Cornell University
Cornell Statistical Consulting Unit

StatNews #56

Managing and Mismanaging Iterative Procedures
February 2003

Although the most basic statistical procedures used involve algebraic equations that are solved directly, many statistical procedures, including common ones such as logistic regression, require iterative algorithms to search for the best solution. Software packages implement these algorithms with default rules to govern their behavior. Most of the time, we do not even think about or understand the underlying algorithms, and simply have faith in the results. But, sometimes the default rules can generate quite erroneous results.

Such a mistake was publicized in the June 14, 2002 issue of the journal Science. A published analysis from the National Morbidity, Mortality and Air Pollution study resulted in the overestimation of the risks of soot. In this landmark air pollution study, five years had past before the mistake was detected. Although the policy implications did not change after the discovery of the mistake, the adjusted estimates were very different and the mistake invited criticism from those opposed to the policy that had resulted from the first reported estimates.

The title of the Science article blamed the mistake on a software glitch rather than on the users failure to reassign default settings. The article claims that the erroneous results were due to a problem with the software application used by the team of very good and very careful researchers. The article warned scientists about using off-the-shelf statistics software without questioning what's inside.

The air pollution study utilized a generalized additive model implemented in S-Plus, which like all iterative procedures, searches for an answer, in this case a pollution effect, until the results no longer differ from the previous one by a certain amount. Since the study dealt with very small values, the so-called software glitch arose from the inappropriately large default setting of 0.001, causing calculations to terminate prematurely.

How do iterative programs decide when to converge? A variety of convergence criteria options are available, such as absolute and relative functions involving first or second derivatives, measures of overall fit, or parameter estimates. Some programs have a default limit on the number of iterations, so the program may stop before convergence. Typically, default settings can easily be changed via option statements.

Default settings are appropriate for most applications. But, when dealing with unusual circumstances, such as the fine-scale modeling being carried out in the air pollution study, attention must be paid to the appropriateness of the convergence criteria when using software programs to analyze data with iterative procedures. Such procedures include logistic regression, multi-level (i.e., mixed) models, cluster analysis, factor analysis, generalized additive models, and many others. It is up to the software user to use reasonable settings for their data and statistical problem because even the best software does not have default settings that are optimal for every problem.

References

- Kaiser J. Software glitch threw off mortality estimates. Science 296:1945-1947, 2002.
- Hurley M. Wetware problem, not a software problem. Science 297:936, 2002.
- Schaeffer A. Unfair characterization of industry response. Science 297:770, 2002.

Authors: Catherine Soria and Edward Frongillo

Back to StatNews Table of Contents

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).