Getting Your Feet Wet with Open-Source Statistical Languages

By Nathan Patrick Taylor

If you are new to healthcare analytics, it can be quite expensive to get up and running with a statistical application. Between the cost of the software itself and the instructional courses—not to mention your investment of time—just getting started could run several thousand dollars. Thankfully, many open source tools are available for the data analytics uninitiated. The term "open source" refers to the fact that the source code is available to be changed and redistributed, typically for free. Various license types exist so you'll need to check the license agreement to determine exactly what can be changed and if any limitations apply.

The first tool I was introduced to was the R programming language. The source software is freely available and several companies have made integrated development environments (IDEs) for R, also typically free. IDEs are simply programs used to write code in a programming language with built-in features like code debugging and version control, among others. To learn the language, there are a few websites that introduce you to R while allowing you to type code directly into the website with explanations on how the code runs and exactly what the code is doing. For introductory books, my first resource was a "free deal of the day" written specifically to welcome new statistical programmers into the world of R. Many publishers offer a weekly or monthly freebie to ease new learners into a programming language.

My second open-source tool was Python. What makes Python a great tool is that it can be used to develop applications outside the realm of statistical computing. For example, let's say that you are charged with creating a dashboard that reports out the number of admissions and discharges by day (a very simple example just for the sake of explanation). With Python, you can hook in to your database, process the data, and write a quick app to display the results to a number of formats, including a web page and PDF. R has some similar functionality that can be achieved with its Shiny package.

The last tool I'll mention here is the KNIME Analytics Platform. KNIME has an intuitive graphical user interface (GUI) that allows you to build data workflows quickly without the need to write code. Once you develop some statistical coding skills, you can dive deeper into KNIME's "nodes" and write your own custom code. In fact, KNIME includes nodes that can run your own R and Python code. Additionally, KNIME includes nodes that can replicate (and even replace) the functionality of extract, transform, and load (ETL) tools. Learning KNIME is easy because the GUI is so straightforward, plus the KNIME website contains a rich set of user guides and community-developed materials. However, there are very few published books that teach KNIME.

Of course, these are not all of the open source tools that might be applicable to a new healthcare data analyst. It may be with exploring other options such as GNU Octave (very similar to MATLAB), Julia, and Haskell.

Regardless of the open source tool you choose, new data analysts will find it helpful to utilize an open source tool to get them started without incurring the costs associated with a proprietary statistical software.

Nathan Patrick Taylor (taylornathan@msn.com) is a clinical informatics consultant with the Symphony Post-Acute Care Network.

Original source:
Taylor, Nathan Patrick. "Getting Your Feet Wet with Open-Source Statistical Languages" (Journal of AHIMA website), January 25, 2016.
