELECTION PERFORMANCE;

(2) *not transforming the data* so as to create a reasonable measure of the variable the sentence/ hypothesis refers to — in particular, not transforming D2PC into INCUMBENT VOTE PERCENT or CLOSENESS OF ELECTION;

(3) making ELECTION YEAR a *variable* (rather than the case label) in a scattergram, with the consequence that the resulting chart was a plot of *a single variable over time*, not a scattergram showing the relationship between two variables;

(4) not plotting *all the cases* for which data is available, e.g., plotting only Democratic victories in connection with #4;

(5) interchanging the *independent* (horizontal) and *dependent* variables (which is more than a cosmetic problem because, as we shall see, interpretation of the *regression slope/coefficient* is dependent on following the standard setup);

(6) in hand-drawn scattergrams, not creating *equal interval scales* for each variable axis (which also is more than a cosmetic problem, since the scattergram may be distorted in a way that makes it hard to see whether and how strongly the variables are associated);

(7) not providing any interpretation of the scattergram and not *presenting substantive conclusions* based on the scattergram about the sentence/hypothesis in question; and

(8) in analyzing the scattergram, establishing arbitrary boundaries (e.g., between "close" and "not close" elections, "popular" vs "not popular" Presidents), when the whole purpose of scattergrams is to reveal the overall relationship between two quantitative variables, *without creating arbitrary categories or class intervals*.

More cosmetic problems included:

(a) not *labeling the axes* with the names of the variables they represent;

(b) not restricting the range of values shown on the axes to approximately that actually found in the data, with the result the scattergram contains a lot of "white space" (horizontal and/or vertical areas in which no data points fall) and/or the scattergram was drawn in such a way that the data range of one variable was much more compressed than that for the other variable (so that the points are jammed together within a small area, making it difficult to plot them accurately and then to see whether there is an association between the variables);

(c) using DEV, rather than D2PV, to measure Democratic (or Incumbent) election success;

(d) drawing in a 45° (and/or $x = y$) line through a scattergram, which had significance in the "Sons' Height by Fathers' Height" and "Obama vote by Kerry vote" scattergrams, but which has no significance in any of the scattergram here (because the variables do not have *matching values)*;

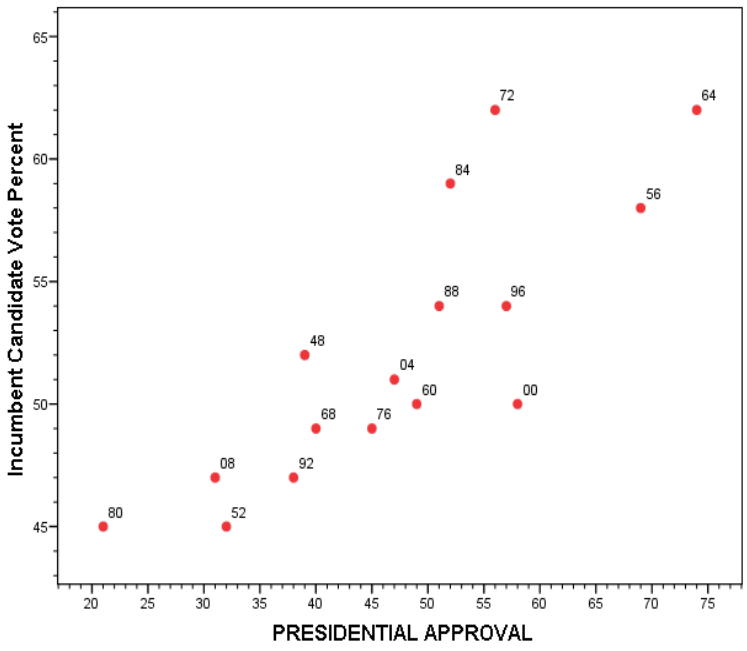
(d) not labeling the points by ELECTION YEAR as requested by Note 2, so I could not readily check whether your points were plotted reasonably accurately; and

(e) plotting points with *missing data* in such a way that it appears that, for example, in 1928 TURNOUT = 0 rather than TURNOUT data is missing entirely (a case with missing values on either variable cannot be plotted in a scattergram).

