Department of History
UNIVERSITY OF WISCONSIN -- MADISON
Semester I, 1991-92

History 795
Quantitative Methods for Historical Research

Tuesday, 10 - 12 Thomas J. Archdeacon

All historians gather data. In recent years, however, a sub-group of historians, composed of researchers who think of themselves as quantitative social scientists, has emphasized the importance of a particular kind of data. Their focus has been on those categories of information that measure, for each member in a group of people or in a collection of political or social units, various fundamental characteristics. Their presumptions are that common patterns in the data pertaining to the individual members will illustrate the true character of a collectivity, that internal dissimilarities will expose fissures undermining its unity, and that systematic comparisons between groups will reveal the essential differences separating them. They believe that, within certain constraints, the observation of patterns can often lead to an understanding of causal relationships. Moreover, what separates these quantitative, social scientists from others in their discipline, including those who use different techniques to study the same groups, is an explicit reliance on theory and on computer-assisted statistical analysis.

The preceding paragraph touches on sensitive issues. It describes the immediate task of historical research as the gathering of a vaguely defined kind of information called "data." It classifies the prime subjects of investigation as collectivities. It identifies history, or at least a legitimate subdivision of the discipline, as a social science that relies on quantitative evidence. It implicitly equates analysis with the study of variation and makes some association between the observation of patterns and the understanding of causation. Finally, it implies that hypotheses about patterns and causation ought to be evaluated according to the criteria of formal statistics.

History 795 addresses the several points made above and invites students to consider the implications of them. Beyond that, the course focuses on the statistical techniques that are at the heart of the quantitative social science approach to history. In particular, it examines the family of analytical techniques associated with the "general linear model," which has been the most important interpretive paradigm in quantitative research in the social sciences. The explication contains enough detail that students will be able not only to assess critically scholarship with major quantitative components but also to build skills necessary for applying statistical techniques in their own research.

Reading Assignments:

In teaching 795 this semester, I intend to rely on a manuscript that I have been preparing over the last few years. The syllabus refers to this work as the ***Text***. We need to go through the text at the rate of one chapter per session. Some of the subject matter is difficult or at least will demand your full attention. In order to understand what we will be discussing each week, you will have to read the assigned chapter of the text *prior to* the class. I shall do my best to present the material at a pace that the class as a whole can maintain. I shall not, however, slow down to accommodate anyone who has not come prepared.

As you read the text, you will find citations to secondary works that make use of the techniques being discussed. Copies of those book chapters and articles will be on reserve at the History Department Library. I shall not make reading those pieces a requirement, but perusing as many of them as possible will undoubtedly benefit you.

Computers:

Learning how to use a computer to manipulate data and to perform statistical analyses on them will be essential for your work in the course. You will have free access to the computers located at the Social Science Microcomputing Laboratory (SSML: 3218 Social Science Building) for computational work related to History 795. For a modest printing fee, you will also be eligible to wordprocess at the SSML whenever the machines there are not being used for classes or statistical analyses.

The staff of the SSML will provide an introduction to the facility and its equipment during the second hour of Meeting #1. I shall provide additional instruction and demonstrations during subsequent sessions. In addition, you will find in the History Library several pieces that I have written to help you understand the computer hardware and software to be used in this class. You will need to practice outside of class in order to reinforce what you have been shown. Working with a partner can be an effective learning technique. Should you encounter intractable problems, the SSML staff will give you additional help. Likewise, the staff of the Data Program and Library Service (DPLS: 3208 Social Science) will assist you to acquire data and put them in usable form.

Grading:

My assumption is that you will prepare for classes and participate in them. Lack of preparation or of participation will obviously hurt your cause. Beyond that, your performance on two written exercises will determine your grade.

I plan to provide you, on 19 November at Meeting #12, information about a data set that will be available in machine-readable format at the SSML. You will be asked to use the data to answer a set of questions and to analyze an historical problem. Doing the exercise will require you to use the Statistical Package for the Social Sciences (SPSS/PC+) to manipulate the data and to carry out a variety of statistical measures and tests. You

will be expected to present your results in a ten-page paper to be submitted on 3 December at Meeting #14. There is no class scheduled for 26 November. Your performance on this requirement will determine approximately one-third of your final grade.

You will also be expected to write a twenty-page paper on a research topic of your choice. Quantitative analysis must play a fundamental role in the research for and the presentation of the essay. You may work on data collected for a project of your own or devise a study to be based on data available through DPLS. If you choose the latter alternative, you may work either with data generated by government agencies, survey organizations, or similar bodies, or with data archived by an earlier scholar after the completion of his or her own research. You will be able to discover such data sets by examining the DPLS catalog and the ***Guide to Research and Services*** produced by the Inter-university Consortium for Political and Social Research (ICPSR).

