Date submitted: 18/06/2010

Using Web-based Software to Promote Data Literacy in a Large Enrollment Undergraduate Course

Harrison Dekker
UC Berkeley Libraries
Berkeley, California, USA

Meeting: **86. Social Science Libraries with Information Literacy**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract:

For the past two years, UC Berkeley's Library Data Lab has played a key instructional role for a large enrollment undergraduate Economic Demography course. In addition to providing assistance to students in locating and evaluating data sets for their term projects, the Lab provides training and support in the use of SDA, a web-based data analysis application freely available on a growing number of data archive websites. This paper will focus on the rationale for using SDA, particularly how it helps us achieve certain data literacy goals, and offer practical advice to those interested in pursuing this type of instruction at their own institutions.

In their 2007 article, “Incorporating data literacy into undergraduate information literacy programs in the social sciences,” Elizabeth Stephenson and Patti Schifter Caravello explain the importance of data literacy in the social sciences and the appropriate role for librarians:

> In many of the social sciences, and particularly in sociology, comprehension of the scholarly record and development of critical thinking rely in part on an understanding of numerical information and its representations. Data literacy is therefore an integral component of information literacy for these disciplines. While it is not up to the librarian or data archivist to teach statistics and quantitative methodologies to undergraduates – faculty teach this in methodology and statistics courses – information professionals can have a role in helping students develop and build data literacy. (Stephenson)

Over the past two years, in working with Economic Demography, a large enrollment (300+ students) course taught at University of California, Berkeley, I've had the opportunity to apply practices similar to those described by Stephenson and Caravillo. The course provides students with a broad theoretical grounding in a subject area that lends itself to quantitative study, even for students with limited statistical knowledge. Here is how it's described in the Spring 2010 syllabus:

> This course will examine various economic and social causes and consequences of population change in an international context. The consequences studied will include the economic impact of immigrants on US workers and taxpayers, the growing pension burden as populations age, the effect of population growth on economic growth, and environmental consequences of population growth. The course will also examine the economic causes of demographic behavior including fertility, marriage, and labor supply. How have the functions of the family changed during the course of economic development, and how do they continue to change today? Why have divorce and extramarital fertility risen so much, while fertility has fallen way below

> replacement in many countries, and marriages are postponed to later ages or foregone altogether? How are these profound changes in family life related to the changing economic roles of women, and to economic growth? Finally, the course will consider whether there is a gap between individual and societal net benefits to childbearing, which would provide grounds for government intervention to alter birth rates.(Lee)

These topics are, as most social science librarians will recognize, rich with relevant, and readily available data sources, such as the U.S. Census, Current Population Survey, the General Social Survey, and many others. Accordingly, the course place a great emphasis on the following quantitative term project:

> The term project gives you the opportunity to get hands-on experience doing research using demographic data to answer a question of your choosing. In the past, students have found it challenging, but very interesting and rewarding. Many find it their favorite part of the course. Those with statistical skills can use them in the project, but these skills are not necessary. While regression models are not necessary, cross-tabulations, charts, correlations, and other data analysis techniques are useful. The topic of the paper must involve demography, but it does not need to be closely related to any theory presented in class.
>
> Unlike most other papers you may have written, this paper is not a literature review or critique. This is to be a brief paper based on an original analysis of primary data. Do not take your data from a published paper or report which has already analyzed it. You should draw your own conclusions from your own analysis of the data. If you are not clear on what I mean by this, please ask me in class. (Lee)

Given the size of the class and the fact that there is no quantitative methods prerequisite, this assignment proves particularly challenging to both students and the graduate student instructors who teach the sections and provide individual consultation. This presents a perfect opportunity for the Library play an important support role that extend well beyond traditional bibliographic instruction. Initially, early in the semester, I make a 10-15 minute presentation to the entire class. This allows me to introduce myself and give an overview of the location, services, and resources available in the Library Data Lab. Students are encouraged to begin their research projects early and are informed that despite the relatively short length of the final paper, the project might well take more time than a conventional term paper, due to the inherent difficulty of working with data. Several weeks later, a series of workshops are scheduled during regular section periods. Attendance is not required but is encouraged, particularly for those students new to data analysis. These workshops are approximately 90 minutes in length and provide in depth instruction in the use of SDA, a web-based statistical application, as well as an overview of how to search for data on the web.

Students are encouraged during both presentations to make appointments to meet with me to discuss their research topic and the availability of relevant data; approximately a third of the students schedule or drop-in for appointments. During these appointments, assistance is provided with both the research portion of their project and analytical aspects. While the SDA software is fairly intuitive to use, the data analysis process requires that the student learn some basic skills fundamental to working with 'raw' data.

SDA is web-based data analysis software developed by the Computer-assisted Survey Methods (CSM) Program at the University of California, Berkeley. The project is ongoing and new features and enhancements are regularly introduced. CSM also maintains the SDA

Archive, which provides free access to such data as the National Opinion Research Center's General Social Survey, the American National Election Survey, and Public Use Microdata from the U.S. Census. In recent years, a number of other data archives have chosen to make SDA available on their websites, making hundreds of data sets available for online analysis. Most significant are the Inter-University Consortium for Political and Social Research (ICPSR), the Minnesota Population Center's Integrated Public Use Microdata Project (IPUMS), and University of Toronto's SDA@CHASS. Many other institutions have licensed SDA to make data sets available for teaching purposes on non-publicly accessible websites.

