Collaborative Automation Reliably Remediating Erroneous Conclusion Threats (CARRECT)

Jonathan C. Lansey*[1], Paul Picciano[1], Ian Yohai[1], Fred Grant[2] and Robert Gern[2]

[1]Aptima inc., Woburn, MA, USA; [2]Northrop Grumman Corporation, Falls Church, VA, USA

Objective

The objective of the CARRECT software is to make cutting edge statistical methods for reducing bias in epidemiological studies easy to use and useful for both novice and expert users.

Introduction

Analyses produced by epidemiologists and public health practitioners are susceptible to bias from a number of sources including missing data, confounding variables, and statistical model selection. It often requires a great deal of expertise to understand and apply the multitude of tests, corrections, and selection rules, and these tasks can be time-consuming and burdensome. To address this challenge, Aptima began development of CARRECT, the Collaborative Automation Reliably Remediating Erroneous Conclusion Threats system. When complete, CARRECT will provide an expert system that can be embedded in an analyst's workflow. CARRECT will support statistical bias reduction and improved analyses and decision making by engaging the user in a collaborative process in which the technology is transparent to the analyst.

Methods

Older approaches to imputing missing data, including mean imputation and single imputation regression methods, have steadily given way to a class of methods known as "multiple imputation" (hereafter "MI"; Rubin 1987). Rather than making the restrictive assumption that the data are missing completely at random (MCAR), MI typically assumes the data are missing at random (MAR).

There are two key innovations behind MI. First, the observed values can be useful in predicting the missing cells, and thus specifying a joint distribution of the data is the first step in implementing the models. Second, single imputation methods will likely fail not only because of the inherent uncertainty in the missing values but also because of the estimation uncertainty associated with generating the parameters in the imputation procedure itself. By contrast, drawing the missing values multiple times, thereby generating m complete datasets along with the estimated parameters of the model properly accounts for both types of uncertainty (Rubin 1987; King et al. 2001). As a result, MI will lead to valid standard errors and confidence intervals along with unbiased point estimates.

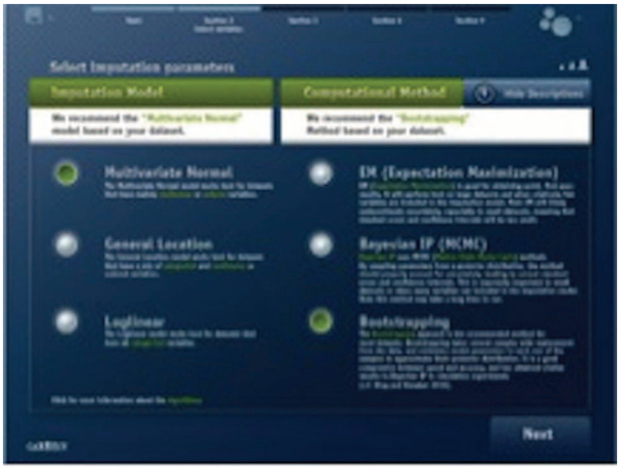
In order to compute the joint distribution, CARRECT uses a bootstrapping-based algorithm that gives essentially the same answers as the standard Bayesian Markov Chain Monte Carlo (MCMC) or Expectation Maximization (EM) approaches, is usually considerably faster than existing approaches and can handle many more variables.

Results

Tests were conducted on one of the proposed methods with an epidemiological dataset from the Integrated Health Interview Series (IHIS) producing verifiably unbiased results despite high missingness rates. In addition, mockups (Figure 1) were created of an intuitive data wizard that guides the user through the analysis processes by analyzing key features of a given dataset. The mockups also show prompts for the user to provide additional substantive knowledge to improve the handling of imperfect datasets, as well as the selection of the most appropriate algorithms and models.

Conclusions

Our approach and program were designed to make bias mitigation much more accessible to much more than only the statistical elite. We hope that it will have a wide impact on reducing bias in epidemiological studies and provide more accurate information to policymakers.



Select Imputation parameters

Imputation Model

We recommend the "Multivariate Normal" model based on your dataset.

● Multivariate Normal [illegible]

○ General Location [illegible]

○ Loglinear [illegible]

Computational Method

Hide Descriptions

We recommend the "Bootstrapping" Method based on your dataset.

○ EM (Expectation Maximization) [illegible]

○ Bayesian IP (MCMC) [illegible]

● Bootstrapping [illegible]

Next

Figure 1 - Screenshot of user selecting imputation parameters.

Keywords

Bias reduction; Missing data; Statistical model selection

Acknowledgments

This material is based upon work supported by the Walter Reed Army Institute of Research (WRAIR) under Contract No. W81XWH-11-C-0505. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the WRAIR.

References

James Honaker and Gary King, "What to do About Missing Values in Time Series Cross-Section Data" American Journal of Political Science Vol. 54, No. 2 (April, 2010): Pp. 561-581.

Gary King, James Honaker, Anne Joseph, and Kenneth Scheve. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation", American Political Science Review, Vol. 95, No. 1 (March, 2001): Pp. 49-69.

*Jonathan C. Lansey
E-mail: jlansey@aptima.com