In approaching the final paper, you should choose a topic that is related to a substantial historical issue. You should identify a set of questions, the answers to which will advance our understanding of the problem of interest. Of course, those questions should be amenable to quantitative analysis, and the data set used should be likely to yield the needed answers. You should employ statistical methods appropriate to the data. Complexity for the sake of complexity is not the goal, but those who successfully address difficult problems that demand the subtle application of advanced techniques will have an advantage over those who do not demonstrate such abilities. Finally, effective written presentation of your findings is important. Turning quantitative research into cogent, readable prose requires considerable effort.

Your performance on the second paper will determine approximately two-thirds of your final grade. The deadline for submitting those essays is 16 December 1991. Some of you, however, may want to incorporate in your papers techniques to be discussed late in the term. Students in that position may seek permission to defer completion of the essay, but they should be willing to accept a temporary grade of "I" if the delay required is more than minimal.

Schedule

Meeting #1: 3 Sept.
Introduction to the Course

The first hour of this meeting will enable us to introduce ourselves and to discuss mutual expectations about the course. For the second hour, we shall go to 3218 Social Science for an introduction to the SSML. Most of you will not have the Text by the time of Meeting #1, but please read its Preface and first chapter as soon thereafter as possible.

Text: Preface and Chapter 1

Meeting #2: 10 Sept.
Measures of Central Tendency

Studying a population or sample largely involves gathering and analyzing data about characteristics that pertain, or potentially pertain, to each of its members. Depending on the collectivity under scrutiny, the researcher may want to record the racial background of each resident of a city, the number of inhabitants in each county of a particular state, or the party affiliations of the 100 members of the U.S. Senate. Traits such as those are known as variables.

This meeting will be concerned with the topic of descriptive statistics, which are techniques for measuring and summarizing, in as little as a single number a critical piece of information contained in data about a variable. The mean or average value is perhaps the fundamental example of a descriptive statistic. Others, less well known to most persons, include the median, mode, variance, and standard deviation.

Chapters 2 and 3 of the ***Text*** contain information that is elemental but of fundamental importance. At least in an informal way, you may know parts of the material already. Therefore, reading both chapters in a single week should not be too taxing. Do not take Chapters 2 and 3 too lightly, however; you will need to develop a thorough understanding of them.

Text: Chapters 2 and 3

Meeting #3: 17 September

The Normal, Student's, and Chi Square Distributions

The normal distribution has enormous importance for statistical analysis. It not only allows researchers to be precise in the measures used to summarize data collected on a whole population; it also makes possible accurate predictions about the values of such parameters from data gathered about a properly drawn sample of those cases. Likewise, two distributions derived from the normal -- "Student's t" and the "chi square" -- play critically important roles in those analyses. The goal of this meeting is to lay the theoretical groundwork necessary to make those points clear.

Text: Chapter 4

Meeting #4: 24 September

Sampling

Sampling makes people nervous. For historians, it is a concept that runs counter to their training and personal inclination to seek the elusive and unique. They naturally worry about the consequences of a sample that misses a key case, an important personality, or a critical document. For just about everybody, sampling has an aura of legerdemain. How can one make generalizations about a phenomenon after examining only a small proportion of the cases involved in it? This meeting addresses that question and presents the rudiments of the most popular sampling strategies -- simple random, systematic, stratified random, and cluster -- and explains why one or the other may be preferable in a given set of circumstances.

Text: Chapter 5

Meeting #5: 1 October

Correlation Analysis

The examination of variables as individual entities rarely satisfies historians, who inevitably want to analyze them in combinations, in order to delve into the potentially causal relationships among such characteristics as class, age, gender, ethnicity, religion, political behavior, and social attitudes. Discerning patterns of association or correlation in the fluctuations of the values of pairs or larger sets of variables is a first step toward that goal. This meeting focuses on correlation measures designed for use with numerical information.

Text: Chapter 6

Meeting #6: 8 October

Statistics for Nominal Variables

Some of the data of greatest interest to social scientists can be measured only at the nominal level. Sex, religion, and race, for example, are fundamental forces that researchers dare not neglect. Likewise, variables representing political party membership and votes on legislation routinely involve one or more categorical variables. This meeting provides a general description of the use of tables to summarize and present data about such nominal data. It focuses especially on the subset of statistics that offer the best insights into the general approach of correlation and regression and that pave the way for later treatment of the incorporation of categorical variables into multivariate analyses.

Text: Chapter 7

Meeting #7: 15 October

Linear Regression Analysis

Regression analysis is one of the most powerful and popular statistical tools for examining linear relationships involving two or more variables. Its basic premise and strategic approach are distinct from those behind correlation analysis. Simple regression analysis, which involves only two variables, assumes that one of them is dependent, that the other is independent, and that the values of the former can be predicted on the basis of changes in the values of the latter. Multiple regression analysis extends the principle beyond two variables to situations in which the values taken by the dependent variable are assumed to respond to a combination of effects exerted by several independent variables.

