Strategically using General Purpose Statistics Packages: A Look at Stata, SAS and SPSS

Michael N. Mitchell

I welcome your comments about this report. You can send them via email to ATSstat@ucla.edu.

The recommended citation for this report is.

Abstract

This report describes my experiences using general purpose statistical software over 20 years and for over 11 years as a statistical consultant helping thousands of UCLA researchers. I hope that this information will help you make strategic decisions about statistical software – the software you choose to learn, and the software you choose to use for analyzing your research data.

Acknowledgements

Revision History

- December 1, 2005. Version .1 (DRAFT) released. Numerous comments received and incorporated.
- December 15, 2005. Version 1.0 released.

Contents

1 Introduction

This report describes my experiences using statistical packages over the last 20+ years, including my experiences as a statistical consultant at UCLA for more than 11 years. As a statistical consultant, I have worked with thousands of researchers and have worked with well over a dozen packages. In any given day, I bounce from helping people using Stata, then SAS, then SPSS, or Mplus, perhaps HLM, maybe LogXact, perhaps LatentGOLD, maybe MLwiN and so forth. I have seen how certain packages have certain strengths and others have certain weaknesses, and that these strengths and weaknesses fall along a large number of dimensions. I have come to believe that data analysts are like carpenters and that statistical software makes up the tools that we use. A carpenter would not buy a screwdriver and conclude that his or her toolkit is complete. Likewise, as data analysts, we may need to draw upon multiple tools (statistical packages) to form a complete toolkit based on the kind of work each of us performs.

While I have worked on creating many web pages that document how to use these statistical packages for a variety of tasks, I realized that I have never documented what I have learned about the strengths and weaknesses of statistical packages. I am not referring to just to whether a package will or will not perform such and such task, but how well it does the task and my experience observing researchers use these different tools for the tasks. While this report will compare (based on my experiences) the features of the packages and consider their strengths and weaknesses, it will also consider many other important dimensions – what operating systems the packages support, what their licensing policies (at UCLA) are, what it is like to install the packages and keep them up to date, what kind of support infrastructure exists for the packages, and so forth. After considering these matters the features (both general and specific) will be considered. The feature comparison will focus on features that I have frequently encountered as a consultant and features where I have found a very substantial difference among the packages. Features that I have not frequently used in consulting or features where I found only minor differences will not be treated here.

I wish to emphasize in the strongest terms that comments and thoughts here are strictly my own and reflect my personal experiences and observations based on visits with UCLA researchers while assisting them with questions about their research projects. While I have worked with people from dozens of schools and departments including Anthropology, Biology, Education, Health Sciences, Linguistics, Political Science, Psychology, Sociology, and many more, these experiences are certainly not a random sample of UCLA researchers and most certainly not a random sample from the broader research community of the world. Please bear this in mind as you weigh how relevant my observations and experiences are for you. Or, to put it another way, *Your mileage may vary.*

Due to changes and evolution of software, this report may become outdated, so please bear in mind the date it was written (shown on the cover page) and the software versions it covers. Note that these packages will be covered in this order, which is based on the size of the user base who comes to our consulting group for help with each of the packages, from largest (Stata) to smallest (SPSS). Section 12 discusses packages not covered in this discussion.

- Stata 9.1
- SAS 9.13 (service pack 3)
- SPSS 14

Statistical software is a tool in the same way that a screwdriver or a hammer is a tool. I know, however, that sometimes people can be as partisan about their favorite statistical software as they can be about their favorite sports team, their alma mater, or their political party. I hope that in reading this report people can be as dispassionate about statistical software as they would be about hammers and screwdrivers. In this way, I hope this report can help people set aside such emotional attachments to their favorite software and take a fresh look at each software package and how well it suits their needs.

2 Making Strategic Choices

When assembling a bicycle or building a doghouse, people select tools from their toolbox **strategically**, choosing a hammer to pound nails, a wood saw to cut wood, a wrench to tighten bolts, or a drill to drill holes. While you can pound a nail with a screwdriver in an emergency, no one would routinely use a

screwdriver to pound nails because that is not the most effective tool. It is a bad strategy. However, I have frequently seen people do the same sort of thing as part of their data analysis, choosing a tool (i.e., statistical package) that is as ill suited for their purposes as hammering a nail with a screwdriver. Rather than choosing their statistical software tools strategically, I frequently see people using other means for choosing the statistical software tool they use. They may choose a tool because they have used it before, or because they have heard that it is good, or because it is what they have always used, or because it was easy to use or had friendly pull-down menus, or it is what has been traditionally used by people in their field. In fact, I think people most heavily weigh three factors: 1) the package their friends use, 2) the package that is used in their department, 3) the package their professor uses. These kinds of peer-based factors are not trivial. They may drive your ability to draw support from your peers or to be able to exchange information and skills. However, your research project may be different from the kind of research that your friends or your department typically performs, and perhaps you have a need for a different suite of tools. Even if you have a good support network, that might not overcome the fact that they are providing you helpful advice on using a screwdriver to pound a nail. I hope that by reading this report more people will say to themselves, "I assessed my needs for this project and compared my needs to the strengths of these packages and decided that this tool offered the best strategies for working on this project."

I have often seen many data analysis projects that have progressed more slowly or more painfully because non-optimal software was used. I have seen many clients who did not realize that they had been pounding nails with a screwdriver for a very long time. Hopefully this report will help you think about your needs and the tools that will most effectively help you meet those needs. You can then, on a project by project basis, assess the particular strengths of different statistical packages and then strategically choose the software you will use. I hope, however, that you will see that this is not just about picking a single package and sticking with it, but it is about learning a toolkit of packages, a toolkit that is composed of general purpose packages and specific purpose packages (although this report will focus on general purpose packages). I hope that after reading this report that you will take to heart the following points (if you have not done so already).

1. It is important to know more than one statistical software package. A single package cannot be the most strategic tool for all of your projects.
2. One should strategically choose which statistical software packages to learn. If you are going to learn multiple packages, then it is useful to take some time and think about your needs with respect to data management and statistical analysis and choose the packages that most strategically address your needs.
3. As you choose packages, not only you want to assess the strengths of the packages alone, but also consider the breadth of coverage of data management and statistical techniques that you achieve with the combination of packages. By considering the coverage the packages provide in concert, you can strategically choose a combination of packages that provides the greatest coverage while learning the fewest number of packages.
4. Selecting a package mainly based on ease of use can deny you features available in programs that may not be that much harder to use. While a package may have an initial ease of use in the first days or couple of weeks, you will likely use a statistical package over months or years, so the initial ease of use in the first days or weeks is trivial compared to the long-term strategic usefulness of the package for your research.
5. It is important to choose the packages that will allow you to use a statistical model that is most appropriate for your data (i.e., a model that most closely fits the behavior of your data).
6. Some statistical packages make it easier to create graphs or tables to interpret your results. It is wise to choose a package that makes it easiest to present your findings in a form that is most easily interpreted.
7. It is penny wise and pound foolish to skimp on purchasing useful statistical software. Statistical software can be one of the most powerful tools for testing your hypotheses and advancing your research. Using an inappropriate tool to save a modest amount of money is more costly in terms of your time, and more costly in terms of handicapping your ability to put forth the most compelling statistical case possible. Consider the value of software by weighing its costs against the benefits it provides.
8. It is unwise to skimp on obtaining useful updates to statistical software. Some updates can provide tremendous increases in functionality that make it substantially easier to analyze your data. Assess the features provided by updates and weigh the cost of updating against the benefits you receive. Look

at the past track record of the statistical software that you use in terms of the usefulness of updates, the cost of updates, and the timing of updates. Use this to plan future budgets for software.

9. When considering the value of software, it is important to distinguish between general purpose packages (such as Stata, SAS, and SPSS) and specialized purpose software. General purpose software sells zillions of copies, while special purpose software sells far fewer copies. The price of specialized software is, of economic necessity, higher than general purpose software. It is not fair to compare the price of special purpose software to general purpose software and conclude that the special purpose software is overpriced. To determine the value of special purpose software, it is important to compare the price and features among software packages in the same league.

10. As a dedicated researcher, you are going to spend years using statistical packages. No matter what stage you are at, it is never too late to reconsider the suite of tools that you use.

11. All researchers should periodically evaluate the suite of tools they use in relationship to the state of the art. Over time some tools that you use may become ill suited to your needs and other tools may evolve that suit your current and future needs.

12. It may be advantageous to use more than one statistical package for a project. Never before has it been so easy to transfer data among statistical packages, so there is only a minimal cost to converting your data to another package so you can exploit the unique features of a different statistical package. Tools like Stat/Transfer permit you to convert data files from one file format to another often in less than a minute. Using a tool like this enables you to move your data seamlessly from one package to another.

As you can see, based on these comments I am taking a long-term perspective focusing on long-term dividends, not short-term benefits. I am assuming that data analysis is a central focus for you and that the extra short-term efforts that you invest to advance your research will pay rich dividends later, yielding much greater long-term gains. This is an important distinction because if data analysis is not your central focus, then some of the advice here may advocate that you make short-term investments for a long-term benefit that is not relevant for you. Also, I do not think this report is more relevant for those who are early in their research career than those who are more established in their career. At any stage of your research career, I think it is important to make strategic choices about the statistical packages that you use.

In addition, I think this report could be of interest to those who mentor people who analyze data, either via teaching courses or advising students in their research. Even if teachers and mentors are not directly using these statistical packages very much, I think that this report would be helpful for them to help students understand the strategic importance of choosing statistical packages and to encourage them to learn multiple packages as part of their training.

3 In Their Own Words...

Before going further into describing my experiences with these statistical packages and my comparison of their features, let's hear the statistical package vendors speak for themselves as to how they portray their company, the features of their products, and the evolution of their company and products over the recent years. By seeing how the companies portray themselves, their products, and their history, I think you can gain important insights into the focus of the companies and the focus of their products. Let's see what Stata, SAS, and SPSS say about their own products.

Stata

The Product Today. You can find the Stata web site at `http://www.stata.com`, which has quite a bit of information about Stata. They have a page entitled Why Use Stata? at `http://www.stata.com/whystata/`, which describes StataCorp's view of the strengths of Stata. This is elaborated further on the Stata capabilities page at `http://www.stata.com/capabilities/`, which describes details about the data management and statistical features of Stata. The page `http://www.stata.com/help.cgi?estimation commands` provides a quick reference of the Stata estimation commands along with more details.

In addition to these pages, I would recommend the Report to Users that Bill Gould (the President of StataCorp) gave to the 3rd UK Stata User Group meeting in 1997. This report describes how Bill Gould

conceptualizes Stata not just as a statistical package but as a Statistical Operating System that allows users to modify and extend its functionality. You can read the notes from that talk at `http://www.stata.com/meeting/3uk/report.html`.

History of StataCorp and Stata. The Stata Journal (Volume 5, Number 1, `http://www.stata-journal.com/sj5-1.html`) contains a number of articles describing the history of Stata, including an interview with William Gould. The Stata web site has very little about the history of StataCorp itself, but does have detailed information about the history of the program. In fact, you can still see the web pages that announced Stata versions 6, version 7, version 8 and the latest, version 9. Stata 6 was released around January 1999 (and each subsequent version approximately every 2 years). You can find these pages at:

- `http://www.stata.com/stata6/`
- `http://www.stata.com/stata7/`
- `http://www.stata.com/stata8/`
- `http://www.stata.com/stata9/`

You can see an even more detailed list of the history of Stata features going all the way back to version 6 by viewing the `whatsnew` help file by visiting `http://www.stata.com/help.cgi?whatsnew`.

SAS

The Product Today. You can find the SAS web site at `http://www.sas.com/`, which has quite a bit of information about SAS. The Tecnologies/Analytics page describes SAS features in Analytics at `http://www.sas.com/technologies/analytics/`. From that page you can find the Statistics page at `http://www.sas.com/technologies/analytics/statistics/index.html`. From there you can find the SAS/STAT fact sheet at `http://www.sas.com/technologies/analytics/statistics/stat/factsheet.pdf`. You can see the success stories that SAS describes regarding the use of Data Mining and Statistical Analysis using SAS at `http://www.sas.com/success/technology.html#DataMiningandStatisticalAnalysis`. The SAS support web site has a number of resources regarding statistical analysis in SAS at `http://support.sas.com/rnd/app/da/stat.html`.

