Toward A Common Framework for Statistical Analysis and Development (Also Known As, "Zelig: Everyone's Statistical Software")[1]

Kosuke Imai[2] *Gary King*[3] *Olivia Lau*[4]

A draft of this paper is available at `http://gking.harvard.edu/files/z.pdf`.

Abstract

We propose a common framework for statistical analysis and software development built on and within the R language. Researchers in different academic disciplines have developed different statistical models, different mathematical notation, different parameterizations, different quantities of interest, and hence different computational implementations. The users of statistical software have much to gain by navigating the babel of R's many packages, but is is often far more difficult than it should be, given that packages have so much underlying statistical theory and computational structure in common.

To address this problem, we have developed a conceptual framework and software package that offers:

- A common formula syntax for specifying univariate response, multivariate response, multilevel, and hierarchical models. In particular, we offer a user-specified, intuitive interface for identifying constraints across equations.
- A method to use this syntax with existing R packages, without modifying those packages (by re-defining function calls on the fly).
- Developer tools to translate this syntax into matrices and arrays useful for programmers. For multiple equation models, we offer three implementation options (all of which work with user-specified constraints): an intuitive option that stacks matrices visually, a computationally efficient option that creates arrays of explanatory variables, and a memory-efficient option that coerces parameters to matrices.
- A framework for calculating quantities of interest with or without conditioning on the observed values of a particular unit, to provide users with substantive interpretation of model output.
- A framework to process lists of multiply imputed data, or stratified data, and combine model estimates when generating quantities of interest, to extend the reach of any one statistical model.
- An application programmer interface that makes it possible to dynamically generate a graphical user interface (GUI) for the models included in Zelig (see, e.g., the Virtual Data Center, for one example of a GUI which has already been implemented).

Or approach creates a common interface between users and developers: users have a common, intuitive syntax for running and interpreting complex statistical models; developers have a set of tools that make writing new models easier, and an interface that makes it easier for other people to use their software.

[1]Our thanks to the National Institutes of Aging (P01 AG17625-01), the National Science Foundation (SES-0318275, IIS-9874747), and the Mexican Ministry of Health for research support. Current software is available from CRAN or `http://gking.harvard.edu/zelig/`.

[2]Assistant Professor, Department of Politics, Princeton University, `kimai@princeton.edu`

[3]David Florence Professor of Government, Department of Government, Harvard University, `king@harvard.edu`

[4]Ph.D. Candidate, Department of Government, and M.A. Student, Department of Statistics, Harvard University, `olau@fas.harvard.edu`