Kangrui Liu

+1 (202)868-9664

✉ krliu67@umd.edu Linkedin Github

Education

University of Maryland **College Park, MD, US**
Master of Science in Survey and Data Science at Joint Program of Survey Methodology *Aug 2023 - May 2025*

Ningbo University of Technology **Ningbo, Zhejiang, CN**
Bachelors of Science in Information and Computing Science at School of Science *Sep 2019 - May 2023*

Research Experience

Research Assistant

Supervised by Prof. Tianzhou Ma and Prof. Edmond D. Shenassa *May 2025 - Present*

- Investigated chronic stress disparities through cohort-based modeling of Allostatic Load (AL) using All of Us biomarker data, integrating epidemiological context with statistical modeling.
- Developed an imputation strategy for incomplete biomarker profiles, enabling valid between-group comparisons across race, income, and age cohorts. Interpreted AL differences to inform health equity research and public health interventions.

Research Assistant

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

- Proposed and evaluated two pseudo-weighting frameworks incorporating gradient boosting to adjust for selection bias in nonprobability samples.
- Designed simulation studies to assess finite-sample bias and variance under different sampling and outcome scenarios. Managed HPC (SLURM) for model tuning and bootstrap variance estimation.

Publications ***(* indicates co-authors)***

1. **Liu, K.***, Wang, L.*, Li, Y. *"Gradient-Boosted Pseudo-Weighting: Methods for Population Inference from Nonprobability Samples."* Working paper, 2025.

Professional Experience

Intern **Guangzhou, Guangdong, CN**

Jianxin Technology *Feb 2023 - May 2023*

- Built a MinIO-based file storage system in Java, reducing upload latency by 30%; designed priority algorithm for high-load file processing.
- Labeled 1,200+ images and built preprocessing pipelines in Python, enhancing feature consistency and model interpretability for gate status classification.

Applied Statistical Projects

Bias Adjustment in Self-Reported Physical Activity Data

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

- Analyzed selection bias in NYC's 2010–2011 Physical Activity and Transit Survey by comparing self-reported physical activity data with simulated accelerometer benchmarks.
- Used machine learning (LASSO) to select key predictors (e.g., age, BMI, neighborhood walkability) and applied ALP weighting to adjust for participation bias.
- Validated approach via simulation, showing improved alignment between adjusted and true physical activity estimates.

National Survey Design of Undergraduate Students

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

- Helped design a split-frame survey for Pew Research Center by integrating USPS ABS with an online panel to improve coverage of U.S. undergraduates.
- Built logistic regression models using ACS and IPEDS data to identify areas with high off-campus student density for stratified sampling.
- Assessed potential coverage errors in both frames and recommended stratification and post-stratification adjustments to improve representativeness and reduce cost.

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.

Interpretable ML for Social Determinants of Fertility

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.

Awards

1. Travel Award, "The Past, Present and Future of Statistics in the Era of AI" — GWU Statistics Dept. 90th Anniversary Conference, May 2025.
2. Successful Participant (Top 40%) — USA Mathematical Contest in Modeling, Feb 2022
3. Zhejiang Provincial First Prize (Top 10%) — China Undergrad Mathematical Contest in Modeling, Oct 2021

Professional Memberships

- American Statistical Association - Washington Statistical Society

Activities

- Grader, Machine Learning for Social Science ($\sim$ 20 Students) in Spring 2025

Skills

- Python, R, SQL; Git, Linux, SLURM; Docker, Java, C, HTML/CSS, JavaScript