DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS

Posc/Uapp 816

MEASURES OF ASSOCIATION AND REGRESSION

I. AGENDA:
- A. Measures of association
- B. Two variable regression
- C. Reading:
  1. Start Agresti and Finlay, ***Statistical Methods for the Social Sciences***, 3rd edition, Chapter 9.

II. THE ODDS RATIO FOR I X J TABLES:
- A. For an I X J table there are I(I - 1)/2 **pairs** of rows and J(J - 1)/2 **pairs** of columns.
  1. Example: in the voting data table there are 2(1)/2 = 1 pair of rows, and 7(6)/2 = 21 pairs of columns.
- B. If we use a pair of rows, **a** and **b**, and a pair of columns, **c** and **d**, we can form an odds ratio

$$\hat{\Omega} = \frac{F_{ac}F_{bd}}{F_{bc}F_{ad}}$$

  1. There will be I(I - 1)/2 times J(J - 1)/2 odds ratios of this type.
  2. But this set contains a great deal of redundant information.
- C. So consider a subset of these odds ratios defined this way

$$\hat{\Omega}_{ij} = \frac{F_{i,j}\ F_{i+1,j+1}}{F_{i,j+1}\ F_{i+1,j}}$$

  1. There are (I - 1)(J - 1) odds ratios of this type. These are odds ratio based on cells in adjacent rows and adjacent columns.
- D. We call this a **basic** set of odds ratios.
  1. This is not, however, the only basic set. Consider another possibility:

$$\hat{\Omega}_{ij} = \frac{F_{i,j}\, F_{I,J}}{F_{I,j}\, F_{i,J}}$$

2. This set uses the last row and column as reference categories.

E. One way to summarize the form and strength of association in an I X J table is to construct a basic set of odds ratios.

1. Example: Consider once again the voting data.
2. We can form (2 - 1)(7 - 1) = 6 odds ratios.
3. Let's use the last column and row cell as the reference category.
   a. That is, we will compare odds of voting for each level of education with the highest level as the reference group.

Education/ voted	1	2	3	4	5	6	7	Totals
Yes	36	80	328	222	115	249	142	1175
No	30	57	147	55	26	32	12	359
Totals	66	137	475	277	141	260	154	1534

4. In other words, let i run from 1 to 2 - 1 = 1 and j from 1 to J - 1 = 6.
5. So we form odds ratios such as

$$\hat{\Omega}_{11} = \frac{(36\ \ 12)}{(30\ \ 142)} = .1014$$

$$\hat{\Omega}_{12} = \frac{(80\ \ 12)}{(57\ \ 142)} = .1186$$

$$\hat{\Omega}_{13} = \frac{(328\ \ 12)}{(147\ \ 142)} = .1886$$

etc.

$$\hat{\Omega}_{16} = \frac{(249\ \ 12)}{32\ \ 142)} = .6576$$

6. The next table contains all of these basic odds ratios and their logarithms.

j	Odds ratio	Log odds ratio
1	.1014	-2.289
2	.1186	-2.132
3	.1886	-1.668
4	.3411	-1.075
5	.3738	-.9840
6	.6576	-.4192

7. Interpretation:
   a. Look first at the first category of education (j = 1) and the last category (j = 7). The odds ratio is .1014 (log odds ratio = -2.289). This means that the odds of being a voter are one tenth of the odds of someone in the highest category of education being a voter.
   b. Now look at the second category of education (j = 2). Now the odds are still only about 1/10 (i.e., .1186) the odds of people in the last category of education.
   c. By the time we reach the sixth category (j = 6) of education, the odds of being a voter are about 6/10--slightly more than half--the odds in the last category.
   d. We see, therefore, that as education increases the odds of voting (compared with the most educated) increase.
   e. There is still quite a gap between the highest and next to highest categories of education: the odds of being a voter are just sixth tenth those in the highest category.
   f. It does not appear that the increase in odds is linear: that is, there is not an equal increase in the odds as we move up one category of education at a time.
   g. Another way of viewing the pattern is to notice that as one moves "down" the table (i.e., from lower to higher levels of education), the odds ratios and log odds ratios approach their "null" values of 1.0 and 0.0, respectively. This means that the odds of voting are getting closer to the odds in the last (highest) level of education, as we might expect.
8. It would be nice to summarize all of this information in a single number, and such indices are available. Still, ransacking a table like this one can be informative.

