Dissertation Paper Abstracts

Ilana Trumble

May 21, 2019

**Paper 1:** Quasilexicographic method

Title: Unbiased, consistent estimation and size inference for longitudinal and multilevel models with missing data

Longitudinal and multilevel studies are frequently subject to missing data. Current approaches, such as mixed models and generalized estimating equations (GEE), are known to have convergence problems and inflated type I error rates when applied to typical longitudinal or multilevel data and unstructured covariance matrices. In this paper, we provide a novel method for analyzing such data without these limitations. We name this the *quasilexicographic* method. The quasilexicographic method can be applied to data with missingness and mistimed outcomes without inflation of type I error rates. We define quasilexicographic model parameter estimates and present derivations for their distributions for continuous outcomes and non-missing predictors. We show that these parameter estimates are unbiased and asymptotically efficient. Additionally, we present a test statistic for testing the general linear hypothesis, and derive its corresponding distribution. Simulations studies confirm that the expected nominal Type I error rate for the general linear hypothesis test statistic are met. Simulation studies are also used to compare the performance of the quasilexicographic model, mixed model, and GEE. The utility of this method is demonstrated for an epigenetics study examining the relationship between differentially methylated positions (DMPs) in fetal cord blood and the subsequent development of allergic disease.

**Paper 2:** Where do the quasilexicographic model, mixed model, and GEE fail?

Title: Analysis methods for longitudinal and multilevel studies with Gaussian outcomes and data that is missing completely at random

Current approaches for analyzing multilevel and longitudinal data, such as mixed models and generalized estimating equations (GEE), have notable limitations. Specifically, these models can fail to converge, can yield biased parameter estimates, and can produce inflated Type I error rates that can subsequently impact statistical inference. We have previously introduced a novel approach called the quasilexicographic method which overcomes these limitations in certain cases. In this paper, we use simulation studies to identify specific situations and criteria for which limitations in convergence, estimation, and inference exist for the quasilexicographic method, mixed models, and GEE for data that is missing completely at random (MCAR). Power is also compared between the three methods. Sample size and missing data are investigated as factors relating to model failure.

**Paper 3:** Tutorial paper applied to epigenetics study

Title: Tutorial on analyzing an epigenetics study with correlated outcomes

Previous studies have shown that maternal diet during pregnancy can effect epigenetic methylation in fetal cord blood. These epigenetic changes have been further shown to impact the development of allergic disease in the child later in life. In this paper, we provide a tutorial for analyzing the association between fetal blood differentially methylation positions (DMPs) and cytokines, which are released during the immune response to allergens. Current approaches for analyzing this type of data do not account for the correlation between cytokines. We prove that when this correlation is ignored, power is lost. Instead, we propose that longitudinal and multilevel analysis methods be applied to this data. Specifically, we present a tutorial for applying our newly introduced quasilexicographic method to this type of data, which accounts for the correlation between cytokines. We present an alpha-spending approach, which distributes the type I error over the duration of sequential tests. We first test the overall hypothesis that there is no association between cytokines and DMPs. If this test is significant, we then step down and test each DMP individually. It is known which DMPs are related to cytokines, and so we are able to compare our results with what is known in the scientific literature. This approach can be applied to many big-data epigenetic studies.