History of SAS Institute and SAS. At one point in time SAS was an acronym that stood for Statistical Analysis System, but now the name is simply SAS. You can learn a great deal about the history of SAS from `http://www.sas.com/corporate/` . This includes information about the history of SAS at `http://www.sas.com/presscenter/bgndr_history.html`. You can see the history of the SAS software by viewing the What's New sections of the SAS documentation at the following links:

- 9 at `http://support.sas.com/documentation/onlinedoc/stat/`
- 9 at `http://support.sas.com/software/91x/statugwhatsnew900.htm`
- 8.2 at `http://support.sas.com/documentation/onlinedoc/v82/whatsnew/zenid-96.htm`
- 8.1 at `http://support.sas.com/documentation/onlinedoc/v81/whatsnew/zenid-30.htm`
- 8.0 at `http://support.sas.com/documentation/onlinedoc/v8/whatsnew/tw5508/znid-162.htm`

SPSS

The Product Today. You can find the SPSS web site at `http://www.spss.com`, which has quite a bit of information about SPSS. The SPSS for Windows page at `http://www.spss.com/spss/data_analysis.htm` describes features of SPSS for Data Analysis. The page `http://www.spss.com/vertical_markets/education/` describes SPSS's role in supporting excellence in education. The SPSS web site has a number of references to their role in Predictive Analytics described, for example, at `http://www.spss.com/predictive_analytics/`.

History of SPSS Inc. and SPSS. You can see `http://www.spss.com/corpinfo/history.htm` for a history of SPSS Inc. The page describes SPSS in 2005 by stating "Today, SPSS is recognized as a leader in the nascent predictive analytics market space. Predictive analytics, which combines advanced analytics and decision optimization, will continue to be a focus for the organization as it seeks to increase marketplace understanding of the business benefits that predictive analytics provides."

For a history of SPSS software, you can see the features that have been added to SPSS over time by visiting `http://www.spss.com/software_version/index.cfm?product=base` and then specifying a version number. For example, specifying version 7.x allows you to see the features added to SPSS from version 7 to the

present. (The page `http://www.ucs.ed.ac.uk/usd/stats/spssversions.html` also has this information presented in a convenient bullet list.)

4 Operating System Support

Although there are a number of operating systems in the world, I mainly see people using three operating systems: Windows, OS X, and UNIX (in its various incarnations, including Linux). This section considers how well each package supports each of these operating systems for use by end users (clients) on each platform.

Stata

Stata is fully supported on Windows, OS X, and UNIX/Linix. Stata is fully up to date on all of these operating systems, and has a long history of fully supporting all of these operating systems. In fact, Stata is the only one of these three packages that is fully supported and fully up to date for OS X, and is the only one of these packages that fully supports all three operating systems.

SAS

SAS is fully supported on Windows and on UNIX/Linux, and is fully up to date on these two operating systems. However, SAS version 6.12 is the latest version for the Macintosh and runs only on OS 9 (does not run under OS X).

SPSS

SPSS is fully supported and up to date on Windows. To the best of my knowledge, SPSS 6.1 is the most recent version of SPSS available for UNIX which allows users to directly run SPSS on UNIX[1]. SPSS does not have a history of keeping the Macintosh version up to date. For a long time, the latest version of SPSS available on the Macintosh was version 11, but recently version 13 has been offered (even as the most current version of SPSS is version 14 on Windows).

5 Licensing Policies

Not all license agreements are equal. In fact, two products might appear to cost the same, but due to extreme differences in licensing, one product may be much more expensive to license compared to the other. While I think that it is more important to focus on the features of statistical packages, in these cost-conscious times it is also important to consider the value of software, comparing features against price. In order to assess the price of a statistical package, a number of factors are important to consider: 1) the cost, 2) the duration of the license (i.e., annual vs. perpetual), 3) how many machines the license permits you to install the software on, and 4) whether there are extra costs for modules that are included for free in other packages. The prices and policies mentioned in this report are for members of the UCLA community, but even if you are not at UCLA this discussion may help you assess the factors that determine the cost and value of statistical software based on the licensing arrangements at your university.

The licensing prices and policies described here are for members of the UCLA community only. The prices and policies may be much different for your university. UCLA cannot help you obtain or license software unless you are currently a member of UCLA. Pricing can vary from year to year. Pricing information is current as of 7/25/06.

---

[1]SPSS offers various server editions that run on flavors of UNIX but these require the SPSS client to run on the users computer, not permitting users to directly connect to the UNIX computer and run SPSS, e.g. in batch mode

Stata

The Stata licensing policies are summarized here. You can see the licensing information for yourself, written in plain English, on your one page Stata license (where your license codes were printed). Stata is available directly from StataCorp.

Cost A perpetual license for Intercooled Stata costs $145, and Stata/SE costs $295.

Duration Perpetual Stata licenses permit you to use the version of Stata you purchased forever[2]. When a new version of Stata is released, you can continue to use your perpetual license but you will not be able to enjoy the benefits of the new version.

Installations You can use your license code to install Stata on multiple computers provided that Stata is not used more than once at any given time.

Extra costs - You do not have to pay extra for separate modules – the price you pay for the Stata license entitles you to all of the features that Stata provides.

SAS

The SAS license terms are summarized here, and you can see `http://www.ucop.edu/purchserv/agree/tas/sas.html` for more details. SAS is available under a UC systemwide agreement that is coordinated by Software Central for members of the UCLA community.

Cost A SAS EAS Workstation license costs $77 per academic year.

Duration The SAS license is an annual license, meaning that it entitles you to use SAS for the duration of the academic year in which you obtained it. If you paid for a license in January, that entitles you to use SAS until June 30 of that year. After that time, your license expires and needs to be renewed (possibly at a different rate).

Installations You need one SAS license per CPU on which SAS will be installed. One copy for home use is permitted for each copy licensed; see `http://www.ats.ucla.edu/software/SAS_Home_Use_Form.htm` for details.

Extra Costs For academic researchers (non-administrative users), your SAS license entitles you to use the SAS Education Analytic Suite which includes over 20 SAS products (in other words, almost everything and the kitchen sink); see `http://www.ats.ucla.edu/software/product_list.htm` for more details. While there may be some SAS products that SAS offers that you could license, I think 99.9% of researchers would have no need for additional SAS products.

SPSS

The SPSS license terms are summarized here, and you can see `http://www.ats.ucla.edu/software/product_list.htm#spss` for more details. The SPSS agreement (excluding GradPack) is a volume purchase coordinated for UCLA by Software Central. The cost and duration is different for students, as described below.

Student Cost and Duration Students may obtain the SPSS GradPack from ASUCLA for $199. This license lasts about 4 years and permits SPSS to be installed twice.

Faculty/Staff/Department Cost and Duration Faculty/Staff/Departments obtain SPSS from Software Central for $62 per academic year. This is an annual license, meaning that a SPSS license from Software Central entitles you to use SPSS for the duration of the academic year in which you obtained it. If you paid for a license in January, that entitles you to use SPSS until June 30 of that year. After that time, your license expires and needs to be renewed (possibly at a different rate).

Installations You need one SPSS license per CPU on which SPSS will be installed. The license code that you receive will work only once. If you reinstall Windows, reformat your hard drive, or replace your computer, you will need to request a replacement license code (and, according to the license agreement, there may be "possible fees associated with the transfer").

---

[2]Students can buy a one year Intercooled Stata license for $89. This is a good choice if you believe that the next version of Stata will be released within a year and you intend on updating to the next version

Extra Costs SPSS software is broken up into numerous modules. Students obtaining the GradPack receive the Base, Advanced, and Regression modules, as well as AMOS. Those obtaining SPSS from Software Central receive the same modules, except for AMOS.[3]

SPSS offers modules, at extra cost, that perform tasks that are included in the price of SAS and Stata. Here are some examples.

- Unlike Stata and SAS, with SPSS you will have to pay extra for tools for analyzing survey data at a cost of $499 per CPU (and as Section 8 notes, the features SPSS provides are not nearly as extensive as those Stata includes at no extra cost).
- Unlike SAS, you will have to pay extra to obtain tools for dealing with Missing Data, at a cost of $399 per CPU (and, as described in Section 8, SPSS only supports single imputation while SAS permits multiple imputation and free user-written multiple imputation routines are available for Stata).
- The SPSS agreement does not include tools for time series analysis. Stata includes tools for time series analysis and the UC agreement for SAS also includes tools for time series analysis. Those wishing to perform time series analysis would need to obtain SPSS Trends at an extra cost ($499 per CPU).
- When purchasing these add-on products, it appears that maintenance is required to obtain technical support and updates for these extra modules at varying costs in the neighborhood of $80 to $100 per license, per year.

Cost Comparison - Scenario 1

Imagine you are a student and you wish to use statistical software on your laptop and home machine for about 1.5 to 2 years. Analyzing survey data is an essential part of your research. Here are the approximate costs associated with meeting these requirements for each package.

Stata You would obtain one Stata license, which you could install on your desktop and home machine, provided that you do not invoke Stata more than once at a time. Stata includes tools for survey data analysis. The total cost is $145 for a perpetual license.

SAS A student would obtain one license for his/her laptop and then be permitted home use for that licence. SAS includes tools for survey data analysis (although not nearly as extensive as Stata, see section 8). The total cost would be approximately $144 for one license for two years (the cost for the license could increase in the second year). Note, however, that SAS is only available via Software Central, which can sell licenses to departments but not directly to students.

SPSS As a student you would obtain an SPSS GradPack license which permits you to install it twice, once for your laptop and once for your desktop. You would also need to obtain an SPSS Complex Samples license (to analyze survey data, although this module is far less comprehensive than what Stata offers at no extra cost) at an additional cost of $599 (including the first year of maintenance). Each subsequent year would cost $100 for maintenance on SPSS complex samples. The total cost after two years would be $199 for the SPSS GradPacks, and $699 for SPSS Complex Samples for a total cost of $898.

Cost Comparison For two years of use, the cost for Stata would be $145 (assuming you did not upgrade), approximately $144 for SAS, and $898 for SPSS.

Cost Comparison - Scenario 2

Imagine you are a faculty member and you wish to know the price to use your statistical software on a desktop, laptop, and your home computer for the next two years.

---

[3]When I installed SPSS it appeared that AMOS was included in the UCLA license. I installed AMOS and used the AMOS License Wizard to apply the license code and was given a message indicating that the product was successfully licensed. When I clicked on AMOS, the product ran successfully. However, after a few weeks, AMOS stopped working. When I contacted SPSS technical support, I was told that the message indicating that the product was successfully licensed was actually spurious and in fact AMOS was not part of the UCLA license agreement. It only appeared that AMOS was running under the license code I supplied because it was using a temporary test license. There was nothing about the licensing process or running of AMOS that even suggested that I was using a temporary license. If you obtain SPSS from UCLA Software Central, be advised that despite all appearances that AMOS is included as part of the license, that your copy of AMOS will stop working at the conclusion of the trial period and you will need to purchase an AMOS license for it to continue working.

Stata You would obtain one Stata license, which you could install on all three computers, provided that you do not invoke Stata more than once at a time. Total cost, $145 for a perpetual license.

SAS You would obtain one license for your desktop and then be permitted home use for that licence, and then obtain a second license for your laptop. The total cost is $144 per year for the two licenses, or approximately $288.

SPSS You would obtain three SPSS licenses for your three computers at a cost of $186 per year, or for a total of approximately $558.

Cost Comparison For two years of use, the cost for Stata would be $145 (assuming you did not upgrade), for two years of SAS use you would pay approximately $288, and for two years use of SPSS you would pay approximately $558.

Cost Comparison - Scenario 3

Imagine you are a PI for a grant and you use statistical software on your desktop, your laptop, and your home office. You use survey data and you have issues with missing data. Your project will last three years, during which time Stata will offer a new release and you will choose to obtain the new version (although you do not have to).