III. SINGLE MEASURES OF ASSOCIATION FOR I X J TABLES:

A. Agresti and Finlay, ***Statistical Methods for the Social Sciences***, 3rd edition, pages 266 to 285 discuss quite a few alternatives.

B. Note these points, however:

1. So-called ordinal measures of association remain common in the social science and policy literature but are not nearly as popular tools as they used to be.
    a. We'll discuss a couple of them briefly when talking about the correlation coefficient.
2. "Proportional reduction in errors" (PRE) measures and interpretations have largely disappeared from the literature.
3. Occasionally one will see a measure of association based on the observed chi square statistic. But they are hard to interpret and are not used too commonly these days.

IV. TESTING INDEPENDENCE IN 2 X 2 AND OTHER TABLES:

A. Read Agresti and Finlay's account of testing for independence in 2 X 2 tables, pages 263 to 266.
    1. In general, use so-called "exact" tests.

B. Be careful applying the chi square test to tables having many empty or nearly empty cells.
    1. Indeed, if there are more than 2 cells with expected frequencies less than 5, the chi square test may not give accurate results.
    2. See Agresti and Finlay page 265.

V. INTRODUCTION TO REGRESSION: A TWO VARIABLE EXAMPLE:

A. To motivate the discussion let's consider a substantive problem.

B. What explains "socially undesirable" behavior such out-of-wedlock births? (Needless to say, what is "undesirable" is a matter of opinion. Nevertheless, let's stick with this example that still troubles people.) Charles Murray, in his book ***Losing Ground***, proposes that part of the increase in illegitimacy rates can be traced to the welfare system that inadvertently offers inducements to change behavior. In particular, welfare benefits are so generous and rules for their receipt are so lenient that unwed mothers have no reason to marry the fathers.

C. One implication of this theory is that illegitimacy should be highest where welfare benefits are greatest.
    1. Murray himself uses time series data and intuition to test his proposition. A better test would be to examine behavior first hand. We can't do that with the available data so as an approximation let's look at aggregates, namely states.

D. Here (on the next page) are some data for a "sample" of 19 American states.
    1. Note that congressional action in 1996 eliminated the Aid to Families with Dependent Children "AFDC" in part for the very reason Murray and others' work suggested: welfare rewarded irresponsible behavior.

State	Illegitimacy Rate 1980	Average Monthly AFDC Payments 1980	Percent of Families Below Level, 1979
AL	221	110	14.8
AR	204	145	14.9
CT	179	358	7.7
DE	241	227	8.9
GA	231	133	13.2
IL	225	277	8.4
KS	122	271	7.4
ME	138	233	9.8
MA	156	341	7.6
MI	161	379	8.2
MO	176	217	9.1
NV	134	207	6.3
NM	160	185	14.0
ND	92	277	9.8
OR	147	318	7.7
SC	203	107	13.1
TX	133	109	11.1

Table 1: Illegitimacy and AFDC Payments

2. These data come from the ***Statistical Abstract, 1985*** (pp. 381 and 457) and ***Monthly Vital Statistics Report***, November 30, 1982, p. 24.

