DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

DATA ANALYSIS

I. AGENDA:

A. Data analysis: some tips

1. Necessity of “theory”
2. “Operational indicators” and measurement errors
3. Preliminaries
4. Model building
5. Corrections: iterative model building.

B. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition, Chapter 14 again.

1.

II. EXAMPLE:

A. Here’s an example much like the one used for Assignment 12.

B. I obtained the data and documentation from the “Data Story Library” at Stalib, an excellent web-based source of information for people interested in statistics and data analysis.

1. It’s located at Carnegie Mellon University and can be access through the course web site.
2. Although somewhat dated, it is still useful for our purposes.

C. The data consist of variables for 47 states. The “response” or dependent variable is crimes known to the police per 1 million inhabitants.

D. Variables:

1. Crime rate: # of offenses reported to police per million population
2.Age: The number of males of age 14-24 per 1000 population
3. Indicator variable for Southern states (0 = No, 1 = Yes)
4.Mean # of years of schooling x 10 for persons of age 25 or older
5. 1960 per capita expenditure on police by state and local government
6. 1959 per capita expenditure on police by state and local government
7. Labor force participation rate per 1000 civilian urban males age 14-24
8. The number of males per 1000 females
9. State population size in hundred thousands
10 The number of non-whites per 1000 population
11 Unemployment rate of urban males per 1000 of age 14-24
12. Unemployment rate of urban males per 1000 of age 35-39
13.Median value of transferable goods and assets or family income in tens of $
14.The number of families per 1000 earning below 1/2 the median income

E. Our task, like the one in Assignment 12, is to build a "model" for the crime rate.

1. That is, we want to explain variation in Y, the number of crimes per 1 million inhabitants.

III. THEORY:

A. Here are some tips I find helpful

1. Data "never speak for themselves."

i. Meaningful efforts to explain variation have to rely partly on "theory," a set of expectations and possibilities that existing scholarly literature, conventional wisdom, common sense, and the like suggest.

ii. As often as not "correlating" everything with everything else will lead to confusion and misinterpretation.

2. Adopt a plan of action based on the "theory" but of course be flexible.

i. On paper a lot of statistical inference assumes that (statistical) hypotheses are stated **before** one "sees" the data.

ii. I doubt that very many practicing social or policy scientists strictly adhere to this guideline. (After all, it's a guideline.)

iii. Still, by thinking ahead you should be able to reduce the number of relationships to be explored and the number of tests to be conducted.

iv. Serendipity as a research strategy is, I think, overstated.

3. Two principles:

i. "Realism"

ii. Parsimony: the simple or two equal or nearly equal models is to be prefered.

B. Example:

1. To build a model of crime I am going to follow what I take to be conventional wisdom, political folklore, and perhaps a bit of scholarship.

i. Note: I fully appreciate the simplicity of this analysis.

2. In particular, I guess crime is mostly explained by the presence and absence of various "opportunities."

i. For example, I've heard that crime is somewhat a behavior of youth and so the more young males that are around the higher the crime rate, other things being equal.

ii. As another example, nearly everyone believes that crime emerges to the extent that people are "trapped" in "impoverishing" circumstances.

iii. I also wonder if crime doesn't have a regional "bias."

1) The murder rate certainly does, doesn't it?

C. Consequently, I will look at the relationships between crime and

1. The size of the young male population;

2. The percent non-white in the state;

3. Wealth;
4. Region.

D. A model and structural equations.

1. All of this suggests that something like crime cannot be explained by reference to a single causal factor that operates in a vacuum.
2. Thus, to model it adequately one most likely will need a system of equations to describe the mutual interactions of various sets of variables.
    i. Sets of equations that involve “endogenous” and “exogenous” variables are called structural equations.
        1) You will also hear the term “simultaneous” equation model: unlike the models we have analyzed, these involve several equations that supposedly hold at one and the same time and for which one can find estimators of the various parameters.
    ii. These models frequently make specific assumptions about measurement errors and the like.
    iii. The seldom treat all of the error terms as simple random disturbances, but instead try to “model” error processes.
    iv. They can encompass “one-way” or reciprocal causation.
3. Unfortunately we have not had time to develop the analytic tools necessary to study them in detail.
    i. That work follows in a third semester of applied statistics.
4. Consequently, we’ll stick with the familiar one equation model and simply keep in mind its (over?) simplicity.

