On Effective and Efficient Survey Data Collection at the Crossroads of Statistics and Artificial Intelligence

Survey research faces escalating data collection costs and declining response rates. To maintain valid population inference, methodologies must evolve. This prompts the adoption of Adaptive Survey Design (ASD), which tailors data collection strategies to population subgroups, balancing quality and costs. While ASD relies heavily on statistical modeling, its execution can be optimized by complementing established methods with modern computational tools. This dissertation explores the synergy between traditional survey statistics and artificial intelligence, developing strategies that effectively mitigate nonresponse bias and efficiently allocate fieldwork resources.

The first study addresses subgroup formation for ASD when response signals are weak and traditional predictive trees fail. It proposes a tree-like stratification method partitioning the auxiliary variable space to minimize the within-group standard deviation of estimated response propensities. Splitting and trimming are governed by a one-sided *z*-test evaluating if this standard deviation is significantly greater than zero. Evaluated through a simulation using the American Community Survey data, the method consistently reduces bias and Root Mean Square Error of survey estimates compared to random allocation of data collection strategies. However, because nonresponse is conditioned on auxiliary variables related to target outcomes, bias remains when adjustments are restricted to observed sample data without external benchmarks. Confidence interval coverage declines as the sample size increases, because the sampling variance shrinks around the remaining bias. Overall, the stratification method provides a robust ASD framework by identifying actionable subgroups that differentiate response behavior.

The second study builds upon this framework by investigating a two-step method for estimating response propensities to further minimize nonresponse bias. Rather than using standard one-step propensity models, this approach first predicts target survey variables and utilizes these predictions as propensity proxies. Utilizing the stratification method from the first study, this study evaluates the approach across diverse

nonresponse scenarios. The findings confirm it provides double robustness: specifically, a strong predictive model for the target survey variable can protect against a mis specified response propensity model. Naturally, the best performance occurs when both models are correctly specified, as seen in scenarios where auxiliary variables correlate highly with target variables under the Missing at Random mechanism. In these optimal conditions, two-step ASD leads to modest reductions in bias for household income compared to the one-step method. These benefits also extend to correlated secondary target variables, demonstrating an effective strategy for mitigating bias when the survey design focuses on a single topic where target outcomes share similar characteristics.

The third study addresses the inefficiency of futile outreach, exploring a pipeline leveraging Active Learning (AL) and Natural Language Processing (NLP) to dynamically prioritize respondent recruitment. In order to simulate this approach, a Large Language Model is used to convert structured auxiliary data into textual profiles. The AL process then uses these profiles to identify individuals most likely to respond for field interviewers. As recruitment outcomes are recorded, the NLP classifier iteratively updates, re-ranking remaining units in real-time. A simulation using the Dutch General Population Survey data (n = 21,200) demonstrates that with individual-level auxiliary variables, 31 percent of the sample could be bypassed from Computer-Assisted Personal Interviewing while successfully recruiting 95 percent of the individuals in the respondent set. Emphasizing data privacy, this AI-assisted fieldwork model significantly optimizes human effort by dynamically learning response patterns without relying solely on historical data.

Yongchao Ma is a Ph.D. candidate in Survey and Data Science at the Institute for Social Research, University of Michigan. His research focuses on survey nonresponse, adaptive survey design, the application of artificial intelligence in surveys, and statistical modeling in the social sciences. His recent work has been published in *Sociological Methods & Research* and *Drug and Alcohol Dependence*. He holds a B.A. in Sociology from China, an M.S.Sc. in Criminology from the University of Macau, and an M.Sc. in Methodology and Statistics from Utrecht University.