E. They are "operational indicators" of the concepts "out-of-wedlock births," "poverty," etc.

VI. FIRST STEP: THE SCATTERPLOT

A. Let's look at the relationship between illegitimacy and welfare "generosity," AFDC payments.

B. Always plot data first.

1. Using MINITAB

a. Go to **Graph** and then **Plot**

C. Each point is state. That is, scores on **both** illegitimacy and AFDC payments have been plotted.

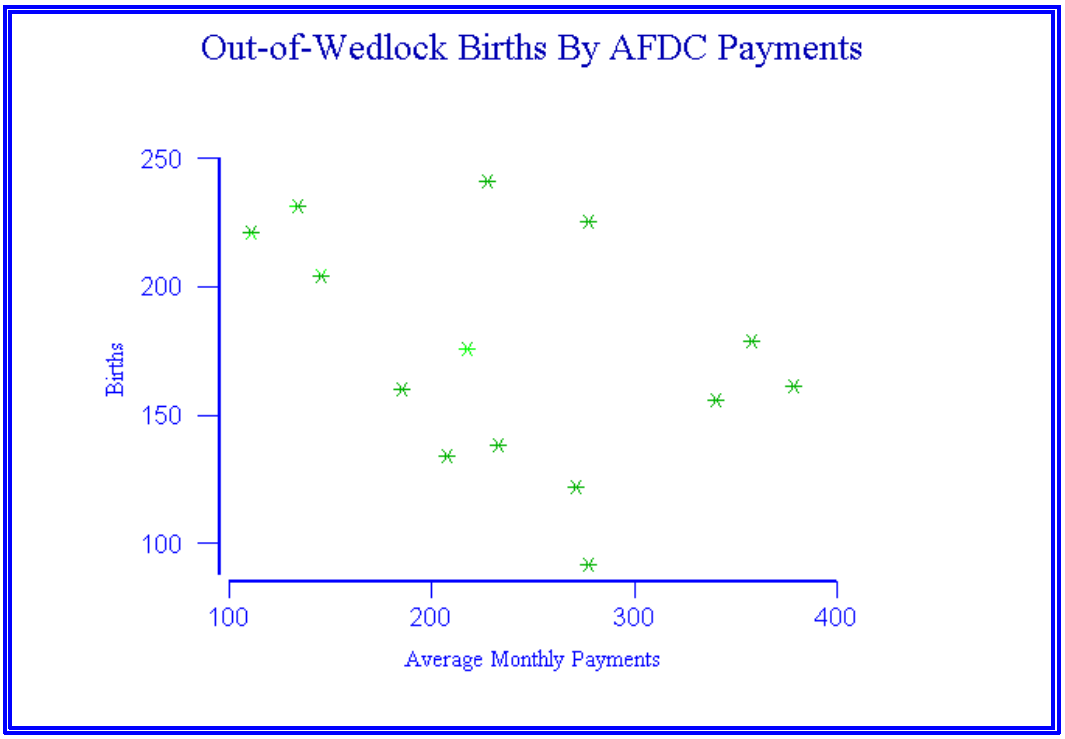

Figure 1: "Scatterplot" of Births by AFDC Payments

1. The pattern suggests that contrary to expectations as a state's AFDC level increases, its out-of-wedlock births drop.
2. But the figure also tells us that the relationship is very weak.
3. In fact it appears as though there are two cluster of points. We may investigate this further.

D. Tips:

1. Most statistical program packages have enhanced graphics capabilities, which means that besides examining data visually from numerous perspectives one can also annotate the graphs.
2. You should title your graphs and label the axes so that it is clear to the reader what is being plotted.
3. If time permits, we'll see some examples using MINITAB.

VII. REGRESSION ANALYSIS:

A. Overview:

1. The next step is to propose a model that might "fit" these data.
2. The simplest model is a "linear" equation:

$$Y_i = {}_0 + {}_1 X_i + {}_i$$

where the parameters, **B**'s show how X affects Y.

3. For the moment, the $\mathbf{B}_1$ shows how a one-unit change in X changes Y. More precisely, as X changes on unit (1 percent, say) Y will change $\mathbf{B}_1$ units.

B. Regression analysis involves a several activities:
   1. Estimating the parameters of the model (i.e., "guessing" their numerical values based on a sample);
   2. "Testing" their significance;
   3. Assessing the goodness of fit of the data to the model;
   4. "Diagnosing" problems and looking for improvements.
      a. Transformation to make the data consistent with the requirements of regression analysis.
      b. As an example, the relationship between Y and X may be "curvi-linear" and/or distorted by "outliers," values that do not "fit" the general pattern of relationship seen among most of the cases. Such contingencies may throw off conventional analysis so we need methods to make sure that findings are not misleading.

C. For the most part we will do two different types of regression:
   1. Conventional "least squares" which is what everyone reads about in the usual literature and which is supported by most computer program such as SAS and SPSS.
   2. Exploratory or robust regression which may in a few cases have some advantages over the least squares version.

VIII. GEOMETRY OF LINES:

A. To understand the linear model let's review some simple math.

B. The equation of a linear (straight line) relationship between two variables, Y and X, is

$$\boldsymbol{Y = a + bX}$$

C. Interpretation:
   1. **a** is the <u>intercept</u>, that is the value of Y when X equals zero. If the line is graphed on an Y-X coordinate system (see below), then a is the point where the line crosses the Y axis.
   2. **b**, called the <u>slope</u>, is the amount of change in Y for a one-unit change in X. It's measured in units of the dependent variable, Y, <u>but</u> its numerical value depends on the measurement scale: if X is measured in dollars, then b will equal some particular value, but if the scale is <u>thousands</u> of dollars, b

will have a different value.

