Linux: International UNIX and Freeware for Statisticians

Charles M. Fleming
Jameson Burt

National Agricultural Statistics Service, U.S. Department of Agriculture

ABSTRACT: Packages for high quality statistical data analysis, mathematical typesetting, networking, and Internet software have been copyrighted under the General Public License and hence are free. By using these packages, an organization anywhere in the world having the right technical talent can build a fully capable statistical workshop comparable to that of a major research institution.

Useful statistical software depends on the operating system of the computer to make it function reliably. Of the available operating systems, Linux is enjoying worldwide recognition for its extraordinary reliability, versatility, development tools, utilities, ease of use, and price.

Linux can be installed on Intel x86, DEC Alpha, DEC VAX, MIPS, Sparc, HP, Apple, and most recently on PalmPilot platforms. Linux supports 29 file-systems including AIX, OS/2, DOS, 16-bit, BSDI, NTFS, Novell, and CP/M. Members of such prestigious institutions as the National Aeronautical and Space Administration (NASA), National Institute of Standards and Technology (NIST), European Laboratory for Particle Physics (CERN), and numerous research universities contribute to the development of Linux.

The lineage of Linux is UNIX, so Linux's remarkable success can be attributed to 30 years of UNIX development and to provisions of the General Public License (GPL). Under the terms of the GPL, the source code of a program must be made freely available to anyone. Furthermore, the source code of a proprietary software package which incorporates some portion of a GPL software code must also be made freely available in its entirety. Therefore, Linux or any derivative of it will always remain freely available.

Experience shows that the facility to inspect and to modify source code, which the GPL guarantees, is more important than the free cost of the software. Because anyone can determine how Linux or any other GPL software executes a function, anyone can modify the source code in order to accommodate a special need. If the improvement is deemed to be a good one, it can be submitted to the Linux community for general implementation.

This collaborative activity has led to the creation of international Linux teams which work around the clock by handing their latest code across time zones to other members of their team. The Debian Linux team alone consists of 210 maintainers from Argentina, Austria, Belgium, Brazil, Canada, Columbia, Czech Republic, Finland, France, Germany, Hong Kong, Hungry, Israel, Italy, Japan, Netherlands, Spain, Sweden, Switzerland, and UK. Such collaboration from tens of thousands of Linux beta testers and contributors has created collectively a development team far larger and more talented than any commercial enterprise can support. For this very reason, which the Linux paper,*The Cathedral and the Bazaar*, *(http://www.earthspace.net/~esr/writings/cathedral-bazaar)* nicely discusses, the owners of Netscape will freely distribute the source code of their product so that it might attract the same level of vigorous development that Linux attracted.

Linux comes to the world as a product designed to be versatile and reliable. As a result, while Linux will run on 2 MB memory computers, it will also run on computers with 2 GB of memory, 16 processors, and 5 ethernet interfaces as demonstrated by researchers at Los Alamos National Laboratory. Because every Linux computer runs peer-to-peer, any single computer in a network can simultaneously serve software and accept both multiple modem and multiple ethernet logins. Linux has proven its reliability in many organizations. In particular, our Linux network of statistical and ancillary utilities has not failed in three years of continuous operation.

Hundreds of standard UNIX commands, by design, have the capacity to work together as one supertool. When placed in the hands of a competent system administrator, they constitute a powerful set of tools for managing a computer network. A Linux network can serve Microsoft and Apple software to Microsoft Windows, NT, and MacIntosh computers; run print servers as Cisco does to its 1600 printers worldwide; operate behind a natively equipped firewall; maintain the popular Apache web server which serves 45 percent of the world's web pages; provide the same ftp server used by 80 GB ftp sites found on the Internet; offer Secure Shell for encrypted peer-to-peer connections; and create a Cryptographic File System to protect any directory.

Linux maintenance tools include a facility to install and upgrade all software over telephone lines. We keep a current 1.4 GB mirror of the latest Linux software by means of these tools. For a small site, even a mirror of this size can be kept current with little difficulty by running Linux and using a simple telephone line connection through a 33.3 Kbps modem. Linux provides other savings; for example, an office need pay for but one IP address if it uses the common technique of IP masquerading. Because of these capabilities and the reputation for being robust, Linux at present accounts for one-third of all UNIX computers in the world.

In any statistical office, a statistician needs adequate software to analyze data, to make graphs, to share information with other researchers, and to publish his/her findings. In a poor country where a professional educator or statistician may earn $300 per month, setting up a statistical workshop that uses a commercial operating system and commercial software may cost more than a year's salary for only one set of licenses. It is in this context that Linux and open source statistical software might be helpful.

Linux includes all the standard programming languages, such as FORTRAN, Pascal, C, C++, and Lisp. AT&T developed the powerful statistical language called S. Carnegie-Mellon University maintains a site to which principally academic statisticians contribute statistical programs written in S. With S no longer freely available, a GPL implementation of S, called R, arose. Like Linux, R draws contributors from around the world. In some crucial aspects, R is superior to and more current than Splus, a commercial derivative of S. The powerful packages Octave and SciLab numerically solve differential equations and systems of linear equations, often with the same commands one would use in the commercial package MatLab.

SQL servers together with object-oriented and relational database management systems abound in the Linux world, including the GNU SQL server that was developed under the auspices of the Russian Academy of Science, the Russian Foundation for Basic Research, and the Free Software Foundation. Baylor University now maintains the freely available geographic information system (GIS) originally developed by the U.S. Army Corps of Engineers called GRASS.

GNUPlot and xfig provide very handy utilities for making graphs and drawings. For extensive visualization of scientific data, one can use Dataplot developed by NIST. The premier typesetting package for the mathematical and statistical communities is LaTeX. Originally developed by the American Mathematical Society, National Science Foundation, and Stanford University, LaTeX has been implemented in Linux better and more thoroughly than in any other operating system.

Creating a free statistical workshop is feasible wherever there is an industrious statistician. Linux and statistical freeware provide the means for a statistical office anywhere in the world to create a sophisticated statistical workshop for free. The executive director of the organization, “Food for the Hungry”, located in Washington, D.C., “firmly believes that Linux is an important asset for the developing world and can hasten the wiring of Africa.”

For insights and problems, Linux has numerous on-line books, quick “HOWTO” documents covering 221 subjects, web pages with their own search engines, and mail archives. Throughout the world there exist Users’ Groups of Linux, R, GRASS, TeX, and other packages. They are accessible via the Internet. Help in answering questions often comes within 4 hours.