SPSS can produce very nice scattergrams, which are displayed below. First clear the Data Editor screen by clicking on File => New => Data. Next type in the data from the Problem Set or download it from the course webpage. Then use the SPSS chart facility: Graph => Legacy Dialogs => Scatter/Dot => Simple Scatter => Define.

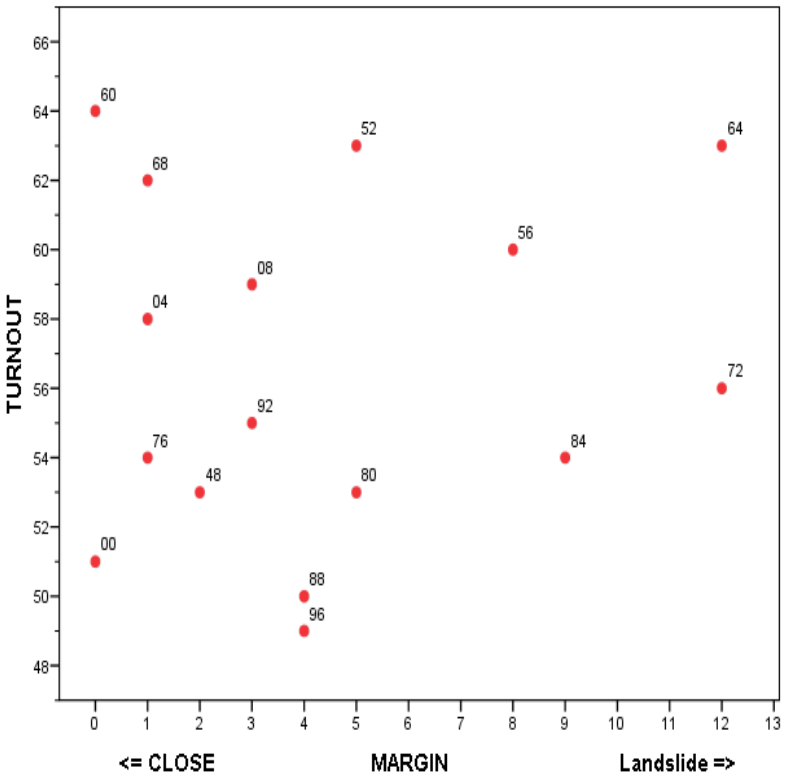
For each of the problems, you should produce (preferably using graph paper for reasonable accuracy, if done by hand) a scattergram that looks about like those that appear below, *with the variable axes appropriately labeled and scaled*. Note that I have scaled each axis so that the scattergram is only a bit wider than tall and is largely filled with data points.

#1. The hypothesis says that "high approval ratings boost a President's re-election performance," so the independent variable is LEVEL OF APPROVAL RATING and the dependent variable is INCUMBENT RE-ELECTION PERFORMANCE. POP measures the independent variable. D2PC gives the vote received by the Democratic Presidential candidate. In those elections years in which the Democrats held the White House (1948, 1952, 1964, 1968, 1980, 1996, and 2000) D2PC gives the value of the dependent variable; in those elections years in which the Republicans held the White House (1956, 1960, 1972, 1976, 1984, 1988, 1992, 2004, and 2008), 100 – D2PC gives the value of the dependent variable. Here is the appropriate scattergram.



There does appear to be a clear (and unsurprising) positive association between the two variables. (The calculated measure of association [correlation coefficient] is +0.83.)