Stata Since you are the only one using your computers, you can obtain a single Stata license. Stata includes tools for survey data analysis, and add-on tools have been created that deal with multiple imputation for missing data which you could use for your missing data issues. Your total cost would be $145 for a perpetual license and then approximately $145 to purchase the new version (the new version may cost more). The total cost is approximately $290. (Note that you probably would be purchasing an entirely new license, so you would then own two perpetual Stata licenses, one for the current version of Stata and a perpetual license for the prior version of Stata, which you could use or transfer to a student or other person at your university, after checking with Stata).

SAS You would obtain one license for your desktop and then be permitted home use for that license, and then obtain a second license for your laptop. SAS includes tools for survey data analysis (although not as extensive as Stata; see section 8) and tools for multiple imputation of missing data. Total cost, $144 per year for the two licenses, or a total of $432 for the three years.

SPSS You would obtain three SPSS licenses for the three computers at a cost of $186 per year or approximately $558. You would also obtain three SPSS Complex Samples Licenses to analyze survey data (although this tool is far less comprehensive that what Stata includes at no extra cost) at an additional cost of $1797 (including the first year of maintenance). You would also obtain three SPSS Missing Values Licenses (for missing data issues, although this tool only handles single imputation while SAS and Stata both perform multiple imputation; see section 8 for more information) at an additional cost of $1,497 (including the first year of maintenance). The second and third year of maintenance for Complex Samples would be $300 per year, and the second and third year of maintenance for Missing Values would be $240 per year. The total cost would be $4,392.

Cost Comparison The cost for Stata for the three years would be approximately $290, for SAS approximately $432, and for SPSS approximately $4,392. It is important to bear in mind that although SPSS is much more costly, the SPSS complex samples module has far fewer features than those included at no extra cost with Stata, and the SPSS missing values module does not support multiple imputation, which SAS and Stata can do at no extra cost.

6 Installation and Updates

Let's compare Stata, SAS and SPSS in terms of installation and updates.

Stata Installing Stata is rather simple. Stata installs from a single CD. The most difficult part of the process is typing in your license code correctly. It probably takes about 10 minutes to install Stata and perhaps another 5 or 10 minutes to bring it fully up to date. I have never encountered a problem installing Stata. I believe that Stata consumes about 30 to 50 megabytes of hard disk space.

Stata Stata comes configured to automatically check over the web for updates every 7 days. When updates are found, Stata prompts you to obtain the updates. If you say yes, Stata downloads and installs the updates for you. The process usually takes under 5 minutes. Updates are posted frequently, and can include enhancements as well as fixes. You can type `help whatsnew` for a list of what is new in the updates.

SAS Installing SAS is rather complicated. First, SAS often needs to install updates to your operating system that can require multiple reboots of your machine. During the installation you are asked numerous questions that you may not necessarily know the answer to. The most difficult question regards which modules you wish to install. There are many choices and the safest choice is to install everything, which consumes in excess of 1 gigabyte of hard disk space on your computer. You can try to select fewer modules, but then you might find you skipped modules that you later wanted. There are about 10 CDs in our SAS installation kit, so you need to switch CDs many times. It has often taken me about 45 minutes to install SAS. Then, after installing SAS you may need to install a service pack which comes on an additional CD with the installation package. Installing the service pack can take an additional 20 or more minutes. I have had problems installing SAS many times and have also had problems installing the service packs. I also have had numerous clients ask me for help with problems installing SAS. However, the SAS technical support team is very familiar with installation problems and my experience, as well as what I have heard from others, is that the SAS technical support is excellent at solving installation problems.

SAS updates are delivered as individual "Hot Fixes", each of which fixes a specific problem. Since each "Hot Fix" solves a very specific problem, there could be quite a few (perhaps even dozens) of "Hot Fixes" that could be applicable to your version of SAS. To help with this, numerous "Hot Fixes" are combined together to form a "Service Pack" which can be downloaded and installed all at once. The "Service Packs" can be hundreds of megabytes in size and are daunting to download except on fast internet connections. The "Service Packs" can be downloaded from the SAS web site or you can obtain a CD from your SAS representative. It can take well over 20 minutes to install the service pack, and I have had and seen problems installing the service packs, sometimes meaning that the service pack had to be installed again, or help from technical support was required to complete the installation. Note that even when you have installed the latest service pack, this does not mean that your copy of SAS is fully up to date – it is only as up to date as the service pack you applied. There could be a number of "Hot Fixes" that have been issued since the last service pack was issued.

SPSS SPSS is generally quick and easy to install. However, applying the patches to update the software is not nearly as easy as it could be. I believe that SPSS consumes about 100 or so megabytes of hard disk space. There is nothing that automatically notifies you of whether there are patches available and how you should obtain and apply them. However, SPSS 14 added a `Check for updates` option under the `Help` menu which tells you if your SPSS is fully up to date. But once you are aware that you need to obtain patches, you need to login to the SPSS support web site to obtain the patches, which is password protected. In order to access the updates, you have to create a user id and password and you must provide your name, company, phone and email address to do so.

7 Support Infrastructure

Another important aspect to consider is what I call "support infrastructure", the suite of supporting materials that help you learn, use, and extend the functionality of your statistical software. As a consultant, I work with this support infrastructure every day, both in seeking information for myself and in referring clients to these extra forms of support. One of the most positive trends I have seen over the last ten years is the incredible growth of this support infrastructure and the way such infrastructure can enhance the usefulness of statistical software Below I describe and characterize my observations and experiences with the support infrastructure provided by Stata, SAS, and SPSS.

Stata

Technical Support All those purchasing Stata are entitled to free technical support (for the current and previous version of Stata). Stata provides excellent and very responsive support.

Web Site The Stata web site (`http://www.stata.com/`) contains numerous useful resources and is laid out in an intuitive manner making the information easy to find. The site gives easy access to the descriptions of the products Stata provides and resources and support for Stata, including FAQs, Tech Support, information on Statalist, training and so forth. The FAQs cover many useful topics of interest to researchers and are organized intuitively, based on the topic area of interest. I often read the FAQs just as a way of expanding my knowledge, often finding useful information that often goes beyond just the nuts and bolts of how to use Stata. The locations of the pages (i.e., the URLs) are incredibly stable and the URLs for the pages almost never change. Almost all of the links I have created to inner parts of the Stata web site have not moved in the last five years. Also, the site is searchable by search engines like Google, so a general web search may display their pages if relevant.

Documentation The Stata documentation (`http://www.stata.com/bookstore/documentation.html`) is easy to use, uses a consistent and intuitive format, includes background explanation for statistical commands, provides examples (sometimes numerous examples) for statistical commands, and offers easy access to the data files via the `webuse` command to allow you to replicate their examples. In short, I feel the documentation is excellent. Please note, however, that while the Stata documentation is excellent, it is only available in printed form, unlike SAS that provides its documentation over the web and SPSS that provides the documentation as .pdf files on installation CD.

Online help Online help is available either via the "help" menu, with topics organized in a hierarchical fashion, or by simply using the `help` command. Typing `help ttest` brings up help on the `ttest` command. You can obtain help for a command within seconds. This online help often quickly provides you instant access to the information that you need to use a command. The online help always ends with cross-references that are hotlinked to related commands, so with a simple mouse click you can jump to related commands. You can also access this online help via the web by visiting `http://www.ats.ucla.edu/stat/stata/` and typing in the name of a Stata command in the search box under *Lookup help for Stata command*. Also, a google search, for example for `stata findit`, will often include hits for the online help near the top of the list. Although the online help is not a replacement for the written manual (and is not intended to be so), but I find this online help to be excellent as a quick reference.

Additional Books The publishing arm of Stata, Stata Press (`http://www.stata-press.com/`), publishes books that go beyond the documentation to help users enhance their use of Stata. The books are often very useful and example driven. (Full Disclosure! I am the author of one of the books published by Stata Press.) While the offerings of Stata Press are not as extensive as SAS Publishing, I think their offerings are terrific and very useful.

Listserver Statalist (`http://www.stata.com/statalist/`) is an independently operated listserver operated and maintained by members of the Stata community. Statalist is generally populated by questions from other researchers on questions that are highly relevant to the kinds of research that I see at UCLA. The answers reflect the incredible talent, generosity, and commitment of the Statalist community. The answers are often very extensive and responsive (i.e., sometimes answers are posted within minutes). I think reading the Statalist is a great way to enhance a researcher's skills. Sometimes Bill Gould, the CEO of Stata, jumps in with answers providing detailed and extremely technical answers to questions posted by users.

Add-on programs Stata provides the infrastructure to permit users to create new commands that can be found, downloaded and installed into Stata that work in the same way as built-in Stata commands. The Stata `findit` command searches for user-written programs and helps you download and install the programs onto your computer. Often in a matter of a couple of minutes you can find additional programs to enhance the functionality of Stata. Nearly all add-on programs use the same general syntax as regular Stata commands and contain help files that can be accessed by via the `help` command by typing `help` followed by the name of the program. Once you have downloaded a program, it is often rather simple to start using it. In addition, Stata interfaces with the Statistical Software Components (SSC) via the `ssc` command to provide a common download site for user-written software on the web, providing a commonly accepted repository that is easy to access.

Stata Journal The Stata Journal (`http://www.stata-journal.com/`) is published quarterly and contains articles about statistics, data analysis, teaching methods, demonstrating how to use the functionality of Stata or to enhance its functionality via user-written programs. I have found these articles very

useful and to be highly relevant for the kind of work UCLA researchers perform.

Training Stata provides online NetCourses (`http://www.stata.com/netcourse/`) on a variety of topics including Introduction to Stata, Stata Programming, and Survival Analysis with Stata. I have found these NetCourses to be very informative and well targeted to the needs of UCLA researchers. The Netcourses are generally priced between around $100 to $150.

SAS

Technical Support All those who have licensed SAS are entitled to technical support. SAS provides excellent and very responsive support.

Web Site The SAS support web site (`http://support.sas.com/`) contains some useful resources, but they are extremely hard to find and can take a painstakingly long time to search. A search of the FAQs sometimes returns hundreds of results and can be overwhelming to search, even for those seasoned with SAS. The phrase "Sometimes Less is More" often comes to mind. The site is heavily loaded with marketing materials and information pitched towards business, so although there can be useful information on the site, finding useful information targeted towards researchers and applied statisticians can be difficult. This is complicated by the fact that the site is frequently overhauled and the URLs for many pages are changed. The half life of a link on the SAS site seems to be somewhere around 3 months. In the past, I would link to useful gems at the SAS web site from the UCLA ATS web site. However, after having these links die so soon after creating them, I gave up on creating such links. Many parts of the SAS site, such as the FAQs, do not appear to be indexed by search engines such as Google, so a general web search using something like Google will not find such pages. However, I have found a useful page that appears to be a source of example problems illustrating how to do a variety of tasks in SAS at `http://support.sas.com/ctx/samples/index.jsp`. This has some very nice examples laid out in a very clever fashion. I hope this link will be a stable one.

Documentation The SAS documentation (`http://support.sas.com/onlinedoc/913/docMainpage.jsp`) is available online, so there is the benefit that many can save money by not needing to buy the documentation. However, the price is that the documentation can be extremely difficult to use. Given that there are well over 30 different products that can be searched, it is often difficult just to figure out which part of the documentation you should be searching for help. While the online documentation includes a search tool, once you use the search and find a relevant portion of the documentation, I have never been able to figure out how to navigate to the table of contents for that chapter so I could access the entire contents of the chapter for that topic. I have found that clients, even experienced clients, frequently have difficulty navigating the documentation to find what they are seeking. However, the documentation is generally very comprehensive and uses a clear and predictable structure. The tone of the documentation is often cold and a touch difficult to digest. The statistical procedures include examples, but I have often found them to be rather complicated. More simple examples would be most welcome.

Online help I have found the SAS online help within SAS to be even harder to use than the documentation over the web. The online help is hard to search, and I feel that I spend more time searching that than it would take to access the web site. Even when I find the right area within the online help, I felt the information was often too sparse to be useful.

