Review of Stata 7

Stanislav Kolenikov
skolenik@unc.edu
117 New West, Cameron Ave, University of North Carolina, Statistics Dept.
Chapel Hill, NC 27599-3260, US

May 10, 2001

1 General introduction

Stata 7 is a general-purpose statistical package that does all of the textbook statistical analyses, and has a number of procedures found only in highly specialized software[1]. Unlike most commercial packages aimed at making it possible for any Windows user to produce smart-looking graphs and tables, Stata is aimed primarily at researchers that possess the knowledge of the statistical tools they are using in their subject fields. The applications of statistics that are covered best by Stata are econometrics, social sciences and biostatistics. The Stata tools for the latter two categories are contingency tables, stratified and clustered survey data analysis (useful also for health and labor economists), and survival data analysis (useful also in the duration studies, such as the studies of the lengths of unemployment or poverty spells).

There are several things that I like most about Stata. First of all, it is just a very good package for doing applied research, with lots of everyday estimation and testing techniques, as well as convenient data handling tools. The unified syntax makes all these things easy to use. It is also a rapidly developing package, with excellent both module extension and file sharing capabilities built into it. Stata features a great support environment, both from the developers, and from the advanced users of the package. The best example of the informal support network formed around Stata is the Statistical Software Components archive (SSC-IDEAS, part of RePEc) that contains several hundred user-written programs. For many users, it is also important that the academic and student prices can be quite low.

[1] Stata software is developed and distributed by Stata Corporation, 4905 Lakeway Drive, College Station, Texas 77845, United States. Tel. 1-979-696-4600. Fax 1-979-696-4601. E-mail: stata@stata.com Corporate website: http://www.stata.com. The worldwide distributors list: http://www.stata.com/info/worldwide/.

2 Specifications and installation

Stata has long been available in two modifications (Intercooled and Small) for all popular platforms: Windows 9x/NT/2000, MacOS and many Unixes (Digital Unix, HP-UX, AIX, Linux for Intel and Power Mac platforms, Solaris; only Intercooled version is available for Unix). Intercooled and Small Stata differ in their capabilities, the Intercooled being the full version, and Small, the restricted version (smaller number of observations/variables; smaller memory addressed, and some other technical limitations) applicable probably only to the undergraduate teaching. The prices of Intercooled Stata for different categories of users may differ quite substantially, with the lowest student license price of $55, academic offers in the range between $100 and $500, depending on the options and eligibility, and the commercial price approaching $1000 (for the first copy of Stata)[2].

Stata for Windows, Unix, and Macintosh are identical, and cross-platform compatibility is one of the long-standing committments of Stata Corp. It is really nice to be able to design your data analysis at an outdated laptop with a 1% sample of the original data and then upload your data file and the program to a university multiprocessor server to run it with the full data set.

The reverse side of the coin is that not too much attention is given to the drug-and-drop operations one might got used to under Windows. This especially concerns the graphics. User can copy-paste Stata output, including graphics, to her favorite word processor, although it only uses Ctrl+C shortcut rather than Ctrl+Ins that I am personally used to.

Stata is a big system in terms of capabilities, but unlike other big systems that are distributed on several CD-ROMs and require all recent Microsoft libraries, Stata is quite tolerable to 486DX/16M system under Linux or Windows 95. This makes Stata great for budget applications and institutions, as well as for the third world countries, where the research institutions might not have much funds for the newest hardware.

Stata is shipped on a CD-ROM where all installation files for all platforms are written into. Additionally, the licensing information is sent along with the documentation. When installed, Stata 7 occupies around 10Mb of the disk space. Installation of updates and additional components would take another 2–5 megabytes. Stata datasets are organized rather efficiently, as they are typically compressed by a factor of only about two by popular archivers.

Stata installs itself in a single directory (e.g. C:\STATA or /usr/bin/stata), so that Windows version does not write anything to Windows directory. This, however, makes Stata rather susceptible to unlincensed copying. When Stata is installed, it requests the licensing information at the first launch. The license information is platform specific, and it also distinguishes between single user and network Stata. As a promotional offer, there is an option for a free monthly trial license (see "Share Stata" on the website). The network version tracks

[2] Prices are subject to change; see http://www.stata.com/info/order/ for the up-to-date information. The prices for multiple user network licenses are also worth checking, as they follow decreasing marginal price pattern.

the number of users currently logged on to Stata, although each user may run several processes.

I found a couple of faults with the installation program. First, it hung when it ran out of free space on the drive where it tried to install Stata to. Second, it did not tell me how to run Stata when it was finally installed.

