a subgraph. This is a joint work with Mathias Drton and Dennis Leung at University of Washington.

A new class of copulas involved geometric distribution: estimation and applications

Kong-Sheng Zhang
Department of Mathematics, Southeast University, China
E-mail: zks155@163.com

Keywords: Copula; geometric distribution; maximum likelihood estimation; interior-point penalty function method

Abstract: Copula is becoming a popular tool for modelling the dependence structure among multiple variables. Commonly used copulas are Gaussian, t and Gumbel copulas. To further generalize these copulas, a new class of copulas, referred to as geometric copulas, is introduced by adding geometric distribution into the existing copulas. The interior-point penalty function algorithm is proposed to obtain maximum likelihood estimation of the parameters of geometric copulas. Simulation studies are carried out to evaluate the efficiency of the proposed method. The proposed estimation method is illustrated with workers' compensation insurance data and exchange rate series data.

CS04 MODELS FOR COMPLEX BIOLOGICAL DATA
Session Chair: Sanjay Chaudhuri, National University of Singapore, Singapore
Venue: Seminar Room 4
Time: 17 Dec, 13:40-15:40

Using robust variance estimation in mixed models: a review

A.A.Sunethra, M.R. Sooriyarachchi
University of Colombo, Sri Lanka
Email: sunethra@stat.cmb.ac.lk, roshinis@hotmail.com

Keywords: Correlated Data, Mixed Models, Robust Variance, Random Effects, Sandwich Variance Estimation

Abstract**:** Presence of Clusters/ sub-groups within datasets is a common phenomenon in statistical data analysis. Examples include repeated measures data, longitudinal data, hierarchical data, and etc. The shared feature in such datasets is that observations within a group are related / similar to each other. Data of this kind is termed as correlated data or non-independent data. When analyzing such data, the methods of analysis should

not rely on the assumption of independence which is a dominant assumption in statistics.

Robust Variance Estimation which is often nicknamed as Sandwich Variance Estimation (SVE) is a method of variance estimation initially proposed by Peter J Huber in 1967 to correct the estimation of standard errors of miss-specified models, i.e. in models that are being fitted incorrectly. These miss-specifications/errors may be due to various reasons such as incorrect distributional assumptions, assuming linear relationships for non-linear data, assuming independency for correlated data and etc. This methods gained more popularity with its derivation in linear regression by H. White in 1980 where he demonstrate its usage for independent, heteroscedastic errors in linear regression models where the miss-specifications was not due to independence but due to errors being heteroscedastic. In contrast, with correlated data modeling, SVE requires to cater for heteroscedatsic, non-independent data. Hence, SVE is being a method for adjusting the standard errors of model parameters; it had been extensively used in correlated data analysis for obtaining standard errors that are adjusted to the correlation of the data where the adjustment made by SVE doesn't rely on the model being fitted to the data. In olden days, when statistical models were not developed for correlated data, models assuming independence were fitted for non-independent data and the model standard errors were adjusted by using SVE. The literature had emphasized that SVE has provided improved inferential results in correlated data analysis in the absence of statistical models for correlated by improving the functionality of independency assumed models fitted for correlated data. In addition to the classical SVE, various adjustments for the classical SVE had been developed for various data scenarios such as small sample data, data with auto-correlation and etc.

Lately, specialized statistical models were developed for correlated data such as Mixed Models and Generalized Linear Mixed Models (GLMMs). Since these models are defined for correlated data, the model parameter estimates and standard error estimates are resultant to the correlation exist in the data. Therefore, the necessity of SVE in such models was at argument by authors in the literature. Mixed models are defined in such a way that clusters/groups that impose correlation to the data is being introduced to the model as random effects that follow a particular statistical distribution (Gaussian, Gamma, t-distribution) where the linear predictor of mixed models consists of a component that represent the grouping/clustering in the data. More over the literature consists of few authors that had demonstrated probable miss-specifications of Mixed Models despite they are defined for correlated data. These miss-specifications are mainly due to the disparity between the correlation structure of the data and the way the random effects are defined in Mixed Models. Upon the identification of miss-specifications of such hybrid models, adoption of SVE in GLMMs becomes remedial since SVE is meant for improving miss-specified models. Though SVE was initially proposed for correcting the standard errors of maximum likelihood estimates, it can be used for parameter estimation methods which obtain parameter estimates by equating the estimation function to zero which doesn't necessarily be a derivative of a log-likelihood. Thus, SVE are feasible with Mixed Models hich mostly accommodate pseudo likelihood methods in parameter estimation.

Researches or studies which have looked at the use of SVE particularly in Mixed Models or GLMMs are very few where a research comparing the use of SVE in GLMMs for analyzing two actual datasets was found which showed up evidence for SVE is being capable of correctly estimating the variance of the fixed effects parameters of GLMMs even when random effects are miss-specified. Researches that had highlighted miss-specifications of GLMMs mainly had exposed the errors of random effects definition of GLMMs not being able to represent precisely the correlation structure present in the data. Therefore, modifying random effect definition can be considered as direct solution for this issue whereas SVE serves indirectly by improving the standard error estimation of Mixed Models. Since SVE can improve the estimation of model standard errors, it improves model adequacy tests and other hypothesis test associated with Mixed Models. Simulating correlated data scenarios with probable miss-specifications in par with Mixed Models' random effect definition and then analyzing those data using suitable Mixed Models while using SVE can be used for evaluating feasibility of using SVE in Mixed Models. Further, the sample size and the level of the correlation present in the data could also impact on the performance of Mixed Models. The development of various adjustments for the classical SVE had mainly taken place for coping up with various correlation structures and for dealing with small sample size where SVE was earlier considered as an asymptotical method which works well for large sample sizes. Therefore, the choice of the SVE adopted should carefully be made with respect to the correlation structure present in the data and with respect to the sample size of the data. It was identified through simulation that Mixed Models with SVE assist on enhancing its functionality than used with standard method of variance estimation while at small sample sizes enhancements can be achieved by using small sample adjusted SVEs. In summary, it can be emphasized though SVE is being a method of variance estimation developed nearly about half a century before, its applicability still resides even with hybrid statistical models like Mixed Model or GLMMs which are very recently developed statistical modeling approaches for correlated data.

Influence analysis of area under ROC curve

Bo-Shiang Ke
National Chiao Tung University, Taiwan

Keywords: AUC, influence function, local influence, cumulative lift chart

This work is collaborated with Yuan-chin Ivan Chang at Institute of Statistical Science, Academia Sinica.

Abstract: Supervised learning is a major issue in statistical learning, especially binary classification problems. Enormous techniques are created to deal with them. In order to quantify the performances of classifiers, an objective performance criterion is indispensable. Area under ROC curve (AUC) is a popular performance measure due to its elegant interpretation; however, potential influential observations may alter its conclusion. To this end, we first