Additional Books SAS Publishing (formerly SAS Books by users) is the publishing arm of SAS, see `http://support.sas.com/publishing/index.html` for more details. They have, by far, the most extensive collection of books that go beyond the documentation including a very large and extensive list of high quality books that explain how to use SAS for data management and to perform statistical analysis. The statistical analysis topics include Regression, ANOVA, Multivariate Statistics, Survival Analysis, Logistic Regression, and more. I have found these books to be an excellent resource for researchers using SAS.

Listserver The SAS newsgroup is an independently operated newsgroup operated and maintained by members of the SAS community. The SAS newsgroup is populated by an extremely heterogenous set of topics mixing questions from those who use SAS in business with those who use SAS in an academic/research oriented setting, meaning that I often have to sift through lots of questions that are not research related. It is unfortunate that there is not a separate SAS newsgroup that targets

academic/research topics which would be more focused on the questions raised by this community. Nevertheless, I believe that the SAS newsgroup taps into an extremely talented community that is very generous with their information. You can find information about the SAS newsgroup by visiting `http://www.ats.ucla.edu/stat/otherresources.htm` and seeing the section on SAS.

Add-On Programs There are many SAS macros that users have written, but SAS does not provide any infrastructure to help users find, download, or install the macros. Even if you do find and install a SAS macro yourself, there is no standard syntax for calling the macros and no standard means for providing help for macros. In short, while there may be good SAS macros, they are much harder to find, install, and use than they could be.

Journals I know of no journals published by SAS on topics of statistics and data analysis. SAS does publish SAScom magazine (see `http://www.sas.com/news/sascom/index.html`) but it is composed of topical articles targeted towards business and IT decision makers, and I think it contains little of interest to academic researchers.

Training SAS has OnlineTutor software covering topics such as data management and creating reports. My experience is that the topics are presented in a "programmed learning" fashion. Small bits of information with lots of graphics are presented, followed by quizzes to test whether you learned the information. The information is presented in a very linear fashion, making it difficult to randomly access the material according to your interests and the speed you wish to proceed. SAS also has live training opportunities at a number of locations, the most convenient for UCLA members being the Irvine location. Most of the topics are geared towards business tasks, such as *Introduction to SAS Business Intelligence Applications* or *Using SAS Web Report Studio for Thin-Client Reporting*. Even a topic like *Predictive Modelling Using Logistic Regression* would sound relevant for researchers, but the UCLA consultants found it very heavily targeted towards business applications and not so relevant for academic researchers. We have had to scrutinize the courses very carefully to find courses targeted towards researchers. In all cases, the experience of our consultants is that whether the topics were well targeted towards researchers or not, the courses are well taught and include opportunities for hands on experiences. The fees for these courses vary considerably, as little as \$500 for a one-day course to \$1500 for a three-day course, but members of the UCLA community should be eligible for a 50% discount on these courses.

SPSS

Technical Support Students buying the SPSS GradPack are **not** entitled to technical support (beyond installation support); however students can purchase the ability to access SPSS technical support. Those who license SPSS via Software Central at UCLA may **not** contact SPSS directly for technical support. Instead, support requests must be routed to Software Central. These support policies are very different from Stata and SAS, which permit all license holders to directly contact their technical support staff for help.

Web Site The SPSS web site (`http://www.spss.com/`) contains very few useful resources to assist academic researchers. The site is heavily oriented towards providing information for those in the business community, with pages, for example, focusing on "predictive analytics" and helping business to build a "predictive enterprise". Likewise, their White Papers (see `http://www.spss.com/downloads/Papers.cfm`) focus mainly on business topics. In order to access the support section you have to create a user id and password and you must provide your name, company, phone and email address to do so. Each time you want to access the support section, you need to login providing your userid and password. Accessing the support section is often annoying since I can never remember my userid and password for this site and need to recover this information. I have not found the information within the SPSS support site to be very useful. The "Resolution Search" searches what appear to be answers to previous email questions. The format of the answers is very hard to follow since it is all plain text and program snippets are provided which are all double spaced and very hard to read. I could not find any answers that integrated program output to help you understand what the code was doing. Even when there are useful answers, they are difficult to access because they are in the password protected portion of the site and would not come up in a general web search using something like Google. The frequently used resolutions and frequently asked questions were largely composed of questions related to licens-

ing problems and installation problems. I was not able to find any "Frequently Used Resolution" or "Frequently Asked Question" that related to problems that I see researchers encountering with data management or statistical analysis. Under the "Statistical Support" section there are a number of well written research related articles; however, they date back to the late 1990's and often refer to the MANOVA command, a command that is no longer actively supported by SPSS.

Documentation The SPSS software comes with a separate CD with the documentation stored as .pdf files at no extra cost. This is a very nice cost savings. But only one of the "books" actually focuses on using SPSS via syntax mode, the SPSS Syntax Reference Guide. The other books all focus on using SPSS via point-and-click, with screen shots showing what to point-and-click to get results. Although these point-and-click manuals are nicely constructed and written, I do not find them useful for learning how to use SPSS for writing syntax to perform my data analyses, and I think other syntax oriented researchers would feel the same. This leaves the Syntax Reference Guide as the only official manual that stresses the use of SPSS syntax for analyzing data[4]. At about 1,500 pages the SPSS Syntax Reference has far less information than the Stata manuals (at about 4,500 pages) and the SAS manuals (SAS/Stat alone is over 5,000 pages). I think there is a real cost to this brevity of pages. The Syntax Reference Guide contains no example analyses, unlike Stata and SAS which include numerous examples with output and substantive explanation, go beyond just the nuts and bolts. The example commands are often fragmented and leave me wondering how the fragments relate to other parts of the command and how to integrate them into a whole command. The syntax examples are sometimes so sparse that despite using SPSS for 20 years I sometimes find myself giving up after reading the manual and just experimenting until I can get the program to do what I want. My assessment is that the SPSS Syntax Reference Guide is much weaker than the syntax oriented documentation provided by Stata and SAS.

Online help The SPSS online help focuses on providing information for point-and-click usage. They also offer a tutorial, statistics coach, and case studies. However, I find these to be targeted towards beginners and often towards people who are not even familiar with statistics. While these might be beneficial for those using SPSS who do not know statistics, I do not think our user community would use these very much.

Additional Books Unlike Stata and SAS, SPSS does not appear to have a publishing arm that fosters the publication of books about SPSS. While there are a number of books on SPSS, I am only aware of two or three such books that focus on the use of SPSS syntax. All other books focus on the point-and-click interface and, to me, seem to be targeted at an undergraduate or first year graduate student audience. I am not aware of any books that feature SPSS syntax and focus on a particular statistical methodology (e.g. regression, logistic regression, or survival analysis) like those available for Stata and SAS. In short, I find these resources to be almost completely absent for earnest researchers.

Listserver For reasons I do not understand, there are both an SPSS newsgroup and an SPSS listserver. You can find information and archives for both of these by visiting `http://www.ats.ucla.edu/stat/otherresources.htm`. Although the SPSS web site has a strong orientation towards business issues, both of these groups seem to be largely composed of questions that are highly relevant to academic researchers. Many questions have multiple answers and the quality of the information on this newsgroup seems timely and excellent.

Add-On Programs There are some SPSS macros that users have created, but the size of this library of contributions is very small compared to the contributions available for Stata and SAS. The SPSS Macro library within the SPSS support site contains very few SPSS macros (9 the last time looked), and in order to find them you have to login to the SPSS support site and perform a number of mouse clicks to locate them. I think the vast majority of such contributions are available at the site provided by an SPSS enthusiast at `http://www.spsstools.net/`. SPSS does not have any mechanism to help you search for and install such macros, nor any standard mechanisms for providing help or documentation for the macros. SPSS 14 added the SPSS Programmability Extension to provide the ability to extend SPSS command syntax with external programming languages. Unfortunately, there is no built-in functionality within SPSS to facilitate the distribution of such tools or finding tools others have created, but instead SPSS provides a forum via the web for users to exchange ideas and tools (see

---

[4]Although not an official SPSS manual, users can download a very nice user written book that emphasizes using SPSS Syntax, *SPSS Programming and Data Management: A Guide for SPSS and SAS Users* by Raynald Levesque, by visiting `http://www.spss.com/spss/data_management_book.htm`

`http://forums.spss.com/code_center/`). The last time I visited this site (around the end of March, 2006), I saw little activity there. I saw about 20 total messages from the last 6 months, and about 6 messages in the last 3 months. While this new functionality adds lots of promise to share programs, so far it does not seem to be gathering much steam. Time will tell the impact that it will have for adding functionality that researchers find useful.

Journals I am not aware of any journals published by SPSS on topics of statistics and data analysis.

Training SPSS offers a number of live training courses: see `http://www.spss.com/training/descriptions.cfm` . The courses appear to typically cost $599 to $1199. Most of the descriptions of the courses seem to indicate an orientation towards business applications, with a handful targeting statistical features. SPSS also offers web seminars, but only the intermediate topics class appeared to be relevant for the sorts of researchers I have met with (at a cost of $599 for a three-day web course). I have no experience with any of these courses so I cannot comment on the quality of the teaching or the quality of the content.

8 Features

This section compares the packages on a selected set of statistical features. This section does **not** consider all statistical features and is not intended as an exhaustive comparison of the packages. The features noted here were selected for two reasons. First, I have chosen features used frequently by our clients; hence, features you might find important may be omitted because I have not seen them used by clients. Second, I only included features where I felt there was a very substantial difference among the packages. To me, a very substantial difference means that as I have worked with clients, I could easily see how they were able to advance their research much more effectively with one package as compared to another. This could take a number of forms, for example: 1) one package offers a feature not available in other packages, 2) one package is **much** easier to use than another, 3) one package provides results that are **much** more interpretable than another, 4) one package provides a superior support infrastructure as compared to others, and so forth. As you can see, this is not solely based on the capabilities but tries to encompass how well I have seen clients use the various packages. The topics are presented in alphabetical order.

ANOVA

About fifteen years ago, SPSS had the finest and most full featured set of ANOVA tools I had ever seen with its MANOVA command. Despite the name, it would do ANOVA, ANCOVA, repeated measures ANOVA, MANOVA, MANCOVA, Canonical Correlation, and a number of other tasks as well. But, most importantly, it would allow you to decompose interactions more powerfully and intuitively than any other tool, enabling you to perform contrasts, simple contrasts, partial interactions, interaction contrasts, simple interaction contrasts, and more. Somewhere around SPSS version 7, the GLM command was introduced to be the successor of the MANOVA command. Seven versions later with SPSS 14, I have not seen much growth in the GLM command, and it still has yet to enable users to *easily* perform the full suite of contrasts that the MANOVA command from 15 years ago could perform. History would seem to suggest that the GLM command will never reach the power of the MANOVA command for its ability to tease apart interactions, an essential tool for researchers who wish to test specific theoretical predictions. Nevertheless, the SPSS MANOVA and GLM commands are still among the best for performing analysis of variance, and the GLM command provides the easiest way to provide graphs of main effects and interactions.

The SAS PROC GLM command has a very rich suite of tools for analysis of variance and a powerful suite of tools for decomposing interactions, and SAS has introduced tools that ease the process of graphing main effects and interactions. SAS continues to grow this suite of tools in the new PROC GLIMMIX that (despite its name) can be used for analysis of variance. If SAS continues to add features to GLM and GLIMMIX in this arena, they may overtake SPSS in power and features in this arena, but SPSS still remains first in this regard.

Stata's support for analysis of variance is not very strong at all when compared to SPSS and SAS. The most powerful tools for decomposing interactions come via user-written programs named `xi3` and `postgr3` (Full disclosure, written by me). However, these programs are more difficult to use and more limited than the tools provided by SAS and SPSS for analysis of variance.

Bootstrap and Jacknife methods

Stata makes it exceptionally easy to execute a command multiple times to compute bootstrapped or jacknife standard errors. Most estimation commands will permit you to specify the `vce()` option to request bootstrap or jacknife methods when computing the variance covariance matrix of the estimators (i.e., bootstrap or jacknife standard errors). For most other commands you can use the `bootstrap` or `jacknife` prefix command. For example, the following command runs the regression of `read` on `write` 100 times, resampling the data with replacement each time and displays the results with bootstrapped standard errors.

```
bootstrap, reps(100) : regress read write
```

While this can be done in SPSS or SAS via macros/programs available from the vendor's web sites, it is a much more complicated and laborious process.