Stata is a very stable software. I do not remember Stata crashing on the machines I was using. If there is a run-time error, Stata displays an error message with a non-zero returned code, and goes back to its normal operation, i.e. command prompt. Users can stop calculation by pressing Ctrl-Break in Windows or Ctrl-C in Unix, although a one line program can be written that captures it to hang Stata. When Stata receives the Break sequence, it needs to complete the operation being performed, delete all unnecessary temporary objects, and restore the status prior to last command issued, so the response to the Break command can at times be quite slow, especially when Stata uses uses virtual memory under Windows.

3 Concept

There are three main conceptual features underlying the operation of Stata statistical software.

First, Stata is an observation oriented package, in the way Gauss is a matrix language, and S-Plus is an object environment. The code is optimized in such a way that all operations are implicitly performed on all observations (unless a subsample is explicitly chosen), and those operations are very fast. This is due to the fact that Stata stores the whole data set in the RAM. It can also use the virtual memory, so there is no formal upper limit to the number of observations in the data set. Virtual memory is OK under Unix, but the use of the virtual memory under Windows results in continual swapping that reduces the computation speed by approximately three orders of magnitude.

Second, Stata is command prompt driven, which means that all programming capabilities are just the same as the interactive ones, and the user does not have to think much converting interactive sessions into programs. Simple Stata programs are sequences of commands one would issue with the keyboard in an interactive regime. They usually have an extension .do (thus called do-files in Stata jargon). Stata can be run in a batch mode, i.e. ordered to execute the specified do-file, write the output into the log-file with the same name, and then exit. Another nice use of the do-files is the specification of the startup profile as a command line option where the user can specify keyboard shortcuts, add paths to her programs, add menus to Stata, or request the updates from Stata web site. Stata 7 also has handy ways of converting the log of the interactive work into a program.

The above does not mean, however, that Stata has no menus. It does have some basic things like file operations, preferences, window management, and help. Users can add their own menus, thus effectively converting Stata into a mouse driven package. In fact, Stata Corp. offers StataQuest, the addition that

adds a lot of menus to help navigating through numerous statistical and data management tools.

Third, Stata is modular by its nature, resembling say MatLAB in that much of Stata is written in Stata code. In fact, Stata has two types of commands. The basic data, graphics and matrix operations, elementary statistical analysis and the command syntax parser are implemented in the main executable file as built-in commands, and everything else, including all advanced statistical modules, are the (interpretable) ASCII text file programs. The latter are referred to as ado-files (Automatic DO-files) and are usually stored in several places under the stata/ado directory. The development of Stata is towards shifting more work to the ado files, and currently, of about 1100 commands in the base release (including undocumented internal ones), only about 200 are implemented as the kernel commands.

The "modularity" has several implications. Stata may be easily updated, and complicated Stata programming can be started by picking the code accumulated previously in the similar problems. Finally, modular character of the language makes it natural to write easily sharable programs providing an unvaluable externality to the research community. As a result, even though Stata is produced by a commercial firm, Stata *capabilities* are produced to a large degree by an 'Open Source'-like effort of researchers all over the world. Several user group meeting have been held around the world (7 in UK, 2 in Spain, 1 in Netherlands, 1 in US).

A lot of user written extensions to Stata come from Stata mailing list and SSC-IDEAS (Statistical Software Components) RePEc archive mantained by Christopher Baum from Boston College. The "officially recognized" user-written additions are published in Stata Technical Bulletin (STB), a bi-monthly publication with a format close to Stata Reference manual. It is also run by Stata users rather than Stata Corp. itself. The editor of STB is H. Joseph Newton (Texas A&M University). STB hard copy publication is accompanied by free Stata modules downloadable from Stata website and a couple of mirrors, at Harvard and in Portugal. Stata has a special file description format that makes it possible to search the Net (including the SSC-IDEAS archive) for Stata components, so that user can find and get additions by typing a few commands / clicking a couple of menu entries. Stata Corp. willingness to provide these facilities in their programming language must be applauded. See also Section 6 below on the online resources.