The appeal of SDA is that it allows instructors to integrate data analysis into their curriculum without the overhead of having to teach or require prior knowledge of a full-featured statistical application like SPSS, Stata, SAS, or R. Not to say that these applications don't have an appropriate place in the curriculum, but for an introductory data analysis course, SDA has some distinct advantages. First of all it's free and runs in a web browser, so students can easily access it from home, classroom, or computer lab. Second, the software, as compared with the full-featured desktop applications, offers a limited, but fast and powerful set of statistical functions, thus minimizing the choices a user must make and simplifying the user interface. Users needing more advanced statistical functions than SDA provides are accommodated by subset and extract capabilities which allow data to be exported from SDA for use in other applications. Third, because loading data into the system is not an end-user responsibility, the potentially technical and time consuming tasks of downloading and preparing data files is avoided. The fourth factor contributing to SDA's appeal is the volume of data and breadth of topical coverage available through free or consortial membership sites like IPUMS, the Berkeley SDA Archive, and ICPSR. This provides students considerable leeway in choosing a topic for their analysis.

An important aspect of the Economic Demography term project is that it does not require the use of inferential statistical methods like multiple regression or comparison of means. Many students limit their analysis to cross-tabulation, a descriptive statistical technique that allows an examination of the effect of one variable upon another. For example, a cross-tabulation could be used to display gender differences in occupational choices. While this is a fairly basic technique that doesn't allow actual hypothesis testing, it's application and mastery by students does address one particular concern of statistical literacy advocates. This concern centers on the ability of students, particularly those in non-quantitative majors, to comprehend basic statistical assertions like the following:

> In terms of people, the number of unemployed workers is much higher in the US than in Canada. But does this mean unemployment is more prevalent in the US than in Canada? No. The number who are unemployed is strongly influenced by the size of the population. To untangle the influence of population on the number who are unemployed we need to look at the unemployment rates: the percentage of those in the civilian labor force that are unemployed. (Schield)

The Economic Demography term project does a good job of building these sorts of competencies. The ability to do an effective job on the assignment boils down to how well the student comprehends the underlying data set, and how well they relate they outputs of their analysis (tables, graphs, etc.) to their topic. This requires the student to pay a great deal of attention to what constitutes a good *statistical* argument, much more so than in a traditional research paper where more weight is placed on rhetorical skill and the use of good sources.

The use of SDA in courses like Economic Demography, promotes data literacy in other ways as well. It teaches a student to answer a research question based on their own analysis of a data set rather than making reference to someone else's research. Often this involves having to make adjustments to a topic based on the availability of data. Other times it requires the student make creative use of an available variable as a proxy for some characteristic that's not exactly represented in the data; a classic example of this is using a certain level of educational attainment as a proxy for literacy. By having to think about such factors as how the data was collected and coded, the student is required to reflect upon the underlying research methods in a way that would otherwise be difficult to teach.

For those considering a similar instructional program at their own institutions, I can offer the following advice. First of all, I feel it's essential to collaborate with faculty to ensure that the statistical work is integrated into the curriculum. The students need to be receive a conceptual framework upon which to build, rather than just crunching numbers for the sake of crunching numbers. For this reason, a traditional statistical methods class, may not be the best type of class to partner with. A methods class focusses upon statistical inference, or the use of statistics to build models and test hypotheses. While SDA can be used for this purpose, it does demand a higher level of expertise on the part of the librarian. In my experience, a particular strength of the SDA package is that it allows a novice user to conduct simple *descriptive* statistical analyses like cross-tabulations and frequency distributions. The ideal types of classes to work with, in my experience, have been those focussed on a topic or theme for which ample data exist for student projects. Some examples are race, public health, employment, voting behavior, and public opinion.

It is also worth mentioning that the librarian's role in promoting data literacy doesn't have to be as involved as mine is with Economic Demography. A potential role could be to raise awareness to social science faculty of the availability of SDA and the scope of available data sets. The software itself has been around for a number of years so many faculty may have heard of it, but the incorporation of SDA into IPUMS, ICPSR, etc. is a relatively recent phenomenon that they most likely aren't aware of. For smaller classes, faculty might not require assistance working with students on the analytical aspects of a data analysis project. There's a natural fit, however, for librarians to be involved in research aspects of such a project, such as helping students find data, or finding publications that cite a particular data set.

For those who feel discouraged from pursuing this type of instruction based on bad experiences with, or lack of training in statistics, I can offer this advice. In my experience, most people's negative associations with data analysis stem from bad experiences in a statistics class or from difficulties they've encountered in trying to work with statistical software, or with data sets in unfamiliar file formats. SDA negates most of these issues. Speaking from experience, I've successfully trained colleagues with very little experience in data analysis to provide effective assistance to students.

References

Lee, R., “Economics 175/Demography 175—Economic Demography”. Course Syllabus, January 2010

“ICPSR Web Site”. Inter-University Consortium for Political and Social Research. June 9, 2010 <http://www.icpsr.umich.edu>.

"IPUMS USA". Minnesota Population Center. June 9, 2010 <http://usa.ipums.org/usa/>.

“Microdata Analysis and Subsetting with SDA”. Computing in the Humanities and Social Sciences, University of Toronto. June 9, 2010 <http://sda.chass.utoronto.ca/sdaweb/index.html>

Schield, M., “Information literacy, statistical literacy and data literacy”, *IASSIST Quarterly,* Vol. 28 No. 2/3, 6-11.

“SDA: Archive”. Computer-assisted Survey Methods Program, University of California, Berkeley. June 9, 2010 <http://sda.berkeley.edu/archive.htm>

Stephenson, E., Caravello, P., “Incorporating data literacy into undergraduate information literacy programs in the social sciences”, *Reference Services Review,* Vol. 35, No. 4, 525-540.