IV. MEASUREMENT AND OPERATIONAL INDICATORS:

A. Theoretical constructs versus observed variables.

1. Almost always our “theories” are not directly “measurable” or testable.
2. We rely on empirical indicators to stand in for the unmeasured, usually unmeasurable concepts.

B. The problem of measurement error.

1. Epistemic correlation: the “relationship” between concepts and indicators.
    i. Validity and reliability.
2. Regression analysis assumes that both independent and dependent variables have been measured without systematic error.
3. For example, in the model $E(Y) = \beta_0 + \beta_1X$, we assume that X’s relation to the “true” or underlying variable is

$$X_i = x_i + \varepsilon_i$$

where x_i is the "true" value,

X_i is the measured value,

and ε_i is random.

4. If this assumption doesn’t hold and there is a more complicated error structure, estimators of parameters may be biased.

5. Random measurement error in Y, the dependent variable, will not produced biased estimators but will show up in the residuals and lead to a less good fit.

C. Again, since we do not have the time or tools to explore measurement errors, we’ll have to assume that our measures of crime and social-economic conditions are adequate.

V. PRELIMINARIES:

A. Explore the variables via descriptive statistics and graphs:

1. Magnitude of variation in Y and X’s.

i. A dependent variable that does not vary cannot be explained.

ii. An independent variable that does not vary cannot explain.

iii. You would be amazed at how important these seemingly trivial statements are.

2. Shape of empirical distributions.

3. Outlying observations.

B. Preliminary analysis of the crime data suggests that a couple of the independent and certainly the dependent variable should be transformed to achieve better fits.

C. Bi-variate relationships:

1. Correlation matrix:

i. Dependent versus X’s

ii. X by X correlation matrix.

1) Values near .8 or .9 suggest multicolinearity.

2) But this assessment is not fool proof.

2. Bi-variate plots.

i. Nature of relationships and possible transformations.

1) Recall the ladder powers.

ii. Outliers and leverage points.

VI. MODEL BUILDING:

A. “Mechanical” versus “guided” modeling.

1. Most computer programs such as MINITAB have regression options that

build models piece by piece by adding or deleting a variable according to whether it makes a statistically significant "contribution."

i. These procedures are usually called "stepwise" regression

2. See Agresti and Finlay, Statistical Methods for Social Sciences, 3rd edition, pages 529 to 534 for an excellent discussion of “forward” and “backward” automatic selection procedures.

3. We’ll use a more intuitive guided approach and rely on diagnostic tools.

B. Regression diagnostics

1. Residuals plots

2. Variance inflation factor.

C. Measures of fit.

VII. ITERATION:

A. Measures of fit.

1. Coefficient of determination.

2. Large residuals and leverage points.

3. Tests of significance.

B. Presenting results:

1. Usually a published paper shows mainly the final results, not the intermediate steps that led to the final model.

i. The diagnostic steps can be described in notes.

ii. If they are essential to establishing a controversial point or perhaps challenging someone else’s findings, the analysis should be more detailed.

2. Tables that contain models should include (at a rock bottom minimum):

i. Parameter estimates (always presented)

ii. Standard errors of estimates (frequently presented).

iii. Attained probability (seldom presented).

iv. “Effective” N: the number of cases upon which **this** model is based, not the number at the outset of the study. (Frequently presented.)

v. Measures of goodness of fit. (Frequently presented.)

3. Most journals will want the “level” of significance reported.

4. Very important: although it is seldom done, I think authors should include a citation showing where one can find:

i. The raw data

ii. Coding and measurement conventions.

iii. The idea is to make your results independently verifiable.

VIII. NEXT TIME:

A. Complete assignments

Go to Notes page

Go to Statistics page