Cross-Sectional Time Series

I must confess that although many clients come to our consulting for questions related to cross-sectional time series, this is not an area that I have emphasized so my expertise in this respect is more limited. While Stata and SAS have tools for such analyses for a variety of models, SPSS is limited to linear models via the `MIXED` command. With respect to the strengths of Stata and SAS, I would recommend a very nice page by Robert Yaffee, a statistician within the Social Sciences, Statistics and Mapping Group at NYU at `http://www.nyu.edu/its/pubs/connect/fall03/yaffee_primer.html`. In the section comparing the packages, the page states, "Among those statistical packages that excel in programs for panel data analysis are LIMDEP, STATA, and SAS. Although all three packages have procedures dedicated to panel data analysis, LIMDEP and STATA appear to have a particularly rich variety of panel analytic procedures." I would defer to this page for further details.

Data, Ease of Distribution

There are three factors that I think when considering the ease of distributing data files: 1) How easy is it to distribute the files over the web? 2) How well do the files work across platforms? 3) How well can you read older versions of the files? Ideally, a package makes it easy to all three of these.

Stata offers the ability to easily read data files right over the web, including raw data files and Stata data files. For example, you can type `use http://www.ats.ucla.edu/stat/stata/notes/hsb2` and that will read in the `hsb2.dta` file from our web site. This is an extremely convenient way to share files over the web, either to share your research data or to share data for teaching. By contrast, while SAS does offer the ability to read data files over the web, the process is difficult and I have often had problems getting it to work correctly. Within SAS, I think it is just easier to download the file to your computer and then access it from your computer. Within SPSS, I am not aware of a mechanism for reading data files right over the internet, so you would need to download and save the file onto your computer and then open it.

With Stata, all files are completely platform independent. With SAS, Version 8 and above files work across Windows and UNIX/Linux platforms, but not so with versions 6.12 and below. Much older versions of SPSS had problems with reading SPSS system files on different platforms, but I believe that modern versions of SPSS can read files from other platforms without a problem.

Stata can read all file formats going back to version 1.0, making Stata excellent in this regard. SAS can have trouble reading older files. For example, I had a SAS format file that was created in SAS 6.04 and tried every means possible to read it, even in SAS 6.12. Eventually SAS technical support told me that they had to install SAS 6.04 on a computer solely for the purpose of being able to read this file. I give very high marks to SAS technical support going the distance to provide help, but poor marks for the lack of ability of SAS to read an older file with a modern version of SAS. I have read SPSS files going back over 10 years with SPSS without a hitch, making SPSS excellent in this regard.

Data, Ease of Exporting to Specialized Statistical Packages

Although SPSS and Stat/Transfer support a large number of foreign file formats, they do not support the file formats used by some special use statistical packages I see people using frequently, such as HLM, Mplus,

and MLwiN. I often find users hit a wall with their general purpose package and need to move into a special purpose package such as HLM, Mplus and MLwiN. Moving data into these packages can be very laborious and none of the packages have simple to use features for reading foreign file formats. However, because of the programmability of Stata, user-written programs have been created to simplify the process of exporting data from Stata to HLM, Mplus, and MLwiN.

Data, Reading foreign file formats

SPSS is by far the best package at reading foreign file formats. You can use the File then Open pulldown menu and select among many kinds of file formats that you wish to read (including SAS and Stata) and point to the file and open it up. I have found this process to be extremely simple and works well. By contrast, reading foreign file formats in SAS is very difficult. Likewise, Stata can read a limited number of file formats including .csv and .xpt (SAS XPORT) files, as well as Access and Excel files via ODBC[5].

Extensibility

By extensibility I mean the extent to which the package supports your ability to add your own features to the package, share these features with others, and obtain programs/features created by others[6]. In this respect, Stata is much stronger than other packages. Bill Gould, the CEO of StataCorp, gave a Report to Users in 1997 at the 3rd UK Stata User Group meeting. You can read the notes from that talk at `http://www.stata.com/meeting/3uk/report.html`. In this talk Bill Gould discussed the architecture of Stata and described Stata not so much as a Statistical Package, but as a Statistical Operating System. The design of Stata is such that anyone can create statistical commands within Stata just as the developers of Stata do. This design means that when Stata adds functionality to make it easier for the Stata programmers to program in Stata, this translates into features that make it easier for users to program in Stata as well. The Stata programming language eases the process of developing your own commands that extend the capabilities of Stata. And, with the addition of Mata, Stata now has a compiled statistical programming language that interfaces with the rest of Stata. This is facilitated by the fact that users have access to the source code for most Stata commands and can read the code to learn how a command behaves. In fact, users can modify the Stata code to add their own features. Stata combines this with an architecture that makes it easy to share the programs that you write with others and easy for you to find the programs that others write. For example, Stata interfaces with the Statistical Software Components (SSC) via the `ssc` command to provide a common download site for user-written software on the web. By comparison, SAS does offer a macro language for automating tasks and PROC IML for writing your own statistical algorithms. However, these two tools are not seamlessly integrated, users do not have access to the source code of SAS procedures, and there is no support for the distribution or access to tools written by others. SPSS 14 added the SPSS Programmability Extension to provide the ability to extend SPSS command syntax with external programming languages. Such programs can control the flow of SPSS syntax programs and can access variable attributes, output of SPSS procedures, and error codes from SPSS procedures. While this is extends the power of SPSS, the interface between the external languages and SPSS is not nearly as seamless as Stata. While SPSS has created a forum for users to exchange programs at `http://forums.spss.com/code_center/`, this forum has had very little traffic so far. Further, once you identify a program of interest, you need to manually download the program and place it into the correct location for it to work. Once you have done this, the syntax to call these external programs is completely different from the syntax to call regular SPSS procedures.

Graphics

Full Disclosure! For the sake of full disclosure, I want you to know that I am the author of a book on using graphics in Stata. I still believe that my assessment here is on target, but you may wish to factor this in as

---

[5]To compensate for this, Stata sells Stat/Transfer, an excellent program for converting data from one file format to another, however this is at an extra cost.

[6]It is important to be mindful that user-written programs may not have been tested as thoroughly as programs that are part of the official software package. While I believe such programs are generally created with great care and diligence, I would not automatically assume such programs produce correct/proper results and would seek ways of validating such programs before putting my full faith in their results.

you read this section.

I believe that Stata offers the best graphics features of all of the packages. It can create an amazing array of graphs. Although the graphics language is extremely powerful and extensive (the manual is in excess of 500 pages), my experience with clients is that they do get the hang of the logic of the graphics commands and are able to develop very rich and complex graphs. If you have forgotten some of the options, you can use the point-and-click interface to generate the command for the graph. Because the syntax is completely clean, you can usually relate your clicks to the contents of the command so you can tinker with the command to get exactly what you want. These graphs are most certainly publication quality and can be easily exported into a number of formats, including Encapsulated Postscript files. While SAS has a very powerful suite of graphics tools that give you immense control and power over the graphs, I have found them very hard to use and I have found that users have the same kind of experience. While some of the SPSS graph commands might be easy to use, I frequently see users exceed the limits of these commands and need to use something more powerful. It feels like SPSS emphasizes graph features that would be impressive to those creating business or marketing graphs, with lesser strengths for graphs that applied statisticians wish to create. SPSS does offer an interactive chart editing tool, but some of the charting features that could be implemented in Stata or SAS via syntax are limited to the interactive point-and-click chart editing environment in SPSS. SPSS Version 14 has added the Graphics Production Language, which offers considerable graphing power, but I have found it to be very complicated to learn. SAS and SPSS both offer the ability to produce three dimensional graphics, a feature not offered in Stata. In fact, SAS permits you to create rotating graphics in the form of a .gif file that you can post on the web site to produce dynamic three dimensional spin plots.

Linear Mixed Models

All three packages offer commands for analyzing linear mixed models, however there are some very substantial differences among the packages. I believe that SAS introduced `PROC MIXED` in the early 1990's, perhaps in version 6.09. This well seasoned procedure has many options and features that permit you to estimate an amazing array of models. Among the features I see used frequently are the many level 1 covariance structures you can choose from and the ability to easily estimate different covariance parameters for different groups. Also, SAS offers a unique suite of influence diagnostics to help you assess the extent to which you might be violating assumptions of the multilevel model. To help users get the most out of this procedure, SAS publishing offers *SAS System for Mixed Models* (see `http://support.sas.com/publishing/bbu/companion_site/55235.html`), a book that combines a level of technical difficulty and applied examples that will help power users understand how to specify many kinds of models and interpret their results. SPSS introduced the `MIXED` procedure in version 11.0, but only version 11.5 permitted you to estimate random intercept/random slope models. While SPSS `MIXED` will handle many common types of mixed models, it lacks many of the rich features found in SAS `PROC MIXED`. SPSS has a more limited number of level 1 covariance structures and does not provide options to easily model different estimates of covariance parameters for different groups. Also, SPSS does not have any supplemental documentation as SAS `PROC MIXED` does. Stata introduced mixed models in version 9.0 with its `xtmixed` command. This command will only analyze the most basic of mixed models. It lacks support for level 1 covariance structures (unlike SPSS and SAS) and has very few additional features, and does not have any supplemental documentation. While basic models can be implemented in Stata and somewhat more sophisticated models can be implemented in SPSS, I think those searching for the greatest level of power, flexibility, and documentation will find SAS `PROC MIXED` most suited to their needs.

Logistic Regression

In my experience, people from many different departments struggle with how to interpret the results from logit models. Logit coefficients can be very difficult to interpret, and even odds ratios are very challenging to interpret. However, almost everyone can understand results presented using predicted probabilities as a means of expressing relationship between the predictor and the probability of the outcome being a 1. The book *Regression Models for Categorical Dependent Variables Using Stata, 2nd Edition* provides excellent information on how to use predicted probabilities for interpreting such models using tools created by the authors that simplify the process. I have found that most clients have a far easier time understanding their results and a far easier time conveying their results to others using this framework. Stata is unique in its

offerings with respect to these tools, and I believe this makes Stata an excellent tool for analyzing logit models and a superior choice to other packages for analyzing such models.

Neither SAS nor SPSS offer this kind of comprehensive suite of interpretive tools to help users understand and interpret their results from logit models. SAS still unconventionally defines a "0" as the event of interest for the outcome, unless overridden via the `descending` option. I have more than once worked with clients who had perplexing results in SAS because they were predicting a 0 on the outcome variable.

Missing Data

SAS and Stata are far better than SPSS for analyzing data sets with missing data. SAS includes (at no extra cost) tools for performing multiple imputation. Stata does not include tools for performing multiple imputation, but tools for multiple imputation have been created by users and are available at no extra cost. UCLA researchers must pay extra for the tools that SPSS provides, yet the SPSS missing values tool only performs single imputation via regression or EM. SPSS does not have built-in support to create or analyze multiply imputed data files.

Monte Carlo methods

Similar to the bootstrap and jacknife commands mentioned above, Stata has the built-in permute and simulate commands to simplify the process of performing permutation tests and Monte Carlo simulations. Although simulations can be a bit tricky, these tools make it substantially easier to perform such simulations as compared to SAS or SPSS.

Non-Linear Mixed models

SAS is the only major package that has built-in features for analyzing non-linear mixed models. SAS offers the NLMIXED procedure which uses numerical integration to estimate these models using maximum likelihood. This procedure is very challenging to use and the syntax can be very daunting, but SAS is the only general use statistical package that will allow you to analyze such models. SAS also offers the GLIMMIX procedure which uses Penalized Quasi Likelihood (PQL). However, under certain situations I believe that PQL may provide misleading results (namely multilevel models for longitudinal data) and I would use it with care. While Stata does not offer such features, the add on program `gllamm` will allow you to estimate such models. But because `gllamm` is such a flexible and general program, it uses algorithms that can take a very long time to complete an analysis. Such analyses can take tens of minutes or even hours.