2. The hypothesis says that "close elections stimulate voting turn-out," so the independent variable is CLOSENESS OF ELECTIONS and the dependent variable is LEVEL OF TURNOUT. We have a direct measure of TURNOUT in the data. We can derive a measure of CLOSENESS from D2PC by calculating the *absolute* deviation of D2PC from 50% for each year and calling the resulting variable MARGIN OF VICTORY. (Alternatively we can calculate the absolute difference between D2PC and 100 – D2PC [i.e., R2PC] and call the resulting variable MARGIN2. MARGIN2 = 2×MARGIN and produces an identical scattergram but with the horizontal scale running from 0 to 24 instead of 0 to 12. Or we could simply user WINNER'S VOTE %, ranging from 50% to 62%. When MARGIN (or WINNER'S VOTE) is *low*, the election is *close* and when MARGIN is *high*, the election is *not close* (i.e., MARGIN is a *negative* measure of CLOSENESS), so if the hypothesis is correct, we expect a *negative* association between TURNOUT and MARGIN. Here is the scattergram.

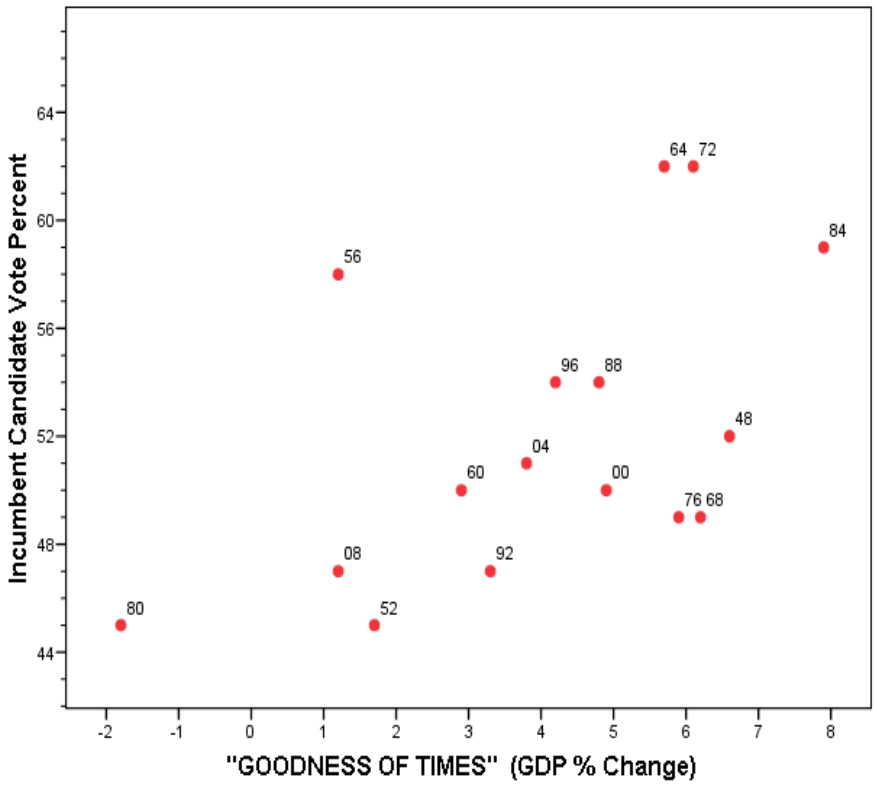


No association is visible to the naked eye, so there is no support for this hypothesis in the data. (The calculated measure of association [correlation coefficient] is +0.14.) *Note*: there are other more refined ways of formulating and testing this hypothesis.

Note 1. The analysis suggested above makes CLOSENESS/MARGIN the independent variable and LEVEL OF TURNOUT the dependent variable, i.e., CLOSENESS ===> TURNOUT. An insightful objection to this analysis is to note that, chronologically, people first decide whether to vote or not (and also how to vote) and only later, after the votes are counted, is the DEGREE OF CLOSENESS of the election revealed. So how can the variable CLOSENESS that "happens later" influence the variable TURNOUT than "happens earlier"? The best response to this objection is to refine the hypothesis to say that it is *EXPECTED* CLOSENESS that influences LEVEL OF TURNOUT. Then the question is how to measure

