thomas.balmat@duke.edu Thomas J. Balmat 919-820-1057

Research Data Analyst and Statistician, Duke University, Durham, North Carolina. Projects in the Social Science Research Institute, the Department of Statistical Science, and Duke University Medical School.

The Social Science Research Institute at Duke University is home to fourteen centers for research in the social sciences. I am involved in the U.S. Federal Employee Human Capital Project, which is composed of faculty and researchers from diverse disciplines including Political Science, Economics, Sociology, and Statistics. The project conducts research into the policy value and effectiveness of the skill, experience, and education assets of non-military employees of the U.S. federal government. This involves intensive data exploration and management, development of novel statistical modeling methods, and development of high performance solutions to computationally complex models. Data sets include large, publicly available and private databases of federal employee career profiles and historical economic, demographic, and socio-political data.

Roles and responsibilities

- Design and implement a SQL database to house and secure over 600,000,000 observations spanning more than forty years of historical data related to U.S. federal workforce human capital, the economy, and governmental policy sourced from Freedom of Information Act (FOIA) requests and sources such as FedScope, The Current Population Survey, The Centers for Disease Control, and customized data sets provided by researchers at affiliate institutions such as Emory University, Vanderbilt University, and UCLA
- Assist researchers with SQL query development and general data access, transformation, and efficiency improvement
- Research and develop high performance computational strategies and functions for fitting high dimension ($10^7 \times 10^5$ and beyond) fixed effects regression, quantile regression, logistic regression, and multinomial models using R, SQL integration, parallel processing, and C
- Assess the performance of research computing servers using high dimension regression and probability models, measure results across wide input parameter intervals, report findings, and recommend alternative configurations

Resources related to the Human Capital project

- Research paper: Efficient Large Fixed Effects Regression Algorithm (author, preparing for submission, includes sections on efficient computation of robust, clustered standard errors for large problems)
  Supplementary material
  R package
  Rcpp code
- Cholesky decomposition and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ algorithms for solving large FE problems (estimates and standard errors)
  R package
  Rcpp code
- Research paper: Elections, Ideology, and Turnover in the U.S. Federal Government (contributor)
- Presentation: Efficient Solution of Large Fixed Effects Regression Problems (presented at useR!2017)
- Presentation: Approaches to Solving Large Computational Problems (presented at Joint Statistical Meetings 2017)
- Study: Research computing server performance evaluation

The Synthetic Data Project, led by faculty of the Department of Statistical Science at Duke University, with participating faculty from the Fuqua Business School and Department of Computer Science, along with members from the Office of Information Technology at Duke and faculty from the Political Science Department at Emory University, develops and applies theory and methods for generating privacy protecting versions of the authentic federal human capital database that maintains career longitudinal and covariate relationships, for 3,500,000 employees over twenty four years, while limiting risk of revealing sensitive information on individual employees.
Roles and responsibilities

- Design and implement SQL instances of six versions of synthetic federal employee human capital data
- Develop SQL queries to integrate authentic and synthetic data for data transformation and research model fitting
- Collaborate with human capital, data privacy, and statistics researchers to develop and fit models for assessing agreement of covariate relationships with those observed in authentic data, risk of private authentic information disclosure, and performance of synthesis probability models
- Analyze numerical and statistical privacy protecting algorithms for improvement of efficiency and compactness; test, certify, and report mathematical properties of resulting probability distributions

- Develop efficient statistical and graphical methods for fitting hundreds to thousands of (typically regression, probability density, covariance, or Markov) models to various subsets of data for wide scale model evaluation
- Develop statistical classifier algorithms, logistic models, and Receiver Operating Characteristic methods to measure probabilities of revealing authentic individual information from generated synthetic data using training patterns derived from publicly available data sets

Resources related to the Synthetic Data project

- Research paper: Providing Access to Confidential Research Data Through Synthesis and Verification (coauthor)
  Synthesis models supplement
  Authentic/synthetic model and covariate distribution validation supplement
  Verification measure sensitivity supplement
  Verification measure regression variance sensitivity analysis
  Verification measure sensitivity simulation
- Research paper: Empirical Evaluation of Privacy in Synthetic Data (coauthor, forthcoming)
- Research article: Improved Efficiency Multinomial Differential Privacy Algorithm (coauthor, in-progress)
- Study: Authentic Gender Disclosure Risk Assessment Using Synthetic Observations

---

The Hi-Host Phenome Project (H2P2) Project at Duke University Medical School seeks to interpret human genetic variation through a study of cell biology. H2P2 uses a genome-wide association study (GWAS) of cellular traits to provide a means to decipher how human genetic variation regulates cellular pathways, while also providing a system for experimental validation and mechanistic studies. The project employs a data set of more than 10,000,000,000 observations that record the association of more than 15,000,000 genetic variants in more than 500 subjects to phenotypic responses for approximately 150 pathogens (diagram). The goal of the project is to provide an efficient and intuitive on-line platform for local and public geneticists and cell biologists to conduct real-time research and genotype-phenotype association hypothesis testing by using underlying fast data query and model fitting processes.
Roles and responsibilities

- Design and implement a SQL database for efficient query of billions of observations across multiple tables
  - Records for more than 15,000,000 genetic variants (SNPs) for each of more than 500 subjects were partitioned into 493 disjoint tables by SNP (the choice of 493 yielded maximum balance of SNP observations per table; choice of partitioning by SNP resulted from a study of expected primary query filtering parameters)
  - More than 2.2 billion GWAS records (one for each SNP and pathogen) were partitioned into 149 disjoint tables by pathogen (partitioning by pathogen provided balanced tables with optimum alignment to expected query parameters)
  - Partitioning of high dimension genotype and GWAS tables reduced typical query execution time from more than ten minutes to less than ten seconds
- Design and implement an on-line R (Shiny) application for public GWAS research query and association exploration
  - Develop query tool to explore associations between researcher specified SNPs, phenotypes, and genes
    - Results presented in tabular format with selectable rows by SNP and phenotype
    - Row selection generates boxplot of phenotypic response by genotype, with filtering features
    - Table download feature permits further local analysis of selected GWAS results
  - Develop interactive heatmap to explore Spearman (rank) correlations of specified phenotypic responses
  - Develop Manhattan plot of specified phenotype-genotype regression estimates for identification of association of phenotype to chromosome-gene position

Resources related to the H2P2 project

- On-line app: H2P2 GWAS application (developer)
- Paper: An Atlas of Genetic Variation Linking Pathogen-Induced Cellular Traits to Human Disease (contributor)

Previous experience

- Laboratory Data Manager and Statistician, Cormetech Inc., Durham, North Carolina
- Operations Analyst, Data Analyst, Hanson Pipe and Precast, Dunn, North Carolina
- U.S. Air Force, Statistical Analyst, four years, honorably discharged, final grade E-5
- Resume of previous experience

- Summary of previous projects

Education

- Undergraduate curriculum in Mathematics, University of Houston, NC State University
- Graduate curriculum in Statistics and Industrial Systems Engineering (Mathematical Programming and Probability), NC State University
- Graduate course descriptions and transcript

Additional resources

- Problem Solver - mostly recreational probability problems and computational pastimes
- Other projects - interesting work related problems and solutions
- An application of Linear Approximation of Variance - useful for modeling dispersion of complex functions (contains generators and theoretical derivations of several arcane, but useful, probability distributions)
- Jamaica storm prep - a hypothetical do/don't go fishing decision process for Jamaican fishermen using World Bank damage estimates, NOAA historical hurricane category risk models, and Markov processes