JASA ACS Reproducibility Initiative - Author Contributions Checklist Form

The purpose of the Author Contributions Checklist (ACC) Form is to document the code and data supporting a manuscript, and describe how to reproduce its main results.

As of Sept. 1, 2016, the ACC Form must be included with all new submissions to JASA ACS.

This document is the initial version of the template that will be provided to authors. The JASA Associate Editors for Reproducibility will update this document with more detailed instructions and information about best practices for many of the listed requirements over time.

Data

Abstract

Verification of models fit to synthetic data requires two data sets: the synthesized observations being studied and authentic observations that the synthesized data represent. Analyst specified models are fit to both data sets and statistical results of fit to synthetic data are presented along with a measure of agreement in the authentic data. The synthetic data and verification system are designed to minimize risk of private information release using principles of differential privacy[*] and at no time are authentic data, or results of models fit to authentic data, released to users of the system. The current paper utilizes synthetic records intended as a substitute for authentic U.S. Office of Personnel Management (OPM) Central Personnel Data File (CPDF) records, consisting of one record for each synthesized federal employee and fiscal year for the period 1988 through 2011. Authentic CPDF records, made available by OPM in response to a Freedom of Information Act (FOIA) request, are used for verification.

Availability

5/15/2017 e-mail from Jerry Reiter to Montserrat Fuentes:

> Hi Montse!
>
> I had a question about materials for reproducibility for JASA. We have a manuscript where we analyze confidential data from the Office of Personnel Management in the government. Part of the innovation in the manuscript is that we develop a redacted (fully synthetic) version of the federal workforce, and also general tools to compare results on confidential and synthetic data using differential privacy. The point of the manuscript is to provide a framework and case study for an integrated approach to sharing confidential data.

[*] A detailed explanation of differential privacy is given in the paper

> We can't release the synthetic data or the confidential data, as the OPM has not yet approved either -- and won't likely do so for the foreseeable future with all the changes from having a new administration. We also do not have approval to release the code for generating the synthetic data. However, we can release the code used to do the analyses of the confidential data and the R package for verification.
>
> Do you have any advice on what we should include for reproducibility materials? Thanks.

5/15/2017 instructions from Montserrat:

> If you send the code to generate the analysis and a simulated version of the data (for reviewers to check code), with the justification provided [above] it will be sufficient.

FOIA Note: A Freedom of Information Act (FOIA) request can be placed with the Office of Personnel Management to obtain the authentic CPDF-EHRI data.

At this time, due to restrictions imposed by OPM, neither synthetic nor authentic data are available to the public. In the future, pending OPM approval, synthetic data and verification system results may be accessible to authorized users. A small set of CPDF-formatted observations, that maintains several basic covariate relationships observed in the synthetic data, is included with the on-line supplemental materials at https://github.com/DukeSynthProj/DukeSynthJASA2017 and should be adequate for testing of verification software, algorithms, and principles.

Description

Sample data file and supplemental materials are available at:

https://github.com/DukeSynthProj/DukeSynthJASA2017

File name: OPMSyntheticTestData.csv
File format: Comma separated ASCII, (CR-LF) line termination
Data dictionary: columns adhere to OPM Guide to Data Standards:
https://www.opm.gov/policy-data-oversight/data-analysis-documentation/data-policy-guidance/reporting-guidance/part-a-human-resources.pdf

Note that a subset of CPDF data elements were made available to Duke University by OPM and, of those, some have been redacted in the on-line supplemental test data.

Code

Abstract

Verification consists of fitting a user specified model to synthetic and authentic observation data (SD and AD, let SF and AF indicate synthetic and authentic fit, respectively) and presenting statistical results of the SF along with a measure of "agreement" in the AF. Generally, AF consists of individual models being fit to disjoint, complete partitions of AD. The proportion of partitions with user selected model parameter estimates within a tolerance band of or with sign, or other significant property, like those observed in the SF is presented. The user decides what proportion level constitutes agreement between the SF and AF.