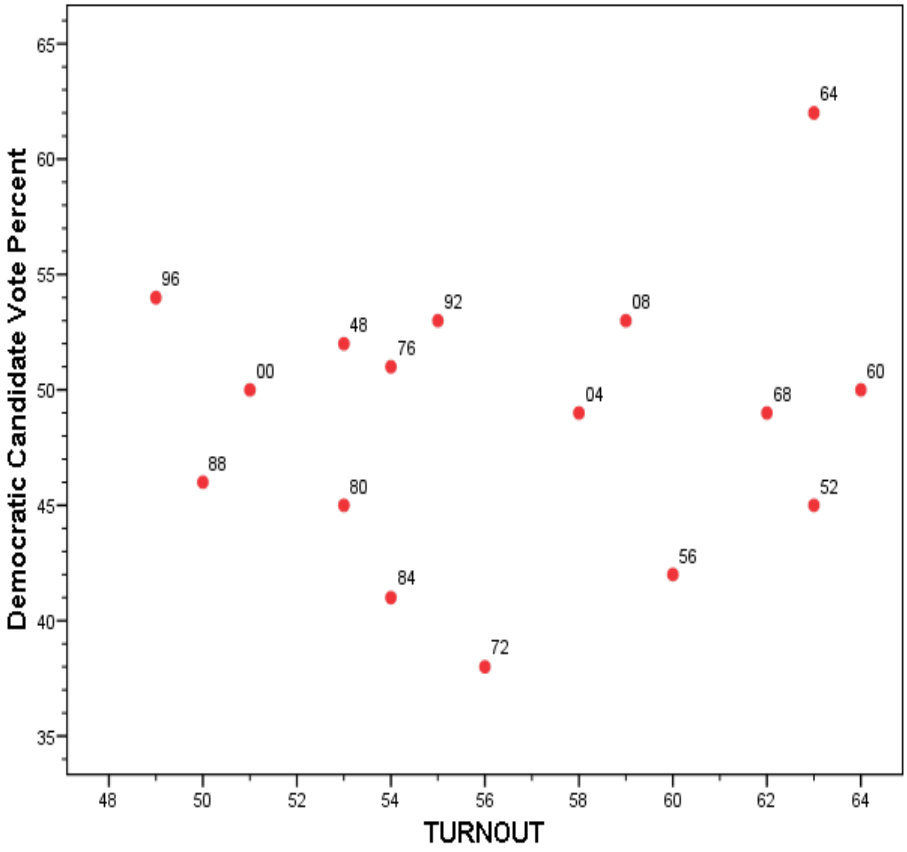
EXPECTED CLOSENESS; this might best be done not by computing MARGIN as indicated above but by using *pre-election polling* data, rather than the actual election results. In the immediate case, however, such polling data is not available. Thus, it can be said that (in the manner discussed earlier in the semester) we are using ACTUAL MARGIN as an *indicator* of (or *proxy* for) EXPECTED MARGIN — since, with few exceptions, elections that are expected to be close in fact are close (1980 was something of an exception) and elections that are expected to be landslides in fact are landslides (1948 was a famous exception — the candidate [Truman] expected to be at the wrong end of a landslide actually won a close election).

#3 The hypothesis says that "when times are bad, incumbent candidates are punished in elections," so the independent variable is DEGREE OF BADNESS OF TIMES and the dependent variable is again INCUMBENT RE-ELECTION PERFORMANCE. GDP provides an indicator of GOODNESS OF TIMES so if the hypothesis is correct, we expect a *positive* association with INCUMBENT RE-ELECTION PERFORMANCE (which variable we have already created for #1). Here is the scattergram.



There is a noticeable positive association, but it is not as strong as that in the scattergram for #1. (Correlation = +0.53.)