Text: Chapter 8

Meeting #8: 22 October

Evaluating the Regression Equation

In performing a regression analysis, the researcher proceeds from the hypothesis that the dependent variable can be expressed as the sum of the weighted effects of a set of independent variables. According to the criteria outlined in the preceding meeting, he or she estimates values for the intercept of that linear equation and for the weights associated with each of those variables. Those calculations, however, are only the first step in the execution of a regression analysis. Using the results, the researcher, through a series of tests, must make judgments about the extent to which all or any of the variables in the model contribute to an understanding of the dependent variable under investigation.

Text: Chapter 9

Meeting #9: 29 October

Regression and Explained Variance

The tests discussed in Meeting #8, albeit essential parts of regression analysis, have minimal meanings. Even when they are "significant," they merely indicate that at least one independent variable, when properly weighted by its coefficient, is a better predictor of the value of individual observations of the dependent variable than is the mean of that dependent variable. What the researcher needs are ways by which to evaluate how good an explanation the model in question provides for the case-to-case fluctuations in the value of the dependent variable. This meeting focuses on such techniques, which build on aspects of the regression analysis that have already been encountered. The procedures, however, involve more than a mechanical application of rules; interpreting the relevant statistics requires subtlety.

Text: Chapter 10

Meeting #10: 5 November

Nominal Independent Variables in Linear Regression

Up to this point, the examination of correlation and regression will have focused on the ideal situation in which the researcher is able to measure both the dependent variable and the full set of independent variables at the interval level. Unfortunately, many of the problems of greatest interest to historians, as well as to other social scientists, do not fit that paradigm. The scholar may want to build a model that has a categorical dependent variable, such as a "yes" vote or a "no" vote on a bill, or that uses one or more independent variables representing ascriptive characteristics like sex and religion. Indeed, nominal data can appear on both sides of the equal sign, as dependent and independent variables in a single equation.

Incorporating categorical variables complicates statistical analyses, and especially those of the multivariate kind. Each variation on their use requires at least minor, and sometimes major, modifications of the procedures thus far discussed. Because the use of nominal data along with interval-level independent variables in regressions involving interval-level dependent variables is so basic to historical investigation, the discussion begins with that topic.

Text: Chapter 11

Meeting #11: 12 November

Residuals and Transformations

Like most statistical procedures, regression analysis is based on a set of specific assumptions. When those do not hold for the data set being examined, a variety of problems can arise. The estimates for the parameters may be misleading, with the result that the researcher misses or misinterprets a relationship. And, even when parameter

estimation for a particular body of data is not seriously affected, the rationale underlying the extension to a population of judgments made about that sample may be invalidated.

Fortunately, researchers can use certain of the results yielded by regression analysis to uncover violations of the procedure's assumptions. Most important, they can look for telltale patterns of error in the predictions that the model makes about the values of the dependent variable. Such inspection may provide clues to altering models or variables in ways that will compensate for trouble-making violations.

Text: Chapter 12

Meeting #12: 19 November
Regression Models and Causation

How can a researcher select for his or her regression equation the set of independent variables that provides the most complete but elegantly parsimonious predictive model? Testing every potential model is impractical, especially when anything more than a few variables are involved. For naive users, relying on the computer to select variables and build equations may seem like a responsible course of action. Unfortunately, automatic selection procedures can be arbitrary and are not magical guarantors of proper decisions. This meeting examines the general problem of the connection between regression analysis and causal explanation. It also discusses the specific approach known as "causal modeling," which uses the techniques of correlation and regression to make explicit, and to test, theories about cause-and-effect relationships in sets of variables.

Text: Chapter 13

Meeting #13: 26 November
NO CLASS

Meeting #14: 3 December
Logistic Regression

Many questions posed by social scientists imply dichotomous divisions. Does a voter back or oppose a particular candidate? Does the woman work inside or outside the home? Does the peasant emigrate or stay? The researcher's natural desire will be to use those dichotomies as dependent variables in regression equations, for which the independent variables will be sundry demographic and socioeconomic measures. Unfortunately, categorical dependent variables violate key assumptions underlying the "ordinary least squares" criteria normally used for solving regression equations. The subject of this meeting, "logistic regression," is a technique developed for addressing the specific difficulties generated by nominal dependent variables.

Text: Chapter 14

Meeting #15: 10 December

Log-Linear Analysis

Investigators often face problems that involve only variables measured either at the nominal level or with the roughest of rankings. Twenty years ago, situations like those seemed to preclude multivariate analyses, and researchers had to content themselves with the piecemeal dissection of tables using statistics like those discussed in Meeting #6. Since then, statisticians and social scientists have developed a set of "log-linear" techniques specifically designed for the multivariate analysis of nominal data. Relatively few historians have made use of the approach, but its capabilities hold great promise for all who must deal with crudely measured data.

Text: Chapter 15