Description

Supplemental software consists of two major components: a Microsoft SQL Server stored procedure script (for data retrieval) and a collection of R scripts for model fitting and verification measure computation and presentation. Source code and instructions on execution are available at https://github.com/DukeSynthProj/DukeSynthJASA2017. Source files include:

- DIBBSDataRetrieval.sql – This file contains SQL instructions to:
    - Create empty authentic and synthetic CPDF observation tables (CPDFNonDODStatusJdF2012, CPDFNonDODStatusJdF2014, and CPDFNonDODDIBBS2017) structured to accept data as delivered by OPM. Note that the supplied test data (OPMSyntheticTestData.csv) can be imported into these enabling SQL data retrieval as executed in the verification functions.
    - Create the view (CPDFNonDODStatusDIBBS2016) referenced by the DIBBSData stored procedure (below) to retrieve versioned synthetic observations.
    - Create the DIBBSData stored procedure that is called by the verification R scripts. There are three primary methods of execution:
        - Retrieve authentic observations: exec DIBBSData @style = 'LargeFixedEffectsModelNonUniqueIDJdF', @WorkSchedule = 'FullTime'

    - Retrieve synthetic observations: exec DIBBSData @style = 'LargeFixedEffectsModelNonUniqueIDDIBBS', @WorkSchedule = 'FullTime'
    - Retrieve observations as requested by verification server functions: exec DIBBSData @style='VerificationData'
    - Additional information on options, parameters, and modes of execution are contained in the scripts
- VerificationMeasure.r – R script that contains example sessions for executing the threshold and three-point longitudinal verification measures presented in the paper. Query and computation functions from VerificationServer.r (below) are called from this procedure. Additional instructions on use are contained in the source file.
- VerificationServer.r – R script containing primary verification system functions, including:
  - queryObservations() – queries SQL synthetic and authentic observations based on user supplied filtering specification
  - fitFEModel() – fits a fixed effects model to user specified synthetic or authentic data set, a data frame formatted as the result of queryObservations(); calls feXTX() (below); returns model fit results
  - thresholdMeasure() – executes the threshold verification measure presented in the paper; expects two queryObservations() style data frames (one for synthetic data one for authentic data); partitions authentic observations; calls fitFEModel() and DP.threshold() Laplace posterior noise algorithm; returns list of verification measure results
  - longitudinalThreePtMeasure() - executes the three-point longitudinal verification measure presented in the paper; expects two queryObservations() style data frames (one for synthetic data one for authentic data); partitions authentic observations; calls fitFEModel() and DP.threshold() Laplace posterior noise algorithm; returns list of verification measure results
- LaplaceDistribution.r – contains pdf, cmf, and random quantile functions based on the Laplace distribution; called by DP.threshold() (below)
- BetaBinomialLaplacePosteriorDP.r – Laplace noise algorithm; primary function is DP.threshold(), called by verification measure functions, calls Laplace distribution functions
- FixedEffectsMatrixSolution.r – contains function feXTX() that fits OLS fixed effects models to large observation, high dimensional data; called by fitFEModel() in verification measure process
- CommonFunctions.r – contains functions for common tasks, such as computing the mode of an empirical distribution

Complete instructions on use of each function are provided in the source files. Questions and comments may be directed to thomas.balmat@duke.edu

Optional Information

Supplemental data and software were developed on a Microsoft Windows 7 server using Microsoft SQL Server 2014 and R x64 version 3.2.1.

R library requirements: RODBC and parallel.

Instructions for Use

Reproducibility

Due to the prohibition on data release imposed by OPM, exact tables and graphs from the paper are not reproducible from supplemental materials. However, for purposes of testing the concepts and methods of verification measures as presented, supplemental procedures can access the supplied test synthetic data (OPMSyntheticTestData.csv). It is recommended, for seamless operation, that the supplied test data be uploaded to a SQL database and accessed using the supplied query functions. An alternative method would be to read the supplied test observations into a data frame formatted as required by the verification functions. Note that, since the test data do not represent covariate relationships observed in either the synthetic or authentic data, results and analyses derived from it should not be considered meaningful.

Complete instructions for execution of verification measure functions are included in corresponding source files. The reader is requested to begin a review of overall execution flow with the VerificationMeasure.r script, function thresholdMeasure() for threshold measure processing and function longitudinalThreePtMeasure() for longitudinal measure processing.