Kangrui Liu

+1 (202)868-9664

krliu67@umd.edu Linkedin/kangrui-liu Github/krliu67

Education

University of Maryland **College Park, Maryland, US**

Master of Science in Survey and Data Science at Joint Program of Survey Methodology *Aug 2023 - May 2025[expected]*

Ningbo University of Technology **Ningbo, Zhejiang, CN**

Bachelors of Science in Information and Computing Science at School of Statistics and Data Science *Sep 2019 - May 2023*

Research

Boosted Pseudo-Weighting for Nonprobability Samples to Improve Population Inference

Research Assistant, Supervised by Prof. Yan Li and Dr. Lingxiao Wang *Dec 2023 - Present*

- Developed and implemented the "PS-GBM" pseudo-weight construction methods, integrating gradient boosting method with traditional propensity-score adjustments to enhance the representativeness of nonprobability samples.
- Conducted Monte Carlo simulations to evaluate the performance of these methods, demonstrating improvements in bias reduction compared to traditional logistic regression-based approaches.
- Using the National Health And Nutrition Examination Survey III as a nonprobability sample and the 1997 National Health Interview Survey as a reference, the 15-year incidence of diabetes was estimated with a 20% improvement in relative bias compared with traditional methods.

Experience

Machine Learning for Social Science **College Park, Maryland, US**

Grader *Feb 2025 - Present*

- Grade assignments and Provide feedback on implementing machine learning workflows in R, including data preparation, model tuning, and evaluation, with a focus on bias-variance trade-offs and performance metrics.

Jianxin Technology **Guangzhou, Guangdong, CN**

Intern *Feb 2023 - May 2023*

- Corporate Resource Library Optimization
  - Designed and deployed a MinIO-based file storage and transfer system, leveraging Fast Transfer and Instant Upload features; reduced average upload latency by 30%, enhancing document accessibility and team productivity.
  - Developed a dynamic file prioritization algorithm based on metadata (e.g., file size, classification level), which improved system throughput and ensured high-priority files were processed within 5 seconds under peak load.
- Power Plant Gate Detection Project
  - Labeled over 1,200 images using LabelMe to identify gate status (open/closed) for downstream ML model training.
  - Implemented preprocessing pipelines in Python (e.g., grayscale conversion, bounding box filtering), improving feature extraction accuracy by 18% for initial model iteration.

Projects

**** stands for ongoing projects***

Correcting Non-Participation Bias in Physical Activity Surveys*

Capstone Project Supervised by Prof. Brady T. West *Jan 2025*

- Investigate non-participation bias in the 2010-2011 Physical Activity and Transit Survey, comparing self-reported physical activity (PA) data from New York City adults with simulated accelerometer-measured PA to assess selection differences.
- Apply Adjusted Logistic Propensity (ALP) weighting with machine learning (LASSO) to select predictive covariates (e.g., age, BMI, neighborhood walkability), reducing bias in PA estimates across work, and transportation domains.
- Conduct simulation study to verify the validity of ALP, comparing ALP-weighted self-reported PA with true PA values, which showed a RMSE reduction of 5 compared with the unweighted and traditional methods.

National Survey Design of Undergraduate Students

Consulted for the Pew Research Center and mentored by Prof. Michael R. Elliott *Jan 2025 - Apr 2025*

- Collaborated with a team to design a split-frame survey for the Pew Research Center, integrating USPS Address-Based Sampling (ABS) and Generation Lab's online panel to optimize coverage of U.S. undergraduates.
- Developed a logistic regression model using ACS, IPEDS, and PUMS data to predict off-campus student density at the census block level, identifying 89% of off-campus undergraduates in 23.5% of Ann Arbor blocks for ABS stratification.

- Evaluated frame-specific errors (e.g., ABS dormitory undercoverage, panel self-selection bias) and proposed stratification by student density and post-stratification with IPEDS dormitory capacity to enhance representativeness and cost-efficiency.

Political Media Consumption and Voting Behavior in the UK

Applications of Statistical Modeling *Dec 2024*

- Analyzed the relationship between political media consumption and voting behavior using data from the European Social Survey. Applied Generalized Linear Models, Generalized Linear Mixed Models with random intercepts, and Generalized Estimating Equations.
- Conducted data normalization and extensive variable selection to isolate key predictors, including age, income, and media exposure. Findings revealed that higher political media consumption, age, and income significantly increased voting likelihood, while regional effects were negligible.

Sampling and Estimation Plan for the Michigan Teen Survey

Applied Sampling *Apr 2024*

- Developed a two-stage sampling plan to support Michigan's Department of Education in monitoring teen smoking and drug use rates. Focused on achieving a coefficient of variation below 0.05, ensuring equal sampling rates across strata, and strategically linking or collapsing strata for effective sample selection.
- Calculated design effects to manage variance inflation, employed Jackknife Repeated Replication for variance estimation, and adjusted for nonresponse to maintain precision and reliability. The plan informed decisions on sample size and budget, ensuring robust point estimates for key variables.

Childlessness Trends in New Jersey

Machine Learning for Social Sciences *Apr 2024*

- Analyzed survey data from Kaggle to investigate economic, social, and health-related predictors of childlessness in New Jersey. Applied machine learning models including LASSO, Decision Trees, SVM, and Gradient Boosting to identify key influencing factors.
- Focused on model interpretability and predictive accuracy; findings informed recommendations on policy interventions for individuals facing barriers to parenthood.

Content Analysis of Reddit User Responses to the SFFA Case

Data Display and Computing *Dec 2023*

- Collected Reddit comments via API on discussions surrounding the *Students for Fair Admissions v. Harvard* case. Cleaned and preprocessed unstructured text data for downstream analysis.
- Applied Latent Dirichlet Allocation (LDA) for topic modeling and conducted sentiment analysis using `NLTK` and `TextBlob`. Performed time series analysis to track sentiment dynamics, revealing shifts in public discourse and emotional trends.

Strength Evaluation on Listed Companies of Zhejiang Province

Data Analysis *May 2022*

- Crawled the 2019 annual report data of Zhejiang companies with Python, and randomly sampled 418 listed companies.
- Chose companies from top 3 industries to conduct Principal Component Analysis to assess the strengths of companies within and across industries.

First Prize in Zhejiang Province in CUMCM (Problem C) (TOP 10%) **Oct 2021**

- Analyzed the order and freight data in the past 24 months; used Greedy algorithm, Monte Carol Simulation and Dynamic Programming to achieve an relative optimal result for order and transportation plans in the next 12 months.
- Building mathematical model, and employing Monte Carol Simulation to obtain quasi-optimal solutions and simplify computation speed.

Skills

Programming & Computing: Proficient in Python and R; working knowledge of SQL, Java, C, HTML/CSS, and JavaScript; experienced in Git, Linux, and SLURM for version control and high-performance computing.
Machine Learning & AI: Applied Scikit-learn, XGBoost, TensorFlow, and PyTorch for supervised learning tasks; expertise in gradient boosting, ensemble methods, and regularized regression (e.g., LASSO, Ridge); actively expanding into representation learning and trustworthy AI (e.g., fairness, causality, robustness).
Statistical Modeling & Causal Inference: Skilled in logistic regression, Bayesian modeling, Monte Carlo simulations, and propensity score weighting; interested in causal representation learning and domain adaptation for biased data.
Survey Methodology & Data Collection: Expertise in survey sampling design, power analysis, variance estimation (e.g., JRR, bootstrap), and total survey error framework; familiar with Qualtrics and API-based data extraction.
Languages: Native in Mandarin Chinese; proficient in English and Cantonese.