3. The figure on the next page shows a picture of the graph of a linear relationship. Notice that the graph is a straight line.
4. The linear relationship described by this graph is:

$$Y = a + bX = 2 + (2)X$$

D. In other words, the intercept of this particular model is 2 and the slope is 2.0.

E. The numbers a and b are called regression parameters; note that they are constants whereas X and Y are variables. The parameters show you how X affects or at least is connected to Y.

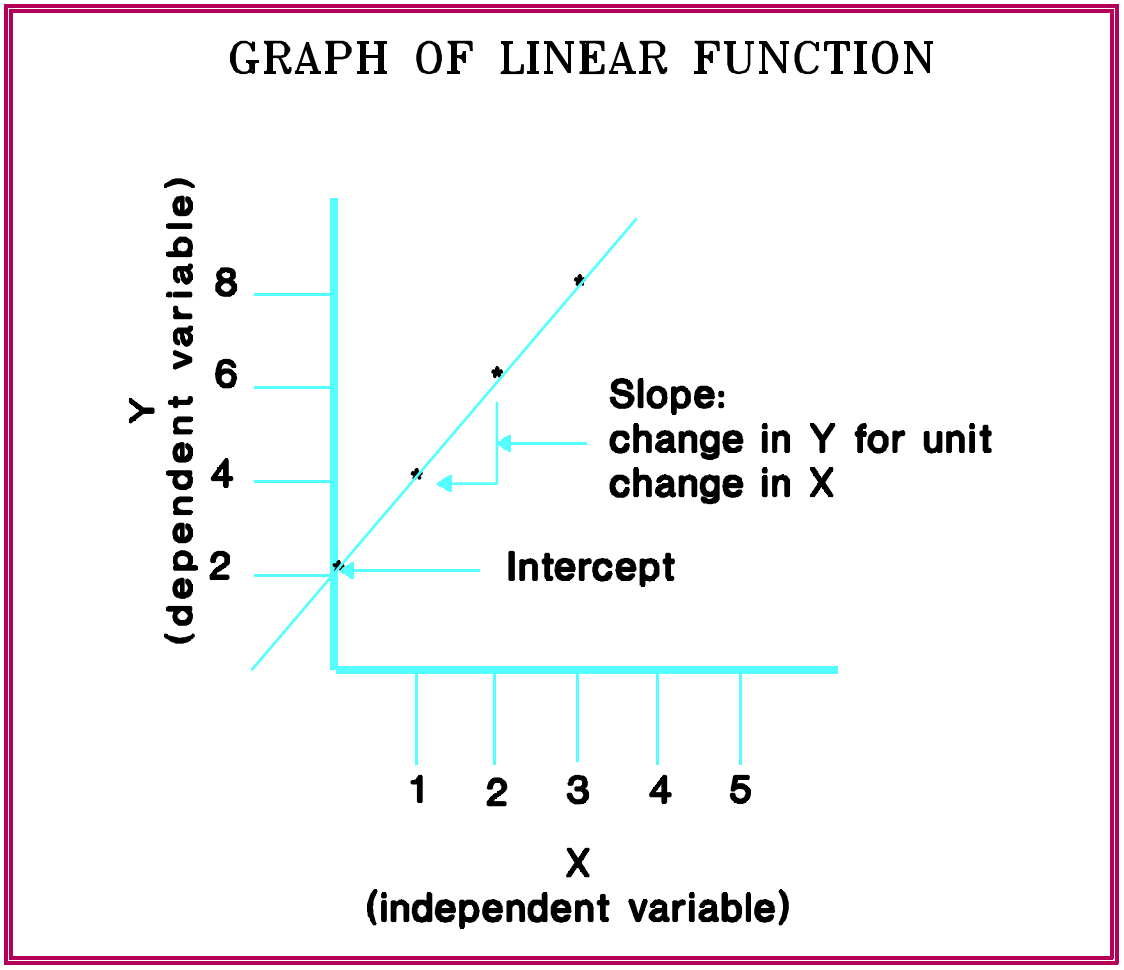

Figure 2: Graph of Linear Equation

IX. NEXT TIME:
- A. Interpretation of regression parameters
- B. Principle of least squares

Go to Notes page

Go to Statistics page