Power Analysis

Both SAS and SPSS offer tools which assist with power analysis. SPSS has an add on module called SPSS Sample Power which is included as part of the UCLA license agreement (i.e., UCLA researchers get this module as part of their SPSS license). SAS 9.1 introduced power analysis tools via `PROC POWER` and `PROC GLMPOWER`. Both of these tools can calculate power for a variety of tests, including t-tests, ANOVA, ANCOVA, correlation, multiple regression (see `http://support.sas.com/rnd/app/da/new/PSS.html` for more information about the SAS power analysis tools and `http://www.spss.com/samplepower/` for more information about SPSS Sample Power). The SPSS module also includes tools for power analysis for logistic regression and survival analysis. The SAS tools are command-based while the SPSS tools are accessed via an intuitive and very interactive point and click interface, permitting you to vary one parameter and see the immediate impact it has on the required sample size.[7] Unfortunately, both tools often make simplifying assumptions that mean that real world designs often exceed their capabilities. With some creative thinking a complex design can be related to a simpler design that these tools can handle so you can still use them to get some useful information. Many designs will so exceed the capabilities of these programs that custom programming may be needed to produce an adequate power analysis. Stata offers very little in the way of power analysis tools, being limited to one- and two- sample comparisons. But if you needed to program a

---

[7]SAS also offers an interactive tool via a web interface, but I think the installation of this can be a bit tricky.

power analysis for a more complex design, Stata's strengths in the area of Monte Carlo methods make it a very good tool for creating a customized power analysis.

Robust Regression

Stata offers robust regression via the `rreg` command. SAS offers robust regression via `PROC ROBUSTREG`. SAS offers more options for estimating robust regressions, however there is little information on how to select among the options offered (perhaps because such information does not yet exist). SPSS offers no such ability to perform robust regression.

Robust Standard Errors

Robust standard errors are useful when heteroscedasticity is present in the residuals. When residuals are heteroscedastic, the statistical tests may lead to inappropriate Type I error rates. Using robust standard errors can produce more appropriate Type I error rates. Many (or perhaps most) Stata estimation commands include a `robust` option to use robust standard errors. Some SAS procedures offer the ability to select robust standard errors, but this often means using a different statistical procedure (e.g., `PROC GENMOD` or `PROC GLIMMIX`). I am not aware of any built-in procedures in SPSS that produce robust standard errors[8].

Special purpose regression models

Up until about five years ago, most questions I handled in consulting in the regression arena concerned OLS linear regression and logistic regression. Since then, there has been an explosion in interest in regression models that go beyond OLS and logistic regression. These special purpose models include models such as censored normal regression, constrained linear regression, conditional logit regression, complementary log-log regression, Heckman selection models (logit and probit), instrumental (two stage) least squares regression, multinomial logit, multinomial probit, negative binomial regression, nested logit regression, ordinal logistic regression, ordinal probit regression, truncated regression, Poisson regression, quantile regression, robust regression, seemingly unrelated regression, Tobit regression, cross sectional time series regression, zero inflated negative binomial and Poisson models, and zero truncated negative binomial and Poisson models. Stata offers very strong support for these kinds of specialized regression techniques (for example, you can see `http://www.stata.com/help.cgi?estimation commands`, which has a quick reference for the Stata estimation commands, including a long list of its special purpose regression models). Stata provides special purpose commands which generally have a syntax that is simple to follow and specialized output tailored to that technique. SAS offers support for many (if not most) of these models, however SAS does this via very general statistical procedures. It can be difficult to identify the SAS procedure that performs the specialized regression that you wish to perform, and sometimes because of the general nature of the syntax it can be difficult to determine exactly what combination of options you need to include to obtain the model you desire. While SPSS has built-in support for some of these kinds of models, I have not been able to find built-in features within SPSS for censored normal regression, Heckman selection models, multinomial probit models, negative binomial, nested logit, truncated regression, quantile regression, robust regression, seemingly unrelated regression, Tobit regression, truncated regression, cross-sectional time series, zero inflated negative binomial and Poisson models, and zero truncated negative binomial and Poisson models. I am sure that some of these can be done by SPSS, either via specifying some options on a general purpose command, in which case finding how to do this is difficult. Or, there may be macros or combinations of built-in commands that can be used to accomplish at least some of these analyses[9]. My experience is that users are much more successful in performing these kinds of specialized regression analyses when they use the special purpose Stata commands as compared to general purpose commands where they need to specify a combination of options or when they need to download a macro or follow a series of statements to coerce the package to perform an analysis it was not intrinsically programmed to perform.

---

[8] Although a user written routine has been created to do this for OLS regression at `http://www.spsstools.net/`.

[9] For example, the SPSS Code Center offers a python module for Poisson regression.

Standard Errors with Clustering

Here I am referring to situations where observations are independent between groups but may not be independent within groups. Using traditional analyses that assume full independence among all observations can yield biased standard errors and inappropriate Type I error rates. As a result, when clustering is present, it is very important to derive standard errors that account for this. Within Stata, many (if not most) of the estimation commands permit a `cluster()` option to adjust the standard errors for clustering. For example, instead of typing `regress y x`, you merely type `regress y x, cluster(classroom)` and the standard errors are adjusted for the fact that students within classrooms may not be independent.

SAS supports analyses that take clustering into account, but this feature is not as convenient or as extensive as the offerings by Stata. It is less convenient because you need to switch to a completely different `PROC`. For example you might run a regression in `PROC REG` but then in order to obtain standard errors that take clustering into account you would need to switch to using `PROC GENMOD`. Also, I believe there are many Stata commands that permit a `cluster()` option that cannot be performed with clustering in SAS; for example, I know that Stata can handle clustering with Cox proportional hazard models, Heckman selection models and two stage least squares models by just adding a `cluster()` option to the end of the command, but I do not know how to do this in SAS.

SPSS offers SPSS complex samples that can handle clustering but only for linear models and logistic models. UCLA researchers must pay extra to obtain this module.

Survey Data Analysis

Stata is far more comprehensive in its support for analyzing survey data than SAS or SPSS. Stata handles a wider variety of estimation commands and a wider variety of survey design types and/or survey aspects than SAS or SPSS. Stata provides the ability to perform survey data analysis for over 18 estimation commands (including regression, logit regression, ordinal logit regression, multinomial logit, Heckman selection models, negative binomial regression, and more). Stata supports surveys with psu/cluster/strata information as well as surveys with replicate weights. In short, the features offered by Stata for analyzing survey data rivals and often exceeds the features provided by packages that are solely devoted to the analysis of survey data. By contrast SAS and SPSS offer extremely limited features for survey analysis. Neither can handle replicate weights, and both have a limited number of estimation commands supporting linear models and logistic regression. Those using SAS could purchase SUDAAN to add extra functionality for the analysis of survey data, but this would be at extra cost.[10]. Also, note that UCLA researchers must pay extra for this functionality in SPSS.

Survival Analysis

Stata offers a very rich set of tools for survival analysis, including Cox proportional hazards (semi-parametric), parametric, and non-parametric methods. The web page `http://www.stata.com/capabilities/survival.html` provides more details about Stata's capabilities for survival analysis. Note that Stata supports Cox models with shared frailty and models with robust variance estimates. Stata also provides a rich set of tools for preparing your data for analysis. Stata also offers the ability to analyze data with clustering (e.g., students within classrooms) where there may be an intragroup correlation. SAS also supports Cox proportional hazards (semi-parametric), parametric, and non-parametric methods, see `http://support.sas.com/rnd/app/da/stat/statspec.html#SURV` for more details. Both Stata and SAS offer well written specialized books showing how to perform survival analyses using their packages, including examples showing how to set up the models and interpret the results. While SPSS can perform survival analysis, it lacks a number of features offered by Stata and SAS. SPSS does not have any specialized data management tools for survival analysis. SPSS does not support parametric survival models. SPSS has only one method for handling tied failures with survival models (Stata and SAS offer four methods), SPSS does not support frailty models and it does not support left censoring or interval censoring.

---

[10]SUDAAN will be discussed in more detail in a forthcoming technical report on special purpose statistical packages

Weights

Stata supports four kinds of weights: frequency weights, sampling weights, analytic weights, and importance weights. Many commands support all four of these kinds of weights, while others may support a subset of these weights. The user explicitly specifies which kind of weight to use for the analysis so there is never any ambiguity about how the statistical procedure is treating the weight. SAS offers the `WEIGHT` statement which is available in most procedures. Because there is only one way to specify a weight, it becomes confusing whether the procedure is expecting a frequency weight or a sampling weight, or some other kind of weight. For any given procedure, weights might be handled differently and you would need to consult the documentation to see how the weights are handled. Aside from the Complex Samples module, SPSS appears to only offer one kind of weighting: frequency weights.

9 Data Management

Another important aspect of a statistical package is the features it offers for data management. Often times data management can consume a substantial portion of time as you prepare your data for analysis. Having a tool that helps you effectively manage your data can be as important as having a tool that effectively analyzes your data. Assessing the data management capabilities of these tools requires being able to define the different kinds of data management tasks that researchers frequently encounter and then comparing the strengths and weaknesses of the packages with regards to these tasks. Before this section can be written, I will need to take some time to create some kind of taxonomy of data management tasks before being able to compare the packages in terms of data management. So, for the time being, this section will be considered *under construction*.

I will, however, say that I believe that the packages can differ substantially in their strengths with respect to data management tasks. Even if you have made a selection of a package that is most suitable for your data analysis, I would encourage you to separately assess the strengths and weaknesses of the packages for your data management and if a different package is most useful for your data management, then use one package for your data management and use a different package for your data analysis. Converting data from one package to another is so easy these days, it makes good sense to select the best tool for data management and the best tool for your data analysis, even if the tools are different. The cost of converting your data from one format to another is so small compared to the benefit that you can get from selecting the best tool at each stage of your analysis.

10 Specific Differences

There are some very specific aspects of the packages where one package is outstanding in the positive sense (i.e., providing an exceptionally useful feature not present in other packages) or in the negative sense (i.e. , you stub your toe on the package because something is much harder to use than the other packages). Here are some examples of these kinds of features that I have encountered while helping clients in consulting.

Value Labels Stata and SPSS save value labels inside of your data file while SAS stores the value labels (called formats) in an external format file. While you can have multiple format files per folder in SAS, there remain a number of complications – 1) you have to manually tell SAS to associate a particular data file with a particular format file, and 2) A data file can become separated from the format file, and 3) you can encounter problems reading older format files while being able to read the data file from the same generation. My experience is that for every 1 Stata or SPSS user who has trouble with value labels, I have seen dozens who have had problems working with value labels (formats) in SAS. In short, value labels in SAS are much harder to use and more prone to problems than in Stata or SPSS.

Reading old Files It can be extremely difficult to read old data files and/or format files in SAS. Reading old version 6 data files can be quite tricky and reading old format files can be almost impossible. By contrast, Stata can read data files going back to version 1.0 of Stata for all platforms. I have never had a problem reading old data files in SPSS and that includes reading files that are over 13 years old.

Mainframe Raw Data Some old raw data files were written on IBM mainframes using data stored in EBCDIC format (as opposed to ASCII). Further, sometimes data files were stored in packed deci-

mal format (to store numbers in as little space as possible). Stata cannot read such files. SAS and SPSS, however, provides the ability to read such files. It can be extremely challenging to read such files, but it would be virtually impossible without SAS or SPSS.

Complex raw data files Some raw data files are stored in a very complex format, perhaps having varying numbers of variables or a varying number of records per case. Without a doubt, SAS is the most powerful tool for reading these kinds of complex data files and is the very best tool for reading very complex raw data files.

Hierarchical raw data files Suppose you have a raw data file that contains information about households and the members of the household. Such files are often structured having one record for the household with information at the household level (e.g., household income, number of members in the household) and then following that would be a record for each person in the household containing person information (e.g., age, income). Such files are not traditional rectangular files because the household records need to be read differently from the person records. Files in this type of format are called hierarchical raw data files. Stata, SAS and SPSS all can read hierarchical data files. In Stata and SAS you need to do some extra effort to teach the program how to read the hierarchical data file. SPSS, however, has a very intuitive means for reading hierarchical data files and is simply the easiest program for reading a standard hierarchical data file. SAS is a close second in this regard. It is harder to read in such files in SAS, however you have additional power while reading the files in SAS that you do not have in SPSS. Stata is the weakest program in this respect, being hard to use (probably equivalent to SAS in difficulty) but not offering the kind of additional power that you get in SAS.

