Semiparametric Estimation of the Cure Fraction in Population-based Cancer Survival Analysis

Ennan Gu
Department of Statistics
University of South Carolina
egu@email.sc.edu

In population-based studies, cured patients is defined as when the mortality rate of the individuals in diseased group returns to the same level as that expected in the general population. The population level mortality can be presented by the known background mortality rate with specified subgroups of interest, such as mortality rate by gender and age in the United States. The background mortality can be incorporated into the mixture cure proportional hazard (MCPH) model to provide estimations of the cure fraction in population-based cancer studies. The semiparametric estimation method based on the EM algorithm for the MCPH model with background mortality is further developed. The proposed method is validated via the comprehensive simulation studies and illustrated via a real data set.

Identification of Differences in Cortical Thickness in Multiple Sclerosis Patients based on Race

Jiajing Niu
Department of Mathematical Sciences
Clemson University
jiajinn@g.clemson.edu

We investigate the association between the cortical thickness and the ethnicity of multiple sclerosis (MS) patients. Our dataset is based on MRI data from Greenville Health System. Our method uses the publicly available MRI preprocessing tools Advanced Normalized Tools (ANTs) and FMRIB Software Library (FSL) to perform the cortical thickness pipeline that contains image registration, bias correction, tissue segmentation, and cortical thickness estimation. We explore the effect of race on the cortical thickness measurements based on analysis of variance (ANOVA) by thresholding the computed t-statistics using (local) false discovery rate (FDR) control based on empirically estimated null distribution, and locate regions in which we observe significantly different thickness between African Americans and Caucasian patients diagnosed with MS.

Hypothesis testing framework for dichotomization

Peter Green
Department of Public Health Sciences
Medical University of South Carolina
greenepe@musc.edu

Dichotomization of a continuous predictor to discriminate a binary outcome is widely-used in clinical settings. Dichotomized variables provide clinicians with easily interpretable decision rules for diagnosis and prognosis among many other benefits. Despite its ubiquitous use, dichotomization of continuous predictors is heavily criticized by the statistical community for resulting in too much information loss. In the event that dichotomizing a continuous predictor is necessary, the current methods available for choosing a cut-point from the data fail to address questions of the appropriateness of the continuous variable for dichotomization. Here we provide a modification to the current information loss paradigm which quantifies the information loss associated with estimating the relationship between a continuous predictor and a binary outcome using either the continuous or dichotomized form of the predictor. Additionally, we develop a hypothesis testing framework for ascertaining the appropriateness of dichotomizing a continuous variable to discriminate a binary outcome and conduct simulations to evaluate the proposed framework and the impact of varying parameters of the associated test statistic.

Equivalence of Maximum Likelihood Estimates of Parameters from Distributions using Two Common Methods in Limit of Detection Data Applications

Lutfiyya Muhhammed
Department of Public Health Sciences
Medical University of South Carolina

AUTHORS: Lutfiyya N. Muhammad, MPH; Viswanathan Ramakrishnan, PhD; Paul J. Nietert, PhD

INTRODUCTION: When measurements fall below a known limit of detection, different approaches have been suggested for estimating parameters from the underlying complete distribution. We propose that methods utilizing truncated distributions yield estimates equivalent to those obtained by approaches that treat the data points as censored.

METHOD: A theoretical relationship paralleling the maximum likelihood estimation approaches for parameters from any left truncated distribution and any distribution with left censored observations is provided. We conducted a simulation study to illustrate the equivalence of the left truncation (LT) and left censoring (LC) approaches. A simulation study is presented with various sample sizes and percentages of below lower limit of detection observations from a normal distribution with a mean of 5 and variance of 4.

RESULTS: In the simulated scenarios, the two approaches produced similar parameter estimates. For example, when 30% of 500 observations were below a lower limit of detection, the means and variances were estimated within 0.1% to 0.3%, respectively (LT: estimated mean of 4.998, estimated variance of 3.990; LC: estimated mean of 4.997, estimated variance of 3.996).

