

A short history of Statistics with Stata

Lawrence Hamilton

The micro revolution was washing across our desktops in 1985, pounding in like surf. Among computer nerds, those were times of high excitement. The previous year, I had set aside my Atari 800 (which still rests in the attic) for a more powerful PC “clone” that boasted two 256-kilobyte floppy disk drives and another 320K of RAM. All that power soon proved not enough. My old notes are full of scheming, as I wheedled small grant after grant from university administrators for peripherals and hardware upgrades. I ordered software, too, almost anything that got good reviews in *Byte*, with only trial and error to teach me what would actually be of use.

This applied to statistics as well. Already a pilgrim, over the previous decade I had drifted from card-sorters and homebrew FORTRAN programs through SPSS, BMDP, and IDA on mainframes, looking for something that worked the way I thought: more visually and more quick to jump sideways, than the number-crunching packages permitted. My pilgrimage continued on micros, where I checked out SYSTAT and Conversational SPSS. When Minitab added crude ASCII box plots, stem-and-leaf displays and other tools from Tukey’s seminal *Exploratory Data Analysis* (1977), it became the best thing I’d seen. Even so, I kept looking.

In August 1985, walking past a booth at the *American Sociological Association*, I saw something new called STATA/Graphics. The graphics made me stop and look closer. The ease with which you could shift back and forth between simple statistics and regression, generate new variables, and visualize everything with clean, publishable graphs all struck me immediately. And it was programmable! STATA/Graphics, designed for the new desktop world, was exactly what I’d been looking for. Within a week I had ordered my own copy, paying for it out of a grant for research on water crises.

That fall, on sabbatical, I studied water crises but also wrote the prospectus for an 8-chapter textbook to be called *An Introduction to Modern Data Analysis*. The book would combine statistical and software instruction, the latter meaning Stata. This prospectus was mailed to half a dozen publishers. In January 1986, I phoned Bill Gould to describe my ideas. He didn’t know me from Adam and seemed understandably cautious at first but soon turned supportive. I signed a book contract with Brooks/Cole in April.

Over the next six years, encouraged by John Kimmel at Brooks/Cole, my 8-chapter book grew alarmingly. It doubled to 16 chapters, then tripled into a trilogy that ate my life. The trilogy’s first two volumes, *Modern Data Analysis* and *Statistics with Stata* (*SwS*), were both published in 1990. *Statistics with Stata* came with a 5.25” floppy disk containing datasets and a student version of Stata 2. Alongside examples from basic statistics through DFBETAS and some Tukey-inspired notes on smoothing and robust regression, *SwS* contained a small piece of original research: a logit analysis of the Challenger disaster, a few years ahead of its time.

The trilogy's third volume, *Regression with Graphics*, appeared in 1992. One reviewer later termed it a "cult classic", which cheered me even though this meant the opposite of "best seller". *Regression with Graphics* was built around examples from my water-crisis research. Its title reflects the same interests that first attracted me to Stata. Stata drew all of its graphs and can be seen on the cover as well: a galactic design featuring a twoway scatterplot with marginal box plots and oneway plots, tricks that Stata no longer recalls. For chapters on robust regression and computer-intensive methods, I rewrote the old `rreg` do-file and then tested it through Monte Carlo experiments. The 4.77-megahertz Intel 8088 processors were so slow that I divided the calculations for one experiment, which required 150,000 regressions, among five different machines over the weekend. Some of the background work for *Regression with Graphics* turned into articles for the *Stata Technical Bulletin*, as well, not only on robust regression and Monte Carlo programming, but also more tangential topics such as definitions of quartiles.

Meanwhile, *Modern Data Analysis* sold poorly, but *SwS* sold well. By 1991, as I mailed off the manuscript of *Regression with Graphics*, my trilogy finally complete, I was already under pressure to write the next *SwS*. Naively imagining that this would be a simple revision, I soon found that a complete reorganization was needed and began to recognize a core problem: how could *Statistics with Stata* keep up with Stata itself? Told to stay below 200 pages, I gave *Statistics with Stata 3* (1993) a denser layout to pack in new material, and ended each chapter with an "Also Type help" section pointing readers towards Stata's on-disk documentation. The new material included full chapters about robust regression and factor analysis, as well as new sections on simple tests, nonlinear regression, multinomial logit and other topics. A few examples from the Arctic crept in, reflecting the new direction of my research.

Over 1994–95, between Arctic trips, I worked on an undergraduate textbook. *Data Analysis for Social Scientists* (1996) was keyed to a menu-equipped student version of Stata 4 called StataQuest. Although StataQuest showed promise, and its menus foreshadowed Stata 8, this unloved child was never updated. *DASS* was not popular, either, leading me to conclude that my writing talents did not encompass undergraduate statistics. I turned back to what I did better, which seemed to be *Statistics with Stata*. *SwS-5* came out in 1998, with over 300 pages, including an expanded chapter on data management, and new chapters on survival analysis and programming. I found ways to fit in additional Stata features, which were multiplying like tribbles, by adding sections of "Example Commands" with short explanations at the start of each chapter and placing long lists of options within. Both approaches were carried farther in *SwS-7* (2003), along with a new chapter on time series, illustrated mostly using Arctic examples.

Stata 8, the most radical upgrade in Stata's history, confronted me with three problems. The foremost was that it made my just-published *SwS-7* obsolete, so no one would buy it now. The others were the new menu interface and redesigned graphics. After some thought, I largely ignored menus for *SwS-8*. It seemed far easier to write out and explain a command, even a long one, than to illustrate with the equivalent sequence of menu selections. The book would become unwieldy if I tried to do the latter. I noticed that Stata's manuals had made the same choice.

The version 8 graphics were another matter. At first I was resistant (Damn! Nothing I know works! Good thing there's `graph7`.) but later came to embrace them, turning graphics into the longest chapter in *SwS-8*. I took an example-based approach, unlike the command-based reference manuals. Writing became a process of self-education that grew more interesting as I gradually caught on. When I had first viewed Stata in 1985, I was immediately drawn by the way Bill Gould had brought Tukey-flavored graphical analysis into the microcomputer age. Nineteen years later, in Stata 8, I saw the spirit of a new graphical guru. Edward Tufte, writing in *The Visual Display of Quantitative Information* (2001) and elsewhere, celebrates the design of clear, creative, and information-rich graphical displays. Stata 8's detail control and overlays opened new doors for such designing, which *SwS-8* could only start to explore.

About the Author

Lawrence Hamilton is Professor of Sociology at the University of New Hampshire. His specialties are the Arctic, human–environment interactions, and statistics and data analysis.