

Using the World Wide Web: Suggested Applications and Precautions for Teaching Multiple Linear Regression

Lynne M. Pachnowski and Isadore Newman
University of Akron

The World Wide Web provides an excellent resource of real data that can be used by research methods instructors. Using this data in class allows more time for students to spend learning how to write appropriate statistical equations. This paper provides two examples of multiple linear regression equations that were written based on data available on the World Wide Web. It also provides a number of Internet references for instructors that would be helpful for obtaining real data that can be applied in a statistics course.

One of the most important processes a research methods instructor can help his/her students acquire is the ability to write a good research question. It is always beneficial for the instructor to have real data to use in formulating a question, since these questions are more likely to be perceived by the students as practical and applicable. The World Wide Web acts as a wonderful source of real data that instructors and students may access in a quick and convenient manner.

Once an instructor and the students have identified research questions using this data and the instructor feels that the students are competent in formulating questions, then the next step is to apply appropriate statistical models to test these questions. Implementing this instructional technique reduces the chance of committing a Type VI error which occurs when there is inconsistency between the research question being asked and the research model used to test it. (Newman & Newman, 1994). When the World Wide Web data is so readily available for demonstration, the instructor and students are able to spend far more class time discussing the analytical questions of which statistical models are appropriate to apply to the research questions developed. Therefore, the students benefit since they are more likely to be engaged in higher-order problem-solving activities.

The following are examples of research questions derived from World Wide Web data and multiple linear regression equations that can be applied to them.

Examples of Sites and Their Multiple Regression Applications

- The Wilmington Institute: Trial and Settlement Sciences

<http://www.wilmington-institute.com/>

The home page of this web site states that the institute was established in order to "help trial lawyers, corporate counsels and governmental agencies forecast the probable outcome of their litigation and trials." By choosing "Jury Talk Survey", the user is able to select a number of on-line survey of recent, well-publicized court cases, such as the O.J. Simpson case or the Timothy McVeigh case. The survey inquires what the visitor's opinion is regarding the nature of the case and requests some demographic information about the visitor. For instance the McVeigh questions are: "Do you believe Timothy McVeigh was involved in the planning and/or execution of the Oklahoma City federal building bombing? and If you answered yes to (1), do you believe Timothy McVeigh was part of a well organized and financed, geographically dispersed anti-government conspiracy?" The participant is then asked to identify his/her age group, gender, ethnicity, and area of residence from a list of possibilities. Once the results are submitted, updated overall results in terms of percentages appear on the screen.

With a data source such as this site, students could look at the results of a survey (guilty, not guilty, etc.) and the demographic information of the respondents to the survey and be encouraged to write sample research questions. For example, are there differences in age, ethnicity, gender, and geographic region and the verdict given? Such a question and the model for it might look like the following:

Is there a profile using age, ethnicity (Black, Hispanic, Asian, and other), gender and geographic region (Georeg1, ... , Georeg5) that differentiates those respondents who respond "guilty" and those who respond "not guilty"?

Model 1: $y = a_0u + a_1\text{Age} + a_2\text{Sex} + a_3\text{Black} + a_4\text{Hispanic} + a_5\text{Asian} + \text{other} + a_7\text{Georeg1} + a_8\text{Georeg2} + a_9\text{Georeg3} + a_{10}\text{Georeg4} + a_{11}\text{Georeg5} + a_{12}E1$
 Model 99: $y = a_0u + E2$

All the independent variables in Model 1 are binary-coded (1, 0).

Looking at the same data, another question that could be raised could be whether ethnicity accounts for a significant amount of variance in their perception of guilt/no guilt over and above age, gender, and geographic region. This would be done by testing Model 1 against Model 2 given below:

Model 2: $y = a_0u + a_1\text{Age} + a_2\text{Sex} + a_3\text{Georeg1} + a_4\text{Georeg2} + a_5\text{Georeg3} + a_6\text{Georeg4} + a_7\text{Georeg5} + aE3$

One can also test for interaction between pairs of variables such as age and ethnicity, age and gender, or gender and ethnicity since examples of addition models could not be tested.

- U.S. Census Bureau
<http://www.census.gov>

This site is provided by the U.S. Department of Commerce, Bureau of the Census. It contains a wealth of U.S. statistics, state and county statistics, and links to international census-related databases. The home page contains, among other items, links to the current U.S. and world approximated population counts, current economic indicators, and census documents.

A powerful link for researchers is the "International Data Base", which can be obtained by clicking through the path: "Current U.S. Population Count", "World", and finally "International Data Base". The visitor is then offered three links which offer three different ways of accessing the data provided. The visitor may either look at the data on the screen, load the data on a spreadsheet, or choose to configure appearance of the output. Once the output manner is selected, the user is then asked to select a statistically-related table (for instance, "life table values, by sex"), to select one or more countries from a table, and to select one or more years or to accept the latest available year as a default. The database will return the requested data or a message stating that the data was not available if it applies.

In one instance, we obtained a table containing the population of Canada by ethnic group and sex and also the U.S. population by ethnic group and sex. (After requesting the latest available year for each, the Canadian data provided was from 1991 and the U.S. data was from 1980. Since significant population

changes probably occurred during those years, another search may want to be done to obtain 1980 data from Canada or similar data from another country from a year closer to 1991.) Data such as this would be helpful in teaching students how to write regression equations to test Chi Squares. A question derived from the data may be:

Proportionally, are there more men than women in the U.S. or Canada?