CONCLUSION: Left truncation and left censoring approaches for computing maximum likelihood estimates of parameters from underlying complete distributions are equivalent.

A Logic Ensemble Model for Identification of Interactions Associated with Continuous Disease Phenotypes

Sherry Livingston
Department of Public Health Sciences
Medical University of South Carolina
livingss@musc.edu

Many diseases have complex etiologies arising from interactions among genetic and environmental factors. If increasing disease severity is due to interactions between factors, identifying those risk factors associated with disease outcome can be difficult using traditional statistical methods as interactions should be selected a priori, and if attempting to evaluate all possible interactions, the number of terms grows exponentially requiring data with a sufficiently large number of observations to model all interactions and main effects. Logic regression, a decision tree method, naturally models interactions among binary predictors without a priori identification and can handle large numbers of variables. However, it is a weak learner in that small changes in the data can results in very different models. Logic forest and logicFS are ensemble adaptations of logic regression that combine multiple logic regression trees to identify important interactions and improve prediction performance. The primary focus of these methods has been on modeling binary outcomes, but for many diseases modeling continuous disease phenotypes is of interest. LogicFS can model continuous outcomes; however, it does not consider the complement of an interaction when assessing the importance of interactions, which is relevant for continuous outcomes. We adapt Logic Forest to model continuous outcomes and develop a quantitative interaction importance measure that accounts for complements of interac

Supervised dimension reduction using Bayesian Hierarchical Modeling: a simulation study and application to ambient air pollutants

Raymond Boaz
Department of Public Health Sciences
Medical University of South Carolina
boaz@musc.edu

Intro: Risk associated with air pollution has typically been evaluated at an individual pollutant level. Researchers understand that the presence of multiple pollutants may have interactive and grouped effects not currently captured in epidemiologic research. Utilizing health outcomes data, we have developed a novel mixture classification and modeling technique to characterize air pollution exposure in a more realistic context that accounts for the simultaneous and joint nature of the exposure. Methods: The model uses a method that informs the grouping of mixtures based on the health outcome of interest within a Bayesian Hierarchical Modeling framework. We are interested in modeling the relative risk as a function of air pollutant mixtures thought to affect the disease outcome. In addition, we observe confounder variables at the locations of the health outcomes. We have conducted simulation studies with ground truth scenarios consisting of mixtures of pollutants X impacting an outcome Y. The pollutants X have prespecified groupings with deterministic impact on the outcome Y, so the model parameters have been evaluated for fidelity to the prescribed relationship. We have also evaluated our model's accuracy and impact using previously developed simulation data sets from National Institute of Environmental Health Sciences. Results: Our model has successfully identified groupings of variables and qualitative effects to this point. We will continue to evaluate the accuracy ...

Threshold Multiple Confidence Interval Procedure with an Application on In-Season Batting Average Data

Taeho Kim
Department of Statistics
University of South Carolina
taeho@email.sc.edu

A thresholding approach in multiple confidence intervals (MCI) is studied. In order to set up a threshold, a hierarchical structure is adopted and the procedure is evaluated by ascertaining the global coverage probability and the average length. Under the normal-normal hierarchical setting, its global properties are investigated and used for optimization. In-season batting average data is used to demonstrate the procedure with real data. In general, the thresholding procedure helps to achieve a significant reduction in the average length while maintaining the acceptable level of global coverage probability. This work has been done under the guidance of Dr. Edsel Pena.

Evaluating Stopping Boundaries for Bayesian Multi-Arm Multi-Stage Design with Binary Endpoints

Zhenning Yu
Department of Public Health Sciences
Medical University of South Carolina
yuz@musc.edu

Multi-Arm Multi-Stage (MAMS) designs may reduce development costs and shorten the drug development timeline in clinical trials. MAMS includes prescheduled interim analysis so that treatment arms could be stopped early if they do not show sufficient promise (futility) or they have overt efficacy (efficacy). Therefore, we need to clearly define the stopping rules such as futility and efficacy boundaries at the beginning for the circumstance under which the trial will be stopped.