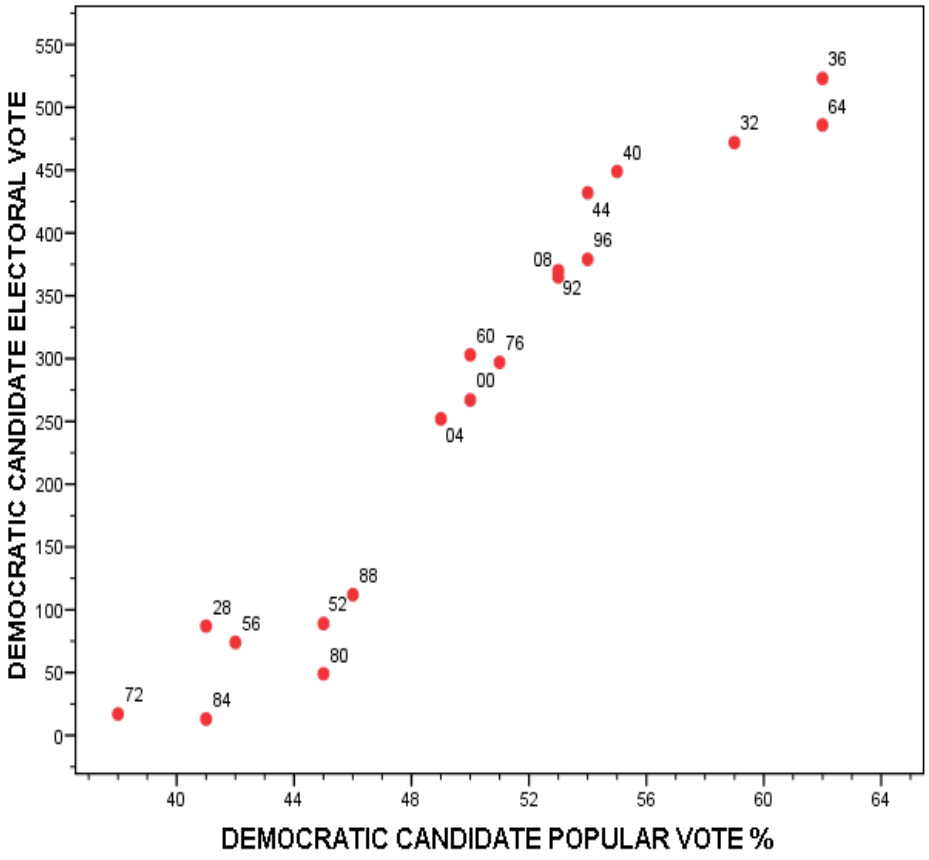
#4 The hypothesis says that "high turnout brings about Democratic victories," so LEVEL OF TURNOUT is now the independent variable and D2PC is the dependent variable. Here is the scattergram.



The message conveyed by the scattergram is that this data exhibits essentially no association between the two variables. (The calculated association [correlation coefficient] is less than +0.1, and this very slightly positive association results almost entirely from the 1964 data point [the greatest Democratic victory based on near-record high turnout]). Thus the hypothesis has no support in the data. (***Note:*** that there are other more refined ways of formulating and testing this hypothesis.)

OVER =>

5. Popular votes get translated into in electoral votes through the workings of the Electoral College, so DEM POP VOTE is independent and DEM ELECT VOTE is dependent. Here is the scattergram.



The scattergram shows very *strong positive* relationship is apparent in the scattergram, which should hardly be surprising. (The calculated association [correlation coefficient] is +0.96, about as high as a "real-world" association ever gets). Among other things, the data indicate that the Electoral College is likely to produce a "reversal of winners" only in very close Presidential elections (like 2000 and 1960).

Political Science Trivia. Under the original Constitution, ELECTORAL VOTES = SIZE OF HOUSE + SIZE OF SENATE. Since 1912, SIZE OF HOUSE has been fixed by law at 435, and of course SIZE OF SENATE = 2 × NUMBER OF STATES. Thus prior to the admission of Alaska and Hawaii in the late 1950s, ELECTORAL VOTES = 435 + 96 = 531. SIZE OF HOUSE was temporarily increased to 437 to give one seat to each of the new states, so in 1960 ELECTORAL VOTES = 437 + 100 = 537. After the 1960 census and reapportionment, SIZE OF HOUSE returned to 435 but the 23rd Amendment giving Washington D.C. three electoral votes came into effect, so since then ELECTORAL VOTES = 435 + 100 + 3 = 538. (The even number of electoral votes since 1964 allows a 269 to 269 electoral vote deadlock even when only two candidates win electoral votes.)