4 Syntax and usage

Most Stata commands have the following format:

`[by` *variable list*`:]` *command* `[`*variable list*`]` `[if` *condition*`]` `[in` *range*`]`
`[using` *file name*`]` `[[`*weights*`]], [` *options*`]`

where the square brackets denote that the argument of a command can be omitted (but the brackets around weights are required if the weights are specified),

the first *variable list* is the list of grouping variables (so that the following command is sequentially applied to the subsets identified by the unique values of those variables, e.g. panels), the second *variable list* is the list of variables to which the command is to be applied (e.g. the dependent variable and regressor list for a regression model), the *condition* and the *range* select the appropriate subsample (e.g. `if year<1990`), the *file name* is used for file operations (load into memory, write to disk, import from or export to a text file), the *weights* specify the sample weights of the observations used for the command, and *options* contain whatever else Stata would need to know to run the program: the choice of the estimation method (e.g., fixed or random effect for panel data, full ML or two-step for Heckman selection model), the values of numeric tuning parameters, request for a more detailed output or for an alternative coefficient estimates covariance matrix to be used, etc. Some additional effort may be required for estimation of systems of equations, especially for specifying the (cross-equation) constraints.

Once installed, Stata can easily be kept up to date. Updates are made available two or three times a month and each includes bug fixes and/or new options for about about a half-dozen commands. The updates are made available over the web and, even better, Stata itself can find them and install them. Installation times are usually minimal since the files transferred turn out to be small (being Stata ado-files, which are reasonably short text files). Once in about two months, the executable itself is also updated (which Stata also handles automatically) and the size of that file is just over 1.5Mbytes.

The issue of numerical accuracy is addressed in Stata website publication — see http://www.stata.com/support/cert/. Stata passed most of the NIST StRD tests, except one for regression with highly collinear regressors. It just dropped some of the regressors, which is recognized as a legitimate response for this type of problem. The built in (pseudo-)random number generator (32-bit KISS algorithm with the period of $\approx 2^{126} \approx 10^{38}$) was also checked with the DIEHARD test and considered adequate.

5 Data management, matrix operations, and graphics

Stata supports its own binary format of data, although it can be taught to read in a text file with an almost any weird formatting one can think of, like several lines per observation. Stata offers optionally (and with an addtional charge) the most recent version of Stat/Transfer along with Stata installation files, so users would have no difficulties converting their data between all popular formats supported by Stat/Transfer. (Another popular data format converter that can deal with Stata files is DBMS/COPY.)

Stata supports file operations over the Internet: the user can simply type the URL of a file just as she would type the local path to it (see example in the Appendix). Instructors may find this especially useful: datasets (or do-files, or

anything else) can be stored on their home pages and they can tell their students to access the materials inside Stata from their home page.

Stata has versatile commands for dealing with data. It has a comprehensive set of statistical, mathematical, string and other functions, although relatively more exotic things like Bessel functions are not implemented. The operations can be restricted to a certain subsample of the data, or to be performed on groups of observartions. It is also quite handy in combining different files, e.g. by some index variables, and doing some special transformations of the data set, like transposing, changing shape, converting the variables from income92, income93, ... × one observation per object format ("wide" format, in Stata jargon) into income + time variables × many observations per object format ("long" format).

Dealing with dates in Stata is not always trivial. The dates are stored as integers (with the value of one assigned to January 1, 1960, or January, 1960, or 1960, depending on the frequency of the data, as specified by the user), not as an additional special format, so the user would need to convert the string dates into those integers to take advantage of lags, leads, differences time series operators, and special time series commands. Stata provides formats to make the integer date recording scheme readable, but if the things are not done properly, the date might appear as say 14900 instead of 17oct2000, which would make most users feel they made an error.

While Stata's programming capabilities allow you to add new command to Stata, they do not allow you to add new functions (such as sqrt()) that can be used in any context. If one wrote a module to compute Bessel functions, one could use the module to assign the values for a new variable **bx** as `egen bx=bessel(n,x)` (where `egen` is the special Stata command[3] to deal with those user-written functions), but one could not use the modifier `if bessel(x,n)` on the end of any Stata command.

Although matrices are not the main objects of Stata, most matrix operations that make sense for computational statistics can be performed, starting from simple arithmetic operations and concatenation down to Kroenecker product down to sweeps and inverse matrices. Various versions of matrix products and accumulated matrices relavant to statistical analyses are also available. All the basic computational linear algebra stuff is covered like Cholesky decomposition, singular value decompositions, eigenproblems. At the time this review was written, one could not save matrices as files for further use, but the new utilities for this operation were being developed at Stata Corp. The maximum size of the matrix can be set by the user, but it cannot exceed 800.

Graphics is one of the most controversial issues in Stata. Given the Stata Corp. committment to provide cross-platform compatible products and formats, Stata does not provide graphs that can be treated as flexible objects in Windows environment, like MS Excel graphs, let alone Java-based S-Plus graphical applets. There are no built-in 3D graphs, either. Still, most 2D capabilities are

[3] Read: Extensions to GENerate, where `generate` is the way to create new variables in simple settings. The modules for `egen` should follow some special syntax.

sufficient to create graphs acceptable for scientific publications. Graphs can be easily converted into (Encapsulated) PostScript or Windows Metafiles formats for further use in TEX or Word documents. Stata Corp. was developing new graphics engine at the time this review was written.