Model 3 $y = a_0u + a_1\text{Males} + E4$

Model 99a $y = a_0u + E5$

In the models above, $y = 1$ if male, and 0 otherwise. Also 1 = "from Canada" and 0 otherwise. By testing Model 3 against Model 99a, we can see if there is a significant difference in the proportion of males and females in the U.S. and Canada. The student would have to take the data given from the screen or spreadsheet and recode it in a manner ("zeroes" and "ones") that would be most effective for the statistical analysis. Since the data are presented in aggregated form, the student will have to learn how to put in the data in individual form from the aggregate. Also, the student will consequently learn how handle data that is provided in different formats. This could be part of the instructional process.

The above examples are only two of the many World Wide Web sites that can be used to obtain data that could be used to make examples more realistic. Many other sites are also available with similar or even more extensive data than the sites mentioned. One excellent archive of links to "data and depiction's of data from throughout the world" is "Dr. B's Wide World of Web Data" found at <http://seamonkey.ed.asu.edu/~behrens/siip/webdata> (then choose "Wide World of Web Data") and created by Dr. John Behrens of Arizona State University. This site contains thirty-one links to sites that contain data sets or data-related information. The data sets are separated into sixteen categories, including "Children and Youth", "Demographics", "Education", and "Social Science -- General". The page encourages instructors to use the data for examples in class and to encourage students to find data that they find interesting.

Instructional Precautions

An instructor that chooses to integrate information of the World Wide Web into a course needs to be aware of the advantages and disadvantages of using such a medium. Because of all the attention the Internet has received in the media, the advantages -- accessibility of data, student convenience -- may seem more evident than the disadvantages. Instructors and students both must be aware of some of the following cautions:

- Each server hosting a web site only has the capability of hosting a finite number of users. Therefore, some sites may not have the capability of hosting a class of twenty students each attempting to visit the site at one time. Therefore, in-class lab time experiences should be planned so that students have a variety of sites from which to choose. If a visit to a particular site is required, instructors should assume that the average student may need to make two to three attempts on different occasions in order to make a connection.

- While new sites are appearing on the Web every day, old ones are often neglected. Sometimes, a promising link title or URL address may, in fact, have no file at the end of it. Furthermore, many sites containing data may left to become obsolete. A user should look for a notation on the page as to when the page was last updated.

- One of the most important cautions to students and instructors of research is the format that "data" can take on the Web. Although the census site has data sets that may be downloaded and the Wilmington site is both interactive and provides overall results, some sites have a much more limited interface. A site that is promoted as a "database" may only have a search engine interface which keeps the entire database hidden from the user. A user may only view pieces of the database based on the parameters of the search he/she submits.

Although these precautions may seem daunting to some new users, a reasonably proficient Web user can address these precautions by simply testing each Web source before a classroom application and creatively designing classroom assignments involving the Web.

Despite the precautions of using the Web within a statistics course, it is still difficult to deny the long-term advantages of using the Internet data as both a teaching and research tool. Among these advantages is that students are able to gain experiences in working with a medium that is increasingly more likely to be a primary source of data in the student's home and workplace.

A paper presented at the 1997 Eastern Educational Research Association's National Meeting related to this topic can be found at:

<http://junior.apk.net/~jurczyk/eera.html>.
The paper contains links to the sites mentioned above as well as the following other related sites:

Government Resources:

U.S. Census Bureau

<http://www.census.gov>

Census reports and links to other federal government and international agencies offering statistical reports.

Fedworld

<http://www.fedworld.gov>

Central location and starting point for finding U.S. government information.

Government Statistics on the Internet (paper)

<http://www.stats.gov.nt.ca/Bureau/General/WWWPaper.html>

Survey of government statistics (Canada, U.S., U.K.) available on the Internet.

SEC (Securities and Exchange Commission)

<http://www.sec.gov>

U.S. government site includes filings by public companies.

Stat-USA

<http://www.stats-usa.gov>

Department of Commerce service offering detailed government statistics-based reports.

Other Resources:

Facts on File

<http://www.facts.com>

Producer of comprehensive studies of modern issues. Reports include some survey results with statistics.

The Gallup Organization

<http://www.gallup.com>

Provider of public opinion poll data.

The Harris Poll

<http://techsetter.com/harris/html/home.html>

Contains the latest Harris poll and comparisons of the previous poll's telephone responses with Internet responses.

CollegeNet

<http://www.collegenet.com/>

A directory of colleges and universities divided into various categories and search parameters.

Texas Lotto

<http://crashdummy.iglobal.net/lotto>

The results of the latest Texas Lotto drawing and the results of the drawing over several years.

References

Braun, E. (1994). *The Internet Directory*. New York: Ballantine Books.

Ellsworth, Jill H. (1994). *Education on the Internet: A Hands-On Book of Ideas, Resources, Projects, and Advice*. Indianapolis, IN: Sams Publishing.

Hahn, H. and Stout, R. (1994). *The Internet Yellow Pages*. New York: Osborne McGraw-Hill.

Newman, I. and Newman, C. (1994). *Conceptual Statistics for Beginners*. Lanham, Maryland: University Press of America.

Place, R. Dimmler, K. Powell, T. (1996).

Educator's Internet Yellow Pages. Englewood Cliffs, N.J.: Prentice-Hall.