

# Quantian: A Comprehensive Statistical Computing Environment

Christopher N. Lawrence\*  
Duke University & Debian Project

Dirk Eddelbuettel†  
Debian Project

September 19, 2005

While political methodologists and other quantitative social scientists are the most obvious beneficiaries of the widespread availability of personal computers and associated software, virtually all social scientists rely on computers for a variety of research-related tasks, including but not limited to empirical analyses; word processing and typesetting; communicating and publishing via email and the web; electronic library and publication searches; and various administrative tasks.

In this article, we present a comprehensive computing environment—**Quantian** (Eddelbuettel, 2003)—that we believe fills these needs at a fraction of the monetary cost of commercial software, while not sacrificing the capabilities and ease-of-use associated with such packages. Unlike most commercial software products, **Quantian** is “free” / “open source” software,<sup>1</sup> which leads to a number of advantages over commercial alternatives, particularly in an academic setting:

**Academic freedom.** Unlike traditional software, free software allows you to build on others’ work by extending and redistributing the results to anyone you choose, without further permission from the original author. For example, the **MASS** package for the **R** statistical language and environment (R Development Core Team, 2005) includes a function named `polr` (*proportional-odds logistic regression*) that estimates the ordered logit model, but until recently there were no functions included in **R**’s standard library collection which estimated the closely-related ordered probit model. With “closed-source” software, the researcher would have to write his or her own ordered probit routine to be allowed to share it with others. By contrast, because **MASS** is “open source,” anyone is free to modify the code of `polr` to estimate the ordered probit model as well, and give that code to anyone else, so long as they give the original authors of the code their due credit.<sup>2</sup> Some free software licenses, most notably the GNU General Public License that **R** uses, also require the distribution of the source for derived works.

---

\*Durham, North Carolina; [lawrenc@debian.org](mailto:lawrenc@debian.org); <http://www.lordsutch.com/polsci/>.

†Chicago, Illinois; [edd@debian.org](mailto:edd@debian.org); <http://dirk.eddelbuettel.com>.

<sup>1</sup>Free software advocates distinguish two aspects of freedom: price (“free as in beer”) and liberty of reuse (“free as in speech”). Free software is always “free” in the political sense; as explained below, it is often “free” in the economic sense of the term as well. This confusion in the English language has led to the advocacy by some of the term “open source” as a replacement.

<sup>2</sup>In recent months, the authors of the **MASS** package added several other link functions to `polr`, including `probit`. The optional **MCMCpack** package also includes an estimator for the ordered probit model.

**Developed by actual users.** Most free software is developed by individuals or groups who use the software in their daily work. `R`, for example, is developed primarily by a core team of well-known statisticians, further extended by researchers and methodologists from different disciplines, including social sciences, and employed in a variety of cutting-edge research projects. Two prominent examples are the `BioConductor` project for the analysis and comprehension of genomic data, and the `Virtual Data Center` software developed at Harvard for sharing data over the Internet which both use `R` as the underlying computational engine.

**Low or zero cost.** As the name implies, free software is just that—free. While distributors of free software are allowed to charge as much or as little as they like for it (often in exchange for some end-user support or making the software easy to install and use), the same software can often be found on the Internet at zero cost. The low cost of free software, however, doesn't come at the expense of quality; in fact, some free software packages, like `TeX` and `R`, are more robust and have more features than their commercial “closed source” alternatives.<sup>3</sup>

These advantages make free software an ideal platform for all social scientists—and in particular for graduate students in the field and teaching methodology in a lab setting. A department could establish a methods lab by recycling discarded computers using free software at a minimal cost, instead of licensing proprietary software.

Until recently, however, such an endeavor would be difficult for an individual department to accomplish. Gathering the required software and installing it on a lab of computers would be a time-consuming process, and would require ongoing maintenance by someone with a strong background in computing—expertise that is not available at many smaller departments and often dependent on an intake of graduate students or junior faculty.

Today, however, there are several options available. A common approach in the recent years has been to gather all of the software on a freely-distributable CD-ROM or DVD, which can be used to install it on computers that will be used as lab machines. While this approach solves the problem of collecting and distributing the software, it still imposes an installation and maintenance burden.

We instead propose the use of a relatively new technology—the “live” DVD-based operating system—to solve the installation and maintenance problem. This operating system, which is called `Quantian`, builds on several (related) open source operating systems: `clusterKnoppix`, a CD-ROM-based operating system for distributed computing (itself based on `Knoppix`, a CD-ROM/DVD-based operating system created by Klaus Knopper), `Debian GNU/Linux`, a non-commercial operating system based on the Linux kernel that is the product of a worldwide network of software developers, and, of course, all the individual applications available in Debian, Knoppix and `clusterKnoppix` that are included in `Quantian`.