6 Documentation and online resources

There are two sorts of documentation available for Stata: built-in (on-line) help, and manuals. On-line help includes all the commands in the base distribution of Stata and explanations of the basic concepts of Stata (like what you can do with the keyboard, and how you use search system, and how you update Stata over the Net, etc.). More detailed explanation is available with the manuals. Stata 7 comes with 386-page User's Guide, 225-page Graphics Manual, 368-page Programming Manual, 180-page Getting Started Manual, and 2239-page 4 volume Reference Manual set. The entries of the latter differ from the online help as they include more examples and the technical details like methods and formulas. In fact, they can serve as a good introduction to the statistical concepts and topics related to the commands described, and there are 5 to 10 references to the literature in the end of each entry to suggest further reading. A cheaper Stata Reference Manual Extract (620 pp.) is available for budget installations and applications complying with the 80/20 Pareto rule (here, 20% of the documentation entries are sufficient for 80% of problems). There were multiple user suggestions to distribute the manuals in the PDF format, and there are some signs of acknowledgement of this initiative by Stata Corp.

The web site of the company (http://www.stata.com) provides a lot of additional information. The Capabilities section fully duplicates the online help. There is a small on-line bookstore where users can order not only additional sets of Stata manuals, but also a number of useful books ranging from [Greene (2000)] and [Johnston and DiNardo (1997)] to [Mooney and Duval (1993)] and [McCullagh and Nelder (1989)]. A number of published recently books use Stata as the primary software, and those are available from Stata bookstore — [Deaton (1997)], [Hamilton (1997)], [Hardin and Hilbe (2001)], [Long (1997)], [Rabe-Hesketh and Everitt (2000)], to name a few. Users can also find errata list for the manuals, register for the online Stata courses, download the official updates and additions to Stata from STB, or subscribe to Stata mailing list.

Another great online Stata resource is the Statistical Software Components archive SSC-IDEAS (http://ideas.uqam.ca/ideas/data/bocbocode.html) mantained by Christopher Baum at Boston College. This archive probably contains more end-user Stata commands (available free of charge) than Stata distribution itself! Coupled with the search and indexing opportunities of the RePEc system, SSC-IDEAS makes search for the statistical routines easy and simple.

7 Support

I have never had any problems getting support for my problems related to Stata, and there are several ways to get it.

The "official" way to get the support is to send a message to tech-support@stata.com. The technicians from Stata Corp. respond pretty fast. I never had response for my troubles delayed for more than few hours.

Second, there is a great mailing list where both Stata developers and experienced Stata users can give an advice for most Stata problems (or at least refer to the relevant software if Stata does not do the required job)[4]. In fact, the founder and president of Stata participates on the list and it is not uncommon that he responds. [5] I personally benefited quite a lot from just reading the statistical discussions on the list. Quite a lot of the list participants have expressed the point of view that Stata is fast, inexpensive and handy, even though there are fields where it can be outperformed by specialty packages.

For users more used to the traditional support tools, a hot line (and a toll-free 1-800 number for users in US/Canada) is available, as well as the support by fax or by mail.

8 Main econometric tools

One of the commands you would probably issue most often is `regress`. It estimates the OLS regression and presents the regression table (see example in the appendix). There are lots of variants of regressions implemented in Stata, including regression with White and Newey-West covariance matrix estimators, robust regression with Huber and biweight loss functions, regression with instrumental variables (2SLS and 3SLS for simultaneous equations), and some other regression-based techniques. The implementation of the GLS is not always trivial, however. Stata has matrix commands that allow for sandwiches with the GLS matrix in the middle, but I am not aware of a command that directly refers to the GLS estimation. Also worth mentioning is the versatile stepwise regression tool that searches for the "best" set of regresors.

Stata has a rich set of tools for regression diagnostics, including the measures of influence (studentized residuals, DFBETAs, DFFITS, COVRATIO) and overall "regression quality" measures (VIFs, heteroskedasticity, non-linearity (RESET) and serial correlation (Durbin-Watson) tests), as well as a number of graphical displays that help visualizing the regression results (added variable, partial residual plots, and some others). Not all of them are routinely accomodated in econometric practice despite their high practical value.

