Data Entry: Going Pro

Seth Masket
University of Denver
smasket@du.edu

Suppose you have a large quantity of data in an unusable format. It could be hard copies of campaign finance reports, printed records of roll call votes, or eighteenth century election returns. Whatever it is, it's of no real use to you unless it's inputted into a computer database, but that will require literally hundreds, if not thousands, of work hours.

You could approach this problem by hiring a team of student researchers. Or you might just sit down in front of your computer with a thermos full of coffee and prepare for a long, hard slog. Before going these routes, though, you might consider hiring a professional data entry firm.

Professional firms offer many advantages over other data entry methods. For one, they are highly accurate in their work, which is particularly helpful with large datasets, for which it is nearly impossible to check every inputted item. Professional firms are also considerably faster than most students or researchers. After all, this is what they get paid to do, while students and researchers do such data entry in their spare time. Finally, professional firms, particularly ones based outside the United States, usually offer quite reasonable rates.

A co-author and I recently hired a data entry firm in India to input more than a century of roll call votes from a state legislature's journals into a database. Through a brief e-mail chat, we were able to establish the parameters of the job and our expectations, and we easily negotiated a reasonable price that was far below those of comparable domestic firms. Despite the size of the job (more than 70,000 votes cast by over 4,000 legislators), the firm was able to meet our expectations quickly and with a very low error rate. Indeed, I am now working on a second state legislature and have hired the same firm to help me.

The fact that the vendor was based outside the U.S. proved not to be a problem. The vendor's employees were proficient in English and highly technically skilled. Shipping can be an issue when working with foreign vendors - it is not cheap to mail hundreds or thousands of photocopied pages overseas. In this case, it proved not to be an issue since the state legislature we were studying makes their journals available on the web. However, when it is an issue, there are ways to mitigate such costs. For example, many universities now have digital senders - devices that rapidly scan pages into PDF format and e-mail them as file attachments. This way, raw pages can be burned onto a CD-ROM, which is much cheaper to mail than reams of paper. The pages can also be uploaded to a server so they can be accessed by anyone in the world with a computer and an internet connection.

One way to improve the accuracy and efficiency of any data entry project is to develop a web-based interface. That is, instead of having a firm directly type your data into a spreadsheet, you can design an web-based program to simplify the data entry. The user's input can be limited to pull-down menus and radio buttons, substantially reducing the potential for error. A web-based interface has numerous other virtues, as well, in that it can be programmed to produce a dataset precisely as you want it and you can keep track of the project as it is being completed.

A web-based interface can be designed by anyone with a reasonable degree of skill with web-based programming. If you don't feel up to this task, it can be quite inexpensive to hire someone to do it for you. Students in computer science departments may be willing to take on the task for a reasonable fee or even course credit. Many professional data entry firms actually have software developers on staff who can help you with such a project. For a recent data entry project, I found a web programmer through the freelance programmer clearinghouse (for example, `http://www.rentacoder.com`). Most such project shouldn't cost much more than $100 to $200.

If you decide to hire a professional data entry firm, here are some tips before you close the deal:

Check references. Many firms have plenty of experience with business clients, but not as much with academic clients. It's helpful if they're familiar with academic researchers' needs and expectations.

Be specific about the output you're expecting. Send some sample material and even mock up some results, if necessary, to show the vendor what you need.

Check the output before it's completed. This is very easy if you have a web-based interface. If not, ask the vendor to send you some results after a dozen or so cases have been entered. Make sure they're doing it correctly.

Figure out your costs ahead of time. You'll need to pay for the data entry. You may need to pay to ship the raw data. You may need to pay for web programming and scanning. Estimate all of this ahead of time to determine if hiring a professional firm is the best use of your research money.

The last point is not a trivial one. Costs multiply as the project grows, so it's good to know what you're getting into. If shipping costs are going to be high, for example, and if the data entry method itself is at least somewhat resistant to errors, hiring graduate or undergraduate research assistants may be the way to go.

However, the adage that you get what you pay for holds true even in academic research. Even if the costs of a professional firm seem high relative to those of student research assistants, the speed and accuracy of the work may make it worthwhile. At the very least, I'd recommend getting a quote or two from professional firms and looking at some of their previous work. Getting more research money usually isn't impossible; fixing improperly-entered data usually is.

Book Review

Review of Janet M. Box-Steffensmeier and Bradford S. Jones' Event History Modeling: A Guide for Social Scientists

Tze Kwang Teo
University of Illinois at Urbana-Champaign
tzeteo@uiuc.edu

Event History Modeling: A Guide for Social Scientists. Janet M. Box-Steffensmeier & Bradford S. Jones. Cambridge University Press, New York, 2004, 232 pages. $24.00, ISBN 0-521-54673-7 (paperback); $65.00, ISBN 0-521-83767-7 (hardcover).

Quantitative political scientists have come to embrace the advantages of analyzing longitudinal over cross-sectional data. There are myriad social and political phenomena where the time-to-event occurrence is of substantive interest, but may be censored due to a variety of (quasi-)experimental conditions. Event history/survival analysis is well-suited for modeling such phenomena, but until last year books on this technique—be they introductory, intermediate, or advanced level—were written by scholars in other fields such as biostatistics (e.g., Collett 2003; Hosmer & Lemeshow 1999) and sociology (e.g., Yamaguchi 1991). Box-Steffensmeier and Jones (B-SJ) can thus lay claim to two "firsts". *Event History Modeling* (*EHM*) is the first in Cambridge University Press' *Analytical Methods for Social Research* series, and more importantly, the first book-length introduction to survival analysis written by political scientists.

B-SJ have written *EHM* with the applied researcher in mind, thus the book's level of difficulty is similar to the above-cited introductory texts. Prior knowledge of statistical distributions and maximum likelihood estimation will be helpful, but is not required. Examples of applications in American, comparative and world politics are favored over extensive mathematical derivations and proofs, and discussed throughout the text, in order to facilitate conceptual understanding of, and familiarity with event history models and data structures. The book is less suitable for those seeking a text that teaches data analysis using commands for a particular statistical package (e.g., Blossfeld & Rohwer 2002; Cleves, Gould & Gutierrez 2004; Tableman & Kim 2004), but note that S-PLUS/Stata code and data for the key examples covered are available online.[1]

The book can be divided more or less into four parts: conceptual underpinnings and basic mathematical functions (chapters 1–2); regression models (chapters 3–5); model selection, diagnostics, and extensions for inclusion of time-varying covariates (TVCs) (chapters 6–8); and models for "complications" of unit heterogeneity or multiple events (chapters 9–10). The concluding chapter (chapter 11) covers a broad range of issues that can crop up in social science settings. I also note that the unfortunate problem of space constraints has led to the omission of nonparametric descriptive methods,[2] and a shorter-than-desired discussion of how event history analysis relates to causal modeling.[3]

I have two minor issues with the book. First, I would have liked, particularly in the parametric models chapter, more elaboration on the differences between (log-relative)

[1] `http://www.u.arizona.edu/ bsjones/eventhistory.html`

[2] The webpage does host a handout on the Kaplan-Meier survivor function. Readers who want to learn more about these methods may consult texts like Collett (ibid.) and Hosmer & Lemeshow (ibid.).

[3] Blossfeld & Rohwer (ibid., ch. 1) discuss this with respect to cross-sectional, panel, and event history data structures.