`Quantian`, like `Knoppix`, is designed to operate on a PC-compatible computer<sup>4</sup> without installing anything on the computer's hard drive. This approach was common in the days of floppy disks,

---

<sup>3</sup>For example, McCullough (2003) compared the accuracy of the free spreadsheet Gnumeric (part of the GNOME desktop environment) and Microsoft's Excel spreadsheet (part of Microsoft Office), and revealed serious numerical inaccuracies in several functions included in earlier versions of both products. However, Gnumeric's problems were rectified in later versions, while many problems in Excel were either not fixed or given incomplete fixes that led to other inaccuracies in Excel 2003.

<sup>4</sup>Virtually all live CD-ROM/DVD projects target x86-compatible PCs, with a handful of exceptions for Mac and 64-bit PC platforms.

but it was generally abandoned as computer hard drives became larger, faster than floppy disks, and less expensive. The increased speed of CD-ROM (and now DVD) drives, coupled with the widespread availability of writable CDs and DVDs have made removable-storage-based operating systems like Knoppix and [Quantian](#) appealing once again. Instead of needing a lengthy and complex installation process, these operating systems run directly from a CD or DVD drive, providing a complete operating system within a couple of minutes of powering on the computer. This reduces the computing expertise needed to use [Quantian](#) to a minimum, as well as allowing potential users to test drive [Quantian](#) without going through an installation process. This approach also allows the use of [Quantian](#) on borrowed computers, such as loaned machines or in public-use computer labs.

[Quantian](#) also builds on the free software packages produced by the [Debian](#) project. [Debian](#) developers have packaged over 10,000 pieces of free software, including a wide variety of packages for statistical and numerical computing. [Quantian](#) takes many of the scientific and mathematical tools from [Debian](#)'s collection of software, as well as a full desktop environment, providing a rather complete set of free software for statistical computing.<sup>5</sup>

At present, [Quantian](#) adds about four gigabytes of scientific software to a basic Knoppix system<sup>6</sup> including a complete  $\text{T}_{\text{E}}\text{X}$  and  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  installation, based on the  $\text{t}_{\text{E}}\text{T}_{\text{E}}\text{X}$  distribution, as well as additional style files and utilities; the [R](#) statistical computing environment, along with an almost complete selection of [R](#) packages from both the Comprehensive [R](#) Archive Network (including [MCMCpack](#) and [Zelig](#), two packages developed by political methodologists) and from [BioConductor](#); [Octave](#), a numerical computing system similar to Matlab; a number of computer algebra systems; [Gnuplot](#) and several other scientific plotting engines; the [OpenDX](#) and [Mayavi](#) data visualization systems; and the [XEmacs](#) editing suite, including the popular [AucT<sub>E</sub>X](#) and [ESS](#) (Emacs Speaks Statistics) packages for editing  $\text{T}_{\text{E}}\text{X}$  and interacting with statistical software, respectively.<sup>7</sup> Many of these software packages will be familiar to readers of *The Political Methodologist*.

For programmers, [Quantian](#) also includes a complete software development environment based on the GNU Compiler Collection (GCC) for programming in C, C++, Fortran, and Java; additionally, the Perl, Python (including the Numeric and Scientific Python modules), Ruby, Lua and Tcl scripting languages are available for less computationally-intensive tasks.<sup>8</sup>

While the [Quantian](#) environment is based on Linux, within [Quantian](#) users can access files on USB-memory devices, floppy, Zip or hard drives connected to the computer, as well as networked resources on Windows and Macintosh networks and on the Internet; there is also support for printing to local and network printers. This versatility allows users to load and save data sets to persistent storage, even though the operating system is only transitory.<sup>9</sup>

---

<sup>5</sup>An overview of [Debian](#)'s thousands of packages is [available](#); of particular interest will be packages in the [math](#) and [science](#) categories.

<sup>6</sup>This is measured relative to the CD-ROM-based Knoppix 3.\* versions which measure around 2.2 GB uncompressed versus about 6.6 GB for [Quantian](#). However, the recently released Knoppix 4.0 is also DVD-based and contains several applications also found in [Quantian](#).

<sup>7</sup>Veterans of UNIX will also appreciate the inclusion of [Vim](#), a clone of the venerable *vi* editor.

<sup>8</sup>A full listing of included software is available at [the Quantian website](#).

<sup>9</sup>However, Knoppix-based systems such as [Quantian](#) also allow for installation to the hard drive, and are frequently used as a means to installing a Debian-based system. Moreover, a USB-memory device can be used for as a 'portable' home directory in which a user can store configuration, application data, or even extensions such as [R](#) packages beyond the hundreds already provided by the base system.