The same futility and efficacy boundaries are used across each stage in most MAMS design. However it may be more beneficial to set increasing futility boundary and decreasing efficacy boundary with respect to interim data, as what we did in Frequentist paradigm to control trial operating characteristics such as type I error and power. In this study, we use simulation to evaluate several operating characteristics including type I error, power, actual sample size and the probability for early termination, from a MAMS with adapting stopping rules and compare the results with those from equal boundaries.

Modeling Birth Outcomes and Food Security: A comparison of skew-normal and skew-t regression models

Carter Allen
Department of Public Health Sciences
Medical University of South Carolina
allecart@musc.edu

In many applications of classical linear regression, the distribution of residuals exibits non-normal qualities such as skewness or heavy tails, making the assumption of normal error terms difficult to justify. The common statistical suggestion in these cases is to implement a transformation of the response variable, but this can result in a loss of interpretability. The skew-elliptical family is a broad class of probability distributions that contain the normal distribution as a special case and allow for flexible modeling when data exhibit skewness. We examine the properties of skew-normal and skew-t models from both a Bayesian and frequentist perspective, and investigate the computational tools available for fitting these models. Finally, we apply skew-normal and skew-t models to data from the Nurture study, a cohort of mothers who gave birth between 2013 and 2016. Skewed-normal residuals are observed when modeling birth weight for gestational age z-score as a function of food security status during pregnancy in these data. The results of models under several different prior structures and using different available methods of estimation are compared with respect to the impact of food security during pregnancy on birth outcomes. We also extend these results to the multivariate case when modeling infant weight longitudinally over the first year of life.

Knot Specification in the Imputation of a Missing Longitudinal Variable

Virginia Shipes
Department of Public Health Sciences
Medical University of South Carolina
shipes@musc.edu

Biologic variables in clinical trials can be important prognostic variables for outcome prediction. These variables are often collected repeatedly over time, which increases the potential for missing data since data may not be collected at all necessary timepoints. Therefore, imputation of missing data is necessary. Fully characterizing the relationship of the patient and the biologic measurement has importance, as these measures could be useful for describing the ongoing condition of a patient. Splines of various knots are suggested as a method for imputation of missing measurements of a continuous, longitudinal variable as splines capture the functional component of the variable. Choosing the appropriate number of knots for a spline is instrumental to correctly imputing the missing values, as too few knots could cause the model to have a poor fit and too many knots could cause the model to over fit the data. Missing data from a continuous biologic variable were imputed using splines of 0 to 6 knots. Summary measures and their interactions were used as predictors in a multinomial model of an outcome measure. The interaction terms of the summary statistics became statistically significant after imputation. Some interaction terms were more dependent on the number of knots used during imputation, while others had more consistent results between imputation methods. This highlights the importance of knot specification in this type of imputation methodology.

Evaluation of Bayesian Multiple Stage Estimation under Spatial CAR Model Variants

Daniel Baer
Department of Public Health Sciences
Medical University of South Carolina
baerd@musc.edu

In this study, an evaluation of Bayesian hierarchical models is made based on simulation scenarios to compare single-stage and multi-stage Bayesian estimation. Simulated datasets of lung cancer disease counts for men aged 65 and older across 44 wards in the London Health Authority were analyzed using a range of spatially-structured random effect components. The goals of this study are to determine which of these single-stage models perform best given a certain simulating model, how estimation methods (single vs. multi-stage) compare in yielding posterior estimates of fixed effects in the presence of spatially-structured random effects, and finally which of two spatial prior models –the Leroux or ICAR model, perform best in a multi-stage context under different assumptions concerning spatial correlation. We found that among the single-stage models, the BYM model is robust to model misspecification. Further, we found that the multi-stage modeling process via the Leroux and ICAR models generally reduced the variance of the posterior estimated fixed effects. Finally, we found that the multi-stage Leroux model has attractive performance properties, and thus conclude that this multi-stage Leroux model should be seriously considered in applications of Bayesian disease mapping when an investigator desires to fit a model with both fixed effects and spatially-structured random effects to Poisson count data.

