Automating SQL queries from formulas: loading data on demand

Thomas Lumley

Department of Biostatistics, University of Washington, Seattle. * Contact author: tlumley@u.washington.edu

Keywords: model formula, database, programming, large data, regression

A relatively common situation with large data sets is that the variables needed for any specific computation will fit in memory, but the entire data set is large enough to be inconvenient. For example, the 2006 NHIS public use data set has about 25,000 observations on 546 variables, and will take up about 100Mb, enough to slow down a computer with 1Gb memory. The 2007 BRFSS public use data has about 430,000 observations and 343 variables. The whole data set cannot be loaded into 32-bit R, but there is no difficulty in handling a dozen variables or so.

I will describe an approach to automated loading of data from a relational database by extracting the names of the necessary variables from a model formula or expression. This approach can be used to wrap existing code that is unaware of databases. Only read access to the database is needed, since newly defined variables are stored as definitions and created as the data is loaded.