**Quantian**, in effect, combines the ease of use of an installed computer system like Windows or Mac OS X with the convenience of having a complete statistical environment that can be used on almost any PC-compatible computer. We envision several use cases for **Quantian** by political scientists:

**Computer labs.** Undergraduate and graduate quantitative methods courses often involve lab sessions. The preparation of computer labs is a time-consuming task, made even more difficult if the computers involved are under the jurisdiction of another department or an academic computing office. **Quantian** solves this problem by allowing the temporary use of a statistical computing environment; all the instructor needs to do is produce one CD (or DVD) for each computer, and when the class begins, each computer can be rebooted into the **Quantian** environment. When the class is over, the students only need to remove the CDs and reboot the computers back to their “normal” operating system.<sup>10</sup>

**Homework.** Undergraduate and graduate methods courses also require students to complete homework assignments. With **Quantian**, students can use the exact environment they use in class for out-of-class assignments on their own personal computers, with no licensing restrictions or lengthy installation process; all the students need is a CD-ROM/DVD containing the **Quantian** system.

**Ad-hoc computing clusters.** **Quantian** includes the openMosix system for parallel computing. This will allow the use of **Quantian** for ad-hoc computing clusters, which can dramatically increase the speed of computationally-intensive estimation procedures such as numerical integration, Markov chain Monte Carlo (MCMC), multiple imputation, and bootstrapping. With **Quantian**, a university computing lab could be put to productive use over a weekend to solve computing problems, without disrupting its use during the week for instruction or other purposes and without permanently installing any software in the lab.

**Bare or “hand-me-down” machines.** Old computers, before they are removed from service or salvaged, often have their entire hard drive wiped clean. Legally, inheritors of these PCs must purchase a new copy of a commercial operating system like Windows to have the right to use it on these machines. **Quantian**, as a truly free operating system, has no such legal entanglements, and thus is an ideal replacement for Windows on salvaged hardware.<sup>11</sup>

**Travel.** Faculty and graduate students often travel to conferences or other campuses to conduct and present research. **Quantian**, as a complete statistical system on one CD-ROM or DVD, can be taken anywhere and used on virtually any PC-compatible computer with a CD-ROM or DVD drive made in the past decade—in hotel business centers, on loaned laptops, and in computer labs at other institutions—without having to travel with a computer.

**Saving space.** One of the apparent laws of computing is that eventually data will expand to fill all available disk space. Having your operating system on a CD or DVD will free space on your computer for email, data sets, papers, letters and other personal information that needs to be stored.

---

<sup>10</sup>Many newer computers have support for network booting using the Intel PXE protocol; this will allow all of the machines in a lab to be booted from a single master computer. For details on how to use this procedure, see [http://dirk.eddelbuettel.com/quantian/howto\\_netboot.html](http://dirk.eddelbuettel.com/quantian/howto_netboot.html).

<sup>11</sup>This of course true for other free Linux distributions, but not necessarily for many commercial Linux distributions, such as Red Hat Enterprise Linux and Novell’s SUSE Linux, which have per-seat licenses.

**Virus protection.** Using an operating system like *Quantian*, where the operating system itself is on a non-writable storage medium, will reduce the ability of a virus to infect important system files and leave your system in an unusable state.<sup>12</sup>

Additional information on *Quantian*, including new releases, is available at the *Quantian* website, <http://dirk.eddelbuettel.com/quantian.html>. You can obtain a copy of *Quantian* on CD or DVD for a small fee from a number of distributors, or produce your own copy for free using a personal computer with a fast Internet connection and a CD or DVD writer. The website also has instructions on how customized variants of *Quantian* can be created.

We look forward to feedback from users of *Quantian*, including suggestions for additional software to be included and reports on its usage by readers of *The Political Methodologist*. We also plan to provide reports on some of the above use cases at future academic conferences and to continue evangelizing the use of *Quantian* and free software by both political methodologists and the broader scientific community.

## References

- Dirk Eddelbuettel. *Quantian: A scientific computing environment*. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 2003. URL <http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/Eddelbuettel.pdf>. ISSN 1609-395X.
- Bruce D. McCullouch. Fixing statistical errors in spreadsheet software: The cases of gnumeric and excel. Technical report, Computational Statistics & Data Analysis Statistical Software Newsletter, 2003. URL [http://www.csdassn.org/software\\_reports/gnumeric.pdf](http://www.csdassn.org/software_reports/gnumeric.pdf).
- R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL <http://www.R-project.org>. ISBN 3-900051-07-0.

---

<sup>12</sup>On the other hand, the read-only nature of the storage medium also implies that security updates cannot be applied. The user will have to wait for (or create) a new release of the CD/DVD with the updated software.