A New Approach to Regression Modeling in Group Testing that Accounts for the Dilution Effect

Stefani Mokalled
Department of Mathematical Sciences
Clemson University
smokall@g.clemson.edu

From screening for infectious diseases to detecting bioterrorism, group (pooled) testing of bio-specimen is a cost efficient alternative to individual level testing. Group testing has been utilized for both classification (identifying positive individuals) and estimation (fitting regression models using covariate measurements). A concern with the estimation process is the possible dilution of one individual's positive signal past an assay's threshold of detection. To account and correct for this dilution effect, we develop a new group testing regression model which explicitly acknowledges the effect. Unlike previous work in this area, this is accomplished by considering the continuous outcome that the assay measures, the individuals' latent biological marker (biomarker) levels, and the distributions of the biomarker levels of the cases and controls without requiring a priori knowledge of these distributions. We develop a novel mixture model and an expectation-maximization algorithm to complete model fitting. The performance of the methodology is evaluated through numerical studies and is illustrated using Hepatitis B data on Irish prisoners.

From mixed effects modeling to spike and slab variable selection: A Bayesian regression model for group testing data

Chase Joyner
Department of Mathematical Sciences
Clemson University
chasej@clemson.edu

Due to reductions in both time and cost, group (pool) testing is a popular alternative to individual level testing for disease screening. These reductions are obtained by testing pooled biospecimens (e.g., blood, urine, saliva, etc.) for the presence of an infectious agent. Though this process may reduce time and cost, it comes at the expense of data complexity, making the task of conducting disease surveillance more tenuous when compared to using individual level data. This is because an individual's disease status may be obscured by a group testing protocol and the effect of imperfect testing. Further, unlike individual level testing, a given participant could be involved in multiple testing outcomes and/or may never be tested individually. To circumvent these complexities and to incorporate all available data, we propose a Bayesian generalized linear mixed model that accommodates data arising from any group testing protocol, estimates unknown assay accuracies, and accounts for the potential heterogeneity in the covariate effects across population subgroups (e.g., clinic sites); this latter feature being of key interest to practitioners tasked with conducting disease surveillance. To achieve model selection, our proposal uses spike and slab priors for both fixed and random effects. The developed methodology is illustrated through extensive numerical studies and is applied to chlamydia surveillance data.

Computer model calibration as a method for design

Carl Ehrett
Department of Mathematical Sciences
Clemson University
cehrett@clemson.edu

A well-established Bayesian framework for the calibration of computer models estimates unknown and/or uncontrolled parameters by attuning the model to data obtained through physical experimentation. In this poster I reconceptualize the framework for model calibration as a method for optimization. That is: rather than calibrating a model to find a posterior distribution on unknown parameters in order to bring the model into agreement with reality, I instead calibrate to find a posterior distribution on controllable model inputs in order to bring the model into agreement with predetermined performance targets. In essence, I treat performance targets as ``desired observations'' which are deployed in calibration where in traditional calibration one would deploy observations obtained through physical experimentation. The methodology incorporates uncertainty quantification. I demonstrate the methodology in an artificial case (without a code surrogate) and in the case of a finite element model of wind turbine blade performance and cost (using a Gaussian process code surrogate).

A Bayesian Multidimensional Trend Filter

Stella Self
Department of Mathematical Sciences
Clemson University
stellaw@clemson.edu

The Bayesian multidimensional trend filter is an extension of the one-dimensional trend filtering technique popular in time series analysis. The method is suitable for data collected over 2 or more dimensions, such as spatial or spatio-temporal data. The multidimensional trend filter estimates a smooth trend function over the entire support space. The technique is computationally tractable, even for a large number of observations or a large support space. Furthermore, the computational expense of the method is determined by a user-specified level of discretization and is nearly independent of the number of observations. An adaptive method of performing the discretization of the support space is developed.