Very large files SAS and SPSS store working data files on disk while Stata stores working data files in memory. SAS and SPSS are much better tools if you are going to work with extremely large data files that exceed the amount of memory on your computer. Stata can work with such files but will use Virtual Memory (i.e. disk) when it runs out of memory. This can be extremely slow. So, for example, if you have one gigabyte of memory, then it could be very slow to use Stata to work with files over about 800 megabytes or if you had two gigabytes then working with files over say 1.8 gigabytes could be very slow. Today a high end Windows machine might have one or two gigabytes of memory, and a high-end UNIX/Linux machine might have perhaps 16 gigabytes of memory. Over time memory gets cheaper and future versions of Windows will permit access to more memory.

Number of variables Intercooled Stata is limited to 2,047 variables. If you will be working with data files that have more than 2,047 variables then you will want to choose Stata/SE which can handle 32,767 variables (the same number of variables as SAS), but note that Stata/SE is more expensive than Intercooled Stata. Sometimes people think they will never exceed 2,047 variables, but when accessing archival data files such as those stored at ICPSR, such files can often exceed 2,047 variables. (When this comes up in consulting we often use SAS to read the large data file and subset it to keep just the variables and observations of interest and then transfer the subsetted data into Stata.)

Reading data over the internet I do not know how to read data files over the internet from within SPSS, and within SAS you can read files over the internet, but it is complicated. By contrast, within Stata you can simply type, for example, `use http://www.ats.ucla.edu/stat/stata/notes/hsb2` and Stata will download the `hsb2.dta` data file from the web location provided. This is an exceptionally easy way to distribute data files.

Locating add-on programs All of the statistical packages have a user base that develops programs that extend the functionality of the statistical packages. However, only Stata has built-in features that allow users to search for such programs from within Stata. For example, you can type `findit regression diagnostics` to search for, and install, tools to aid in performing regression diagnostics. Searching for and installing such tools in SAS and SPSS is much more difficult.

Example Data Stata comes with a number of example data files which you can access via the `sysuse` command. The Stata manuals use dozens (perhaps hundreds) of datasets to illustrate how to use Stata commands. Most (if not all) of the datasets are available via the `webuse` command; for example, `webuse school` brings up a data file named `school` used in the survey manual. You can also type `help dta_manuals` and you can browse and download the data files to which the manuals refer. You can also visit

http://www.stata-press.com/data/, which has links to the datasets used in all of the Stata manuals

and Stata books.

SAS. The SAS manuals also refer to and use dozens if not hundreds of data files. The program comes with hundreds of sample programs, but I cannot find any "master index" that organizes the sample programs and relates them back to the example data that comes with the manuals or other sources. If you are reading the manual online, the data files are often provided as a data step that you can copy and paste into the program editor, but then this means that the data files are often of a very limited size. The books published by SAS publishing `http://support.sas.com/publishing/` all seem to have a companion web site where you can access not only the sample data files used but also the code to replicate all of their results.

SPSS. SPSS comes with a number of example data files stored in your example data directory. I have spent over 20 minutes trying to find a central repository of SPSS example data files and could not find such a place. I performed a Google search for "spss example data files" and found a link to the SPSS web site that purported to provide access to data files, but that link was dead. Searching the SPSS web site did not help to find the new location of such, if it exists.

Subsetting Suppose you want to run a statistical analysis, but you want to run it for a subset of your data (e.g., just for people over 18 years of age). In Stata, after the estimation command you can add an `if` clause which specifies a logical expression that defines the observations to be included in the analysis. Similarly, in SAS you can add a `WHERE` statement. Regardless of the statistical procedure you are running, be it a regression, logistic regression, multinomial logistic regression, factor analysis, or t-test, these subsetting techniques work for most (if not all) statistical procedures in Stata and SAS.[11] However, in SPSS, only the `REGRESSION` and `LOGISTIC REGRESSION` commands appear to support the `/SELECT` option to specify an arbitrary logical condition for subsetting your data[12]. I am not aware of any non-statistical procedures in SPSS that support the `/SELECT` statement for subsetting data. In SPSS, the general means of subsetting your data is by creating a filter variable that is 1 if the cases are to be included. Then you need to use the `FILTER` command to filter out these cases, run your analyses, and then remove the filter condition (and drop the filter variable if you no longer want. You can also use `TEMPORARY` followed by `SELECT IF ()`, however this works only for the one command that follows it. If you wished to execute two different commands governed by the same subsetting criteria, you would need to again issue the `TEMPORARY` followed by the same `SELECT IF ()` command. In SPSS, unless you are running the `REGRESSION` or `LOGISTIC REGRESSION` commands, the subsetting process is not uniform across commands and is needlessly cumbersome. In SAS or Stata, most commands support the addition of an `if` clause or `where` statement.

Limits Sometimes it is very useful to be able to determine the limits of a statistical package. In Stata, you can type `help limits` to learn about its limits (number of variables, number of observations, and so forth). We have tried to maintain a page describing such limits for Stata, SAS, and SPSS over the years and were never able to find such a similar "one stop shopping" resource for limits on SAS or SPSS.

11 Coverage when you combine packages

When you consider the statistical and data management features of these packages, perhaps you can see why I describe these as tools like a hammer, saw, or a screwdriver. If you were a carpenter, you would not buy a saw and believe that you have a sufficient toolkit to do your work. Likewise, it is important to be able to use more than one statistical package to have a more complete statistical software toolkit at your disposal. Given the emphasis I have placed on strategic choices, the choice of which packages to learn, I believe, should also be made strategically. To help you do so, let's reflect back upon the features mentioned above where there are substantial differences among the packages and describe the omissions you would have in your toolkit by focusing your efforts on a pair of these packages.

---

[11] In fact, these subsetting techniques work for many (if not most) non-statistical commands in Stata and SAS

[12] The `DISCRIMINANT` and `FACTOR` commands also support the `/SELECT` statement, but its syntax is different only allowing you to select cases based on a single value of a single variable

Stata and SAS

Stata and SAS complement each other nicely in terms of their features. Where one of the packages might show a weakness, the other package complements it with a strength. SAS is weak in terms of survey data analysis, but Stata is very strong in this area. Stata is not good with extremely large (1+ or 2+ gigabyte) data files, but SAS is strong with such files. It is difficult to use SAS for bootstrap, jacknife or Monte Carlo methods, but Stata excels in these areas. Stata is not very good at reading hierarchical or complex raw data files, but SAS is very good at this. SAS is not as good at handling weights, while Stata is very good at handling weights. The major omissions I would see would be in the area of reading foreign file formats (which could be remedied by purchasing Stat/Transfer) and that SPSS is stronger at some kinds of tasks in ANOVA than SAS. But, overall, when you consider the coverage that you get with these two packages combined, it is extremely impressive the breadth and depth of data management and statistical features that these two packages, in concert, provide.

SAS and SPSS

SAS and SPSS do not complement each other very well at all. Both packages are weak in the analysis of survey data and neither offers the interpretive tools to help interpret logit models in terms of predicted probabilities. Neither package is strong with bootstrap/jacknife and monte carlo methods – they can do these tasks, but can be difficult to use. There are many special purpose regression models that may be either unsupported or difficult to implement. Neither is as strong as Stata with respect to weights. There are models that cannot be estimated with clustering and/or such models would be difficult to specify. Neither package has tools for exporting data to special purpose statistical packages. Combining these two tools in your toolkit leaves a number of weaknesses.

Stata and SPSS

Stata and SPSS do not complement each other very well. While SPSS adds its strengths of reading foreign file formats and ANOVA to what Stata can do, it does not offer much more in terms of the what I have seen clients need. Further, this combination would leave weaknesses with respect to non-linear mixed models, reading complex raw data files, reading old IBM style raw data files. This combination also omits any of the features for missing data that SAS has built-in that may not be available in the Stata user-written tools for multiple imputation for missing data, and any of the methods for robust regression that SAS offers that is not included with Stata.

12   Packages omitted from this discussion

This paper focused on the three major general purpose statistical packages that I see used at UCLA. This, however, omits some other statistical packages, discussed briefly here.

R

Perhaps the most notable exception to this discussion is R, a language for statistical computing and graphics. R is free to download under the terms of the GNU General Public License (see `http://www.r-project.org/`). Our web site has resources on R and I have tried, sometimes in great earnest, to learn and understand R. I have learned and used a number of statistical packages (well over 10) and a number of programming languages (over 5), and I regret to say that I have had enormous difficulties learning and using R. I know that R has a great fan base composed of skilled and excellent statisticians, and that includes many people from the UCLA statistics department. However, I feel like R is not so much of a statistical package as much as it is a statistical programming environment that has many new and cutting edge features. For me learning R has been very difficult and I have had a very hard time finding answers to many questions about using it. Since the R community tends to be composed of experts deeply enmeshed in R, I often felt that I was missing half of the pieces of the puzzle when reading information about the use of R – it often feels like there is an assumption that readers are also experts in R. I often found the documentation for R quite sparse and

many essential terms or constructs were used but not defined or cross-referenced. While there are mailing lists regarding R where people can ask questions, there is no official "technical support". Because R is free and is based on the contributions of the R community, it is extremely extensible and programmable and I have been told that it has many cutting edge features, some not available anywhere else. Although R is free, it may be more costly in terms of your time to learn, use, and obtain support for it.

My feeling is that R is much more suited to the sort of statistician who is oriented towards working very deeply with it. I think R is the kind of package that you really need to become immersed in (like a foreign language) and then need to use on a regular basis. I think that it is much more difficult to use it casually as compared to SAS, Stata or SPSS. But by devoting time and effort to it would give you access to a programming environment where you can write R programs and collaborate with others who are also using R. Those who are able to access its power, even at an applied level, would be able to access tools that may not be found in other packages, but this might come with a serious investment of time to sufficiently use R and maintain your skills with R.

Other General Packages

I have also omitted discussing S-Plus, JMP, and Statistica. I have very little experience with these packages because they are very rarely used by the clients I see. Because I have so little experience with these packages, I am not in much of a position to comment upon them.

Specialized packages

This paper overlooked a number of specialized packages, including HLM, LEM, LatentGOLD, Limdep, MLwiN, Mplus, SUDAAN, and WesVar. This is not to suggest that these packages are unimportant or not relevant. To the contrary, I think the role of these kinds of specialized packages is so important that I will dedicate a separate technical report to discussing them. To try to address these packages within this report would, I think, do them an injustice. In this report I have stressed the analogy of the statistical toolkit, and these specialized packages are an important part of that toolkit. But, for the sake of focus, this paper has focused on the part of your toolkit that is covered by general purpose statistical packages.

13   Additional Thoughts

These are some additional thoughts related to the overall structure and usability of the packages.

To Point-and-Click or Not to Point-and-Click

The first issue here is the desirability of a point-and-click interface for people who are earnest about their research. A point-and-click interface makes it difficult to replicate results. To repeat an analysis via a point-and-click interface a week later, you need to repeat the point and clicks that you did the previous week, and it is very possible that you might not repeat the points and clicks in exactly the same way. In short, while a point-and-click interface might be useful for rank beginners who have mild computer/statistics phobia, I have found it to be a downright impediment to earnest researchers who wish to create reproducible results. With this in mind, let's briefly look at each of the packages with respect to point-and-click interfaces.

SPSS emphasizes its point-and-click interface as a primary means of interacting with SPSS. Aside from the Syntax Reference Guide, the rest of the SPSS documentation emphasizes the point-and-click interface, and the user has to go to extra lengths to configure SPSS to emphasize command syntax. For example, by default SPSS does not display the syntax generated by the points and clicks along with the output, and I frequently have seen users who come to me for help with their SPSS printed output and all I see is output and I cannot see the commands that produced the output.[13]. I then need to ask them to go back and change the setting to display the syntax in the output and re-run their analyses and bring that printed output.