[4] User opinions on Stata, as well as comparisons to other statistical packages and references to the statistical methodology discussions, are selected occasionally from the list archive and made available at:
http://www.komkon.org/~tacik/stata/opinions.html.

[5] An actual quote from one of his letters on the list to characterize the attitude to users: "If we have made the wrong decision in pushing this problem to the side, we an certainly reconsider. Let me know."

The discrete and limited dependent variable commands include most of the things one could imagine in this area, except probably *rara avis* like ordered logit and probit models for panel data. Readily available are standard logit and probit models, Heckman sample selection model (with ML and two stage options) and treatment effect model, tobit regression, bivariate probit model, ordered logit and probit models, McFadden conditional (fixed effect) logit model, complementary log-log (Gompertz) model, and generalized linear models.

The set of commands that deal with the panel data is also exhaustive: random effect, fixed effect (within and between) linear regression, including an option for autocorrelated errors; GLS models with various patterns of within and across panel correlations; Prais-Winsten regression; instrumental variables and two-stage least squares; Arellano-Bond linear dynamic panel data estimator; fixed and random effect logit and random effect probit and gompit models; panel Poisson, negative binomial, and some other generalized linear models. Specification tests (LM test for random effect, Hausman test for fixed effect) are available. The panel data can also be converted into a special form that allows the use of the aforementioned `regress` with all its capabilities of regression diagnostics.

Time series capabilities were seriously introduced to Stata with the release of version 6 in January 1999 and are sufficient for most intermediate level time serires analyses. The canned routines include correlograms, ACF and PACF; general ML-based ARIMA and conditional heteroskedasticity (ARCH, GARCH, asymmetric, threshold, power and nonlinear ARCH) models; white noise (Bartlett and portmanteau) tests and unit root (ADF, Phillips-Perron) tests. The ARCH-type and ARIMA models allow the list of regressors, so the interpretation of these commands by Stata is that the residual process should follow ARCH-type or ARIMA. If only one variable is specified then of course the model is applied to that variable. Programs performing VAR are not a part of the official Stata, but rather are implemented as user-written additions available from SSC-IDEAS. Converting the data set to the format understood as time series by Stata may be somewhat cumbersome, and getting used to the date formats and conversion options will take some time.

Some users might be interested in analysing duration data like unemployment or poverty spells. Stata's tools for this type of analysis are comprehensive (Cox propotional hazard model; parametric survival models with exponential, weibull, Gompertz, lognormal, loglogistic, gamma families; variuos data management and graphical routines), but converting the data with the observed times to the survival time data set would require even more effort than in the time series case.

One of the greatest advantages of Stata is its maximum likelihood estimator [Gould and Sribney (1999)]. Users can write their own likelihood functions and let Stata maximize them. There are several versions of the procedure. In the simplest i.i.d. case, the user just codes the "observation-wise" likelihood function; Stata then performs summation (excluding missing observations if need, or accounting for weights if they are specified) and maximizes the function numerically. More difficult cases can include computation of the full likelihood function

(e.g. in the panel data with intrinsic dependencies). Finally, to increase computational efficiency, the user can code analytical first and second derivatives. The maximization algorithm seems to be quite involved and includes random search at the early stages, then some unidirectional optimization, and then some combinations of Newton-Raphson and steepest ascent algorithms with diagnostics of concavity.

Stata has unified syntax for hypothesis testing after all estimation commands mentioned above. It can test simple linear restrictions, combine them into F- or χ^2-tests, or perform nonlinear hypothesis testing based on the delta method.

One of the main Stata's weaknesses as an econometric package is the lack of the readily available GMM estimator, even though Stata already has a very good maximizer routines implemented in the maximum likelihood estimator. Also, Stata has a number of estimators built on the GMM principle, like Arrelano-Bond estimator for dynamic panel data models.

9 Getting to know Stata

As all softwares, learning Stata requires some initial impetus. The very first steps in Stata can be observed with its built-in tutorials (just type `tutorial` when you launch Stata for the first time in your life). The set of tutorials is rather limited, however: there is an introduction to Stata, tutorials on tables and graphics, tutorials on regression, ANOVA, logit, survival and factor models, and data input. Several toy data sets are supplied with Stata to illustrate linear regression analysis and diagnostics, survival analysis and ANOVA, as well as tabulation and graphic capabilites.

A good way to learn Stata is through the NetCourses offered by Stata Corporation. The Introduction to Stata, Stata Programming and Advanced Programming courses are given twice a year. Other courses offered include Survival Analysis and ML Estimation.

