Washington Statistical Society Seminar

Title: A Regression Tree Approach to Missing Data in Sample Surveys

Speaker: Wei-Yin Loh, PhD, Department of Statistics, University of Wisconsin, Madison
Chair: Katherine E. Irimata, Mathematical Statistician, National Center for Health Statistics
Date: Monday, January 29, 2024
Time: 3:30 – 4:30 PM EST

Abstract:
Missing data are ubiquitous in sample surveys. The usual approach is imputation, but sometimes the unobserved values do not exist (e.g., age of spouse when there is no spouse). Other times, missingness per se is much more informative than the unobserved value and imputation would remove the information (e.g., missing blood type for reasons other than the unobserved blood type). Even when sensible, standard approaches, such as hot deck and sequential regression imputation, are hampered by modeling and computation difficulties. If parametric regression models are used to perform the imputations and predictor variables have missing values, the latter need to be imputed as well. Then problems with model misspecification and missingness assumptions multiply because they affect every variable with missing values. For example, the MAR (missing at random) assumption frequently used to justify imputation needs to hold individually for each variable imputed. Besides, propagation of errors resulting from sequential imputation are impossible to track.

The talk will present a new approach to missing data using the GUIDE regression tree algorithm. Because a regression tree model is nonparametric and asymptotically consistent, the question of model misspecification does not arise. More importantly, because GUIDE does not perform imputation of missing values in predictor variables, problems with propagation of imputation errors and MAR assumptions in the latter variables disappear. GUIDE can also deal with multiple types of missing values, as often occurring in survey data. Examples used for illustration include the BLS Consumer Expenditure Survey, the USDA Agricultural Resource Management Survey, and the electronic health records of a large set of COVID-19 patients. Publications and software for GUIDE are available from
https://pages.stat.wisc.edu/~loh/guide.html

About the speaker:
Wei-Yin Loh is Professor of Statistics at the University of Wisconsin, Madison. He has a PhD in statistics from the University of California, Berkeley, and is a fellow of the American Statistical Association and the Institute of Mathematical Statistics. He has held visiting fellowships at AbbVie, the Bureau of Labor Statistics, and IBM Research. He has consulted for some major biopharma companies. Professor Loh has been researching classification and regression tree algorithms and their properties for almost 40 years. He is the author of the GUIDE algorithm and software.

For additional information about this event, please contact Katherine Irimata (oui4@cdc.gov), WSS Methodology Program Chair.

Join Zoom Meeting
https://gwu-edu.zoom.us/j/99767670330

Meeting ID: 997 6767 0330

One tap mobile
+13017158592,,99767670330# US (Washington DC)
+13052241968,,99767670330# US