Further, in order to be able to edit the syntax, one has to select "Paste" instead of "OK" and then go to the syntax window and separately execute the pasted syntax. Also, the pasted syntax is unnecessarily long,

---

[13]However, if the users brought their `.spo` files I could see their syntax by clicking on the *Notes* to see the underlying syntax prior to the command

including numerous options that merely request that SPSS perform its default behavior with respect to the command. My experience with users is that when they try to transition to SPSS syntax via pasting the syntax they are daunted by the complexity of the commands shown which do very basic tasks. It conveys to them that if they want to replicate the command themselves by just typing in the syntax, that they would need to type in multiple lines of code and it only confirms to them that syntax is a difficult means to obtain what they desire. Here is an example of the syntax that SPSS pasted when I used the point-and-click interface to run a simple multinomial logistic regression predicting y from x1 x2 x3.

```
NOMREG
  y  (BASE=LAST ORDER=ASCENDING) WITH x1 x2 x3
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001) SINGULAR(0.00000001)
  /MODEL
  /STEPWISE = PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE)
  /INTERCEPT =INCLUDE
  /PRINT = PARAMETER SUMMARY LRT CPS STEP MFI .
```

I cannot tell which options displayed are necessary vs. superfluous. I think most users would come away with the impression that they will need to type all of these options just to run a simple multinomial logistic regression. Or, if they don't need to type all of this, they won't know which parts they can safely leave out. This makes the transition from point-and-click to syntax within SPSS much harder than it needs to be.

Stata offers a point-and-click interface with its software via the pulldown menus. Once you have made your selections you can click "Submit" or "OK"; both display and execute the command implied by the points and clicks, but the Submit button leaves the window available for modifying your points and clicks, while OK closes the window. The syntax is placed into the review buffer (a list of previously executed commands) and is shown in the output. This leaves no ambiguity about the command last executed, and provides a means to help the user transition from point-and-click mode to command mode. The user can select among previously executed commands and modify them and submit them as a means of transitioning to using commands. The commands that are executed have a clean syntax, including no extra baggage. For example, I used the Stata point-and-click interface to perform a multinomial logistic regression and it created this syntax.

```
mlogit y x1 x2 x3
```

If I wanted to run a multinomial logistic regression with different variables, it would be easy to simply type the command instead of using the point-and-click interface. And, in fact, I have found that the small number of folks who have used the Stata point-and-click interface transition to Stata syntax very easily.

SAS has tried a number of different point-and-click interfaces including SAS/ASSIST, SAS/INSIGHT, and most recently SAS/Enterprise Guide. We have very little experience with this product. My brief impression is that this is geared towards users who would want to use SAS as a point-and-click product and not transition to using it via syntax. I think that the transition from Enterprise Guide to syntax mode would be a difficult transition, but I do not have any direct experience with this.

Comparison of Command Structure

Stata commands have a very common and simple command structure. The commands are often very short and are typed as a single line. The Stata philosophy is "Type a little, get a little." The commands have a consistent structure (even the user-written commands) and I find that users pick up the structure quickly and can generalize this structure to new commands. You can use the [Page-Up] key to bring up a previous command to make minor changes to it. This makes it easy to tinker with commands and quickly fine tune the options.

SAS commands also have a common and regular command structure, but it is not a simple command structure. The statistical commands come in the form of procedures which can include a number of additional statements. The structure of commands is very regular, so once you know how to use, for example, the "class" statement for one procedure, it works in generally the same way in other procedures. However, the procedures can require quite a bit of typing. By contrast to Stata, SAS is much more like "type a lot, get a lot."

Unlike Stata and SAS, SPSS does not have a well defined regular command structure. While there may be syntactical elements that are shared among some of the commands, it is hard to describe a general

framework that is shared among all SPSS commands. Even some of the rules are not applied consistently. For example, I can type `descriptives write`, `descriptives var=write` or `descriptives /var=write`, and all work. While some may feel that this shows flexibility, I think this lack of a regular structure and inconsistent application of syntax rules makes it much more difficult for users to envision a schema for the structure of SPSS commands.

For the sake of comparison, let's compare the three packages in terms of obtaining a multinomial logit regression predicting `y` from `x1 x2 x3`. Although this is a single example, I do not think this example is atypical in comparing the packages in terms of the syntax used for statistical analyses. As you can see, the syntax of Stata is much shorter and easier to remember than the syntax of SAS or SPSS to obtain a simple multinomial logit regression.

Stata

```
mlogit y x1 x2 x3
```

SAS

```
proc logistic;
  model y=x1 x2 x3 / link=glogit;
run;
```

SPSS

```
NOMREG race WITH educ paeduc maeduc
  /PRINT = PARAMETER SUMMARY CPS MFI .
```

Follow-ups to Models

Often times when you run a model, you want to run some follow-up commands. Perhaps you ran a regression, and would like to generate residuals and examine the residuals. Perhaps you ran a multinomial logistic regression and would like to generate the predicted logits. Or you might have run a logistic regression and want to jointly test the significance of a group of variables. Stata, SAS and SPSS have a very different approach to these "followup" commands.

Stata uses what they call "postestimation" commands (because they are run after estimation). These commands do not require that the model be run again and there are a wide variety of postestimation commands that apply across a wide spectrum of estimation commands. With SAS, some procedures can be run in interactive mode, where even after you issue the `RUN;` statement, the procedure is still running and permits you to issue further commands, but not all commands can be run in this interactive manner (so followups would require re-running the entire model). With SPSS, there are no followups, so to generate residuals or additional tests, the entire model always needs to be re-estimated.

Ease of Teaching

I have participated in teaching the Introduction to Stata, SAS, and SPSS classes dozens of times. These hands on classes cover nearly identical material and provide an opportunity to compare the packages in terms of ease of teaching and ease of learning in the first 4 hours for the rank beginner. Our latest version of these introductory classes can be seen online at `http://www.ats.ucla.edu/stat/seminars/`.

Given that SAS is such a power user program, it comes as no surprise that SAS proves to be one of the most challenging programs to teach at first. Students require lots of help at first and when teaching a class of about 20 people we would strive for having 3 "helpers" who would assist the students as they get stuck. The helpers are constantly on their toes trying to help the students. Students often have trouble with the lengthy syntax and have difficulty recovering from errors, especially those involving omitted semi-colons. Not only are such errors common for beginners, but they are also very difficult for them to troubleshoot.

SPSS is taught via its point-and-click interface and it would seem that it would be the easiest program to teach. However, our repeated experience is that SPSS via the point-and-click interface is as difficult to teach as SAS. The reason is that with the point-and-click interface, the students have to be following along at exactly the same pace as the teacher. If they fall behind even one simple mouse click (or make a stray mouse click), they have a very difficult time "finding their place" again and getting back to be at the same

place as the teacher is. The SPSS class, like the SAS class, requires 3 "helpers" who are constantly on their toes throughout the class.

Stata, like SAS, is taught as a command-based program. One would expect that as a command-based program that it would require the same amount of help as SAS does. Quite the contrary. When we teach the Stata class, we generally bring only two helpers and repeatedly find that we can quickly dismiss one of the helpers. In fact, after about the first 30 minutes, the first helper has very little to do since most of the users are able to follow the teacher. If the students fall behind a little bit, they are able to catch up. If they make mistakes, they are generally able to fix them on their own without intervention from the helper. This is due to the fact that the Stata commands are shorter and have a simpler structure (e.g., do not have semicolons to forget, as with SAS).

Text Output

In past days, statistical packages stored their output as simple plain text files. Plain ASCII text is perhaps the most durable format that can be used for storing computer files because it can be read on almost all computers using nearly any text file editor. Storing output in such a format I think is very advantageous, for the ease with which you can share output with others and the durability of being able to read the data years or decades in the future. By default, Stata output files are stored as `.smcl` (Stata Markup Control Language) files which are kind of like .html files. The files can be read with a plain text editor, but there would be quite a bit of markup language that would be in the way. However, users can choose to store their output files as plain `.log` files that are plain text files that can be read in any editor. SAS, by default, stores its output files as plain text files and the log files are also plain text files. SAS permits you to store your output in other formats such as `.html` or `.rtf`, but I think plain text files are a very good default setting. By default, SPSS files are stored as `.spo` files (SPSS output files). Such files require SPSS or an SPSS viewer program to be able to read the files. Such files could only be read on computers which had such programs installed and available. It is possible to export SPSS output files to other formats (including text files and .html files) but this requires deliberately exporting the output. I do not know of any setting to make SPSS save output as text files by default.

Integration of Commands and Output

Stata, SAS and SPSS differ tremendously in the way that their commands and output are integrated. When I see output from a statistical package, the first question I ask is *What commands produced this output?*. How I read the output depends on the commands that produced the output. I believe it is essential to be able to easily relate the statistical commands to the output to be able to clearly interpret the output. So, let's compare these packages in this respect. In Stata, the output is organized with the Stata command followed by the output that the command produced. It is very simple to see the relationship between the command and the output it produces, even when the Stata point-and-click interface was used. In SPSS, by default the syntax is suppressed in the output. However, you can easily configure SPSS to display the command syntax prior to the output it produces. Once this is done, it is easy to see the relationship between the syntax and the output it produces. In SAS, the commands and error messages are stored in the log window and the output is stored in the output window. The commands are not integrated into the same file. While it is possible to integrate the log and output results into a single printout, doing so is not as simple as with Stata or SPSS and doing so makes SAS difficult to use interactively.

Integration of Graphs in Output

Each package differs in the way that graphs are integrated in output. In fact, there is a tension between this issue and storing output as text files (noted above). Storing output as text files makes it difficult to integrate graphics and output, while storing output in a non-text file format makes it easier to integrate graphics with output. In Stata, the graphs are stored in windows completely distinct from the rest of the statistical output. It is extremely difficult to produce an output file that integrates regular output with graphs. In SAS, the default output mode is plain text, but if you produce output in `.html` mode you can get output and graphs integrated together into a `.html` file. SPSS is the only package that by default integrates the

output and graphs together into a single unified output, and this output can be easily exported as a `.html` file that integrates the graphs and output.

Early Roots

I think it is useful to consider the early roots of these packages. Sometimes it is hard to understand why the packages behave as they do until you consider their origins. Stata originated as one of the first statistical packages specifically developed for the PC running under DOS. As such, it has an interactive style that is reminiscent of DOS programs which often involved typing short commands and getting immediate output. Data files were accessed directly by specifying (optionally) the directory name and the name of the file, and default directories could be established to indicate that when directory names are omitted, the files are presumed to reside in the current default directory.

SAS originated as a mainframe product, looking at the mainframe version and the Windows version shows many similarities in its interface. Many commands/conventions that seem very bizarre to Windows users are really holdovers from the mainframe roots of SAS. For example, the "libname" statement originated as a means of accessing external files without the need for complex Job Control Language on the mainframe computer. Windows users still often need to contend with this anachronistic conceptualization of data storage, and it is difficult to explain this conceptual model since they do not understand its mainframe origins. Likewise, the two part file naming strategy, while very advantageous for the conceptual model of the mainframe, appears very unintuitive and foreign to those who have no such mainframe experience.

Although SPSS also has mainframe origins, its point and click interface has a completely modern interactive windows feel to it. However, its command structure still has a mainframe feel to it. The forward slashes to indicate continuation cards, the need to indent certain portions of commands (or use + signs in column 1 to permit indenting). While there are many holdovers of such conventions, modern versions of SPSS do not appear to read old legacy SPSS mainframe syntax files without a fair amount of modification. I have never found a system setting to tell SPSS to be able to read 20 year old mainframe data deck setups.

14 Final Thoughts

I hope this report helps you think more about the different dimensions which can be used to compare statistical packages and helps you to compare the strengths and weaknesses of the packages along these different dimensions. By assessing these strengths and weaknesses, I hope this paper helps you determine whether you feel that your statistical toolkit is well stocked, or whether it may be missing some important and essential tools. By looking at the strengths of each of the packages and the strengths of the packages when used in combination, I hope this paper helps you make strategic decisions that will enhance the strength of your statistical toolkit and enable you to perform data management and data analysis tasks that extract the greatest amount of information from your data.