Some references to Stata resources are collected at http://www.stata.com/links/resources1.html. One of the most extensive online descriptions of Stata outside Stata Corp. itself is probably http://www.ats.ucla.edu/stat/stata.

Author's own experience shows that it takes about a couple of weeks to get acquainted with the main commands and start writing your own do-files to automate the basic data analysis; then about several months plus a couple of NetCourses to get used to program sequences of commands for your data management or estimation purposes; and then the rest of your life to get to know all tricks and bells and whistles of Stata through your own programming and participation in the mailing list. This pace was mainly achieved through self education with Stata being launched at startup on my office PC every morning and being used at least a couple of hours a day. My goals in knowing Stata might have been too ambitious, though, so the learning curve for students getting a well designed introductory statistics course with Stata need not be the same.

10 Historical developments and previous versions

Stata is developing at a very steep rate, and several initial comments had to be removed from this review during the weeks it had been prepared, as the new capabilities were introduced. For instance, one of my initial intents was to mention the lack of VAR routines, and that niche was filled by the `vecar` routine by Christopher Baum, Boston College. Still, in Stata history, the main landmarks are related to releases of the new versions of the package.

The main new features of Stata 6 (released in early 1999) as compared to the previous version 5 were introduction of a new set of time series concepts, operators, and routines; improvements in the ML optimizer; new programming syntax that allowed explicit parsing of the command line; Internet capabilities of Stata; and a number of new statistical routines. Still, graphics remained on the same level, and there were some incompatibilities between Windows and Unix versions of the software due to the way X-windows work with graphics.

In Stata 7 (released in December 2000), new GUI was developed for Unix version of the software, and these incompatibility bugs were fixed. Some users found it a nuisance that Stata did not calculate some quantiles and/or tail probabilities of statistical distributions with double precision. This was also fixed in Stata 7. Among other new features of Stata 7 were: a set of procedures for cluster analysis, long variable names, and a new output engine (SMCL: Stata Markup and Control Language). Also, Stata improves its speed from one release to another by some 5 to 10%.

Prospective developments that were outlined by William Gould, the president of Stata Corp., at the North America User Group Meeting [Gould (2001)], included: LM test for omitted variables as one of the statistical developments; HTML and TEX formatted log files; new graphics; Big Stata with many restrictions weakened (most importantly, the one on the matrix size); connection of Statas over the Internet.

11 Final remarks

I would boldly recommend Stata for research, especially in cross-sectional, panel and survey data. Teaching Stata might be difficult at the first stages, as you would probably have to explain all commands and some of their details to the students if they have no knowledge of Stata. Learning Stata is an investment, not necessarily with an immediate return, but as all investments into human capital, pays off really well, and in couple of months time you would probably do things faster and more efficiently than with a mouse driven package.

Stata is not the best package for:

- presentation graphics (although publications graphics is OK);
- intricate macro and financial time series analysis (goes only as far as unit root tests);

- GMM or such relatively exotic methods as neural nets, simulated anneal, CARTs, or genetic algorithms;

but it is certainly great at:

- pre-programmed long sequencies of simple commands (maybe tried previously in the interactive work);
- panel and survey data analysis;
- discrete and limited dependent variable analysis;
- survival and duration analysis;
- maximum likelihood estimation;
- regression analysis and regression diagnostics;
- data management and conversion;
- desktop scientific and statistical calculations.

And when you get Stata, do not forget to look up "endless loop" in the manual.

12 Acknowledgments

I am grateful to Stata Corp. for the version 7 release for review purposes; William Gould, Nick J. Cox, Christopher Baum and the Software Review Editor for their most helpful comments and suggestions for this review.

