Bivariate Analysis of Distribution Functions Under Biased Sampling

Hsin-wen Chang[1], Shu-Hsiang Wang[1]

[1]*Institute of Statistical Science, Academia Sinica, Taipei, Taiwan*

ABSTRACT

We compare distribution functions among pairs of locations in their domains, in contrast to the typical approach of univariate comparison across individual locations. This bivariate approach is studied in the presence of sampling bias, which has been gaining attention in infectious disease studies that over-represent more symptomatic people. In cases with either known or unknown sampling bias, we introduce Anderson--Darling-type tests based on both the univariate and bivariate formulation. A simulation study shows the superior performance of the bivariate approach over the univariate one. We illustrate the proposed methods using real data on the distribution of the number of symptoms suggestive of COVID-19.

Keywords: Bootstrap; Empirical distribution function; Size bias; Two-sample test

Regression in 2-Wasserstein Distance

Li-Shan Huang[1], Zhezhen Jin[2], and Ting-Wei Hsu[1]

[1]*Institute of Statistics and Data Science, National Tsing Hua University, Hsinchu, TAIWAN*

[2]*Department of Biostatistics, Columbia University, New York, USA*

ABSTRACT

With the advancement of technology, there is a growing need for regression tools capable of modeling relationships in which the response variables are histograms or probability density functions, and the covariates lie in Euclidean space. In this work, we revisit classical linear regression and reformulate it from the perspective of minimizing 2-Wasserstein distances. The results lead to the development of 2-Wasserstein distance regression for histogram or density response objects with Euclidean predictors. The performance of the proposed methodology is evaluated through simulations, and its practical application is demonstrated through an analysis of mortality data.

Keywords: Best linear unbiased estimator; Gauss-Markov Theorem; Least squares

Debiased Inference for High-Dimensional Censored Quantile Regression

Yu Guo, **Tony Sit**

Department of Statistics and Data Science, The Chinese University of Hong Kong

ABSTRACT

This paper introduces a novel methodology for constructing confidence intervals in high-dimensional censored quantile regression, where the number of covariates may substantially exceed the sample size. Building upon the weighted loss function proposed by Wang and Wang (2009), we incorporate an L1 penalty to handle high dimensionality and apply a debiasing procedure to correct the inherent bias introduced by the LASSO estimator. The resulting debiased estimator is shown to be asymptotically normal, forming a solid foundation for valid statistical inference. Notably, our approach relaxes the conventional global linearity assumption to a local linearity condition near the quantile of interest, enhancing model flexibility and robustness—especially in the presence of heteroskedasticity or violations of global linear effects. Simulation studies demonstrate the superior performance of our method in terms of coverage accuracy and efficiency when constructing confidence intervals, compared to existing approaches. The practical utility of the method is further illustrated through an application to the lung cancer dataset of Shedden (2008), yielding the identification of potentially significant genes associated with survival outcomes.

Keywords: Conditional quantiles; Right-censored data; Post-selection inference; Debiasing

Data Integration in Survey Sampling and Official Statistics

Changbao Wu

Department of Statistics and Actuarial Science, University of Waterloo

ABSTRACT

We discuss issues with and techniques for combining information from multiple sources under settings involving probability samples, non-probability samples, multiple-frame survey samples, known population controls from a census or estimated population controls from other surveys. Two main principles, namely, *validity* and *efficiency*, for methodological developments are discussed under different scenarios. Inferential techniques are reviewed under three general frameworks for analysis of non-probability survey samples, namely, inverse probability weighting, mass imputation, and doubly robust estimation. Some recent research on using modern machine learning techniques for data integration will be briefly discussed.

Keywords: calibration; estimating equations; non-probability samples; variance estimation