A Example of a short Stata session

This example checks for the updates at Stata website, loads the data over the Net, runs a regression with this data, finds the tool to format output regression (the standard errors in the parentheses, etc.), performs a number of diagnostic checks, and repeats estimation with White covariance matrix.

```
. log using example.log, replace
(note: file d:\stata\review\example.log not found)
-------------------------------------------------------------------------------
       log:  d:\stata\review\example.log
  log type:  text
 opened on:  06 Mar 2001, 09:49:02

. update query
(contacting http://www.stata.com)

Stata executable
    folder:               D:\STATA\
```

```
    name of file:         wstata.exe
    currently installed:  05 Feb 2001
    latest available:     05 Feb 2001

Ado-file updates
    folder:               D:\STATA\ado\updates\
    names of files:       (various)
    currently installed:  05 Feb 2001
    latest available:     01 Mar 2001

Recommendation
    Type -update ado-
------------------------------------------------------------------------------

. update ado
(contacting http://www.stata.com)

Ado-file update log
    1.  verifying D:\STATA\ado\updates\ is writeable
    2.  obtaining list of files to be updated
    3.  downloading relevant files to temporary area
        [... output omitted ...]
    4.  examining files
    5.  installing files
    6.  setting last date updated

Updates successfully installed.

Recommendation
    See help whatsnew to learn about the new features

. use http://www.stata.com/users/vwiggins/auto.dta
(1978 Automobile Data)

. desc

Contains data from http://www.stata.com/users/vwiggins/auto.dta
  obs:            74                          1978 Automobile Data
 vars:            12                          7 Jan 1999 17:49
 size:         3,478 (99.6% of memory free)
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
make            str18  %-18s                  Make and Model
price           int    %8.0gc                 Price
mpg             int    %8.0g                  Mileage (mpg)
rep78           int    %8.0g                  Repair Record 1978
hdroom          float  %6.1f                  Headroom (in.)
trunk           int    %8.0g                  Trunk space (cu. ft.)
weight          int    %8.0gc                 Weight (lbs.)
length          int    %8.0g                  Length (in.)
turn            int    %8.0g                  Turn Circle (ft.)
displ           int    %8.0g                  Displacement (cu. in.)
gratio          float  %6.2f                  Gear Ratio
foreign         byte   %8.0g       origin     Car type
------------------------------------------------------------------------------
```

```
Sorted by:  foreign

. regress price mpg wei for

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  3,    70) =   23.29
       Model |   317252881     3   105750960           Prob > F      =  0.0000
    Residual |   317812515    70  4540178.78           R-squared     =  0.4996
-------------+------------------------------           Adj R-squared =  0.4781
       Total |   635065396    73  8699525.97           Root MSE      =  2130.8

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |    21.8536   74.22114     0.29   0.769    -126.1758     169.883
      weight |   3.464706    .630749     5.49   0.000     2.206717    4.722695
     foreign |    3673.06   683.9783     5.37   0.000     2308.909    5037.212
       _cons |  -5853.696   3376.987    -1.73   0.087    -12588.88    881.4933
------------------------------------------------------------------------------

. net search regression output
(contacting http://www.stata.com)

10 packages found (STB listed first)
------------------------------------

sg97_3 from http://www.stata.com/stb/stb59
    STB-59 sg97_3.  Update to formatting regression output / STB insert by
    John Luke Gallup, Harvard University / Support:  jgallup@hiid.harvard.edu
    / After installation, see help outreg

[... further output omitted ...]

. net from http://www.stata.com/stb/stb59
------------------------------------------------------------------------------
http://www.stata.com/stb/stb59/
STB-59 January 2001
------------------------------------------------------------------------------

DIRECTORIES you could -net cd- to:
    ..                Other STBs

PACKAGES you could -net describe-:
    dm73_2            Contrasts for categorical variables: update
    dm83              Renaming variables: changing suffixes
    dm84              labjl: Adding numerical codes to value labels
    dm85              listjl: List one variable in a condensed form
    dm86              Sampling without replacement: absolute sample sizes and
                      keeping all observations
    sbe36_1           Summary statistics for diagnostic tests
    sbe41             Ordinary case-cohort design and analysis
    sbe42             Modeling the process of entry into the first marriage
                      using Hernes model
    sg158             Random-effects ordered probit
    sg159             Confidence intervals for correlations
    sg97_3            Update to formatting regression output
    ssa14             Global and multiple causes of death life tables from
```

```
                    complete or aggregated vital data
-------------------------------------------------------------------------------

. net descr sg97_3

-------------------------------------------------------------------------------
package sg97_3 from http://www.stata.com/stb/stb59
-------------------------------------------------------------------------------

TITLE
      STB-59 sg97_3.  Update to formatting regression output

DESCRIPTION/AUTHOR(S)
      STB insert by John Luke Gallup, Harvard University
      Support:  jgallup@hiid.harvard.edu
      After installation, see help outreg

INSTALLATION FILES                                (type net install sg97_3)
      sg97_3/outreg.ado
      sg97_3/outreg.hlp
-------------------------------------------------------------------------------

. net install sg97_3
checking sg97_3 consistency and verifying not already installed...
installing into d:\stata\ado\stb\...
installation complete.

. * Stata still holds the estimation results in memory...
. * ... and by the way you can use asterisk to insert comments
. * into your do- or log-files
.
. outreg using autoreg, replace bdec(2) se

. type autoreg.out
        Price
Mileage (mpg)   21.85
        (74.22)
Weight (lbs.)   3.46
        (0.63)**
Car type        3,673.06
        (683.98)**
Constant        -5,853.70
        (3,376.99)
Observations    74
R-squared       0.50
Standard errors in parentheses
* significant at 5%; ** significant at 1%

. * the output is not very nice, though, because the columns are
. * tab-delimited for inclusion in a word processor document
.
. search regression diagnostics

[R]     logistic . . . . . . . . . . . . . . . . . . . . Logistic regression
        (help logistic, lfit, lstat, lroc, lsens)

[R]     predict  . . . . Obtain predictions, residuals, etc., after estimation
```

```
        (help predict)

[R]     regression diagnostics . . . . . .  Regression diagnostics
        (help regdiag, avplot, cprplot, lvr2plot, rvfplot, rvpplot, cont.)

[R]     regression diagnostics, continued from above
        (help ovtest, hettest, dwstat, dfbeta, vif)

STB-2   srd3 . . . . . . . . . . One-step Welsch bounded-influence estimator
        (help bound if installed)  . . . . . . . . . . . . . . . .  R. Goldstein
        7/91    STB Reprints Vol 1, page 176
        ols regression output presented, regression diagnostics
        computed, dffits is used to weight the data and estimate
        a one-step Welsch bounded-influence regression

http://www.komkon.org/~tacik/stata/ . . . . . . . . . . . . . . . . Atkinson plot
        (help atkplot) . . . . . . . . . . . . . . . . . . . . . .  S. Kolenikov
        The plot of Box-Cox score test vs added observations
        suggested by Atkinson & Riani (2000)

. hettest

Cook-Weisberg test for heteroskedasticity using fitted values of price
     Ho: Constant variance
         chi2(1)      =      6.34
         Prob > chi2  =      0.0118

. ovtest

Ramsey RESET test using powers of the fitted values of price
       Ho:  model has no omitted variables
                  F(3, 67) =     15.31
                  Prob > F =      0.0000

. predict res, res


. sktest res

                  Skewness/Kurtosis tests for Normality
                                                    ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
         res |      0.000         0.042           14.51       0.0007


. regress price mpg wei for, robust

Regression with robust standard errors                Number of obs =      74
                                                      F(  3,    70) =   15.23
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4996
                                                      Root MSE      =  2130.8

------------------------------------------------------------------------------
             |               Robust
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
         mpg |    21.8536   80.74674      0.27   0.787    -139.1907    182.8979
      weight |   3.464706   .7776165      4.46   0.000     1.913799    5.015613
     foreign |    3673.06   664.9361      5.52   0.000     2346.887    4999.234
       _cons |  -5853.696   3873.723    -1.51   0.135    -13579.59      1872.2
------------------------------------------------------------------------------
```

```
. exit, clear
```

References

[Deaton (1997)] Deaton, A. *The Analysis of Household Surveys*. Johns Hopkins Univ Press (1997).

[Ferrall (1994)] Ferrall, C. A Review of Stata 3.1. *J. of Applied Econometrics*, **9**, 469–477 (1994).

[Gould (2001)] Gould, W. *Report to users at The 1st North America Stata User Group Meeting* (unpublished) (2001).

[Gould and Sribney (1999)] Gould, W., and W. Sribney. *Maximum Likelihood Estimation with Stata*. Stata Press, College Station, Texas, US (1999).

[Greene (2000)] Greene, W. H. *Econometric Analysis*. Prentice Hall (2000).

[Hamilton (1997)] Hamilton, L. *Statistics with Stata 5*. Duxbury Press, 1997.

[Hardin and Hilbe (2001)] Hardin, J., and J. Hilbe. *Generalized Linear Models and Extensions*. Stata Press (2001).

[Johnston and DiNardo (1997)] Johnston, J., and J. DiNardo. *Econometric Methods*. McGraw-Hill (1997).

[Long (1997)] Long, S.J. *Regression Models for Categorical and Limited Dependent Variables*. SAGE (1997).

[McCullagh and Nelder (1989)] McCullagh, P., and J.A.Nelder. *Generalized Linear Models*. 2nd edition. Chapman and Hall (1989).

[Mooney and Duval (1993)] Mooney, C.Z., and R.D.Duval. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage Publications (1993).

[Rabe-Hesketh and Everitt (2000)] Rabe-Hesketh, S. and B.Everitt. *A Handbook of Statistical Analyses using Stata*. 2nd edition. Chapman and Hall (2000).

[StataCorp (2001)] StataCorp. *Stata Statistical Software: Release 7.0*. College Station, Texas, US: Stata Corporation (2001).