Statistical Analysis and Applications IT

J. Anil Kumar,
Assistant Professor,
Department of Management Studies,
KMM ITS,

S. Ismail Basha,
Assistant Professor,
KMM Institute of Technology &
Science,Department of Management Studies,

***Abstract*- Software development is often perceived as a unique activity, residing at the boundary of the technical world and the world of the developer's individuality. To fully understand the phenomena occurring during the software development, a researcher should address the technical, as well as "people problems" associated with the development of software.**

We develop various statistical methods important for multidimensional genetic data analysis. Theorems justifying application of these methods are established. We concentrate on the multifactor dimensionality reduction, logic regression, random forests, stochastic gradient boosting along with their new modifications. We use complementary approaches to study the risk of complex diseases such as cardiovascular ones. The roles of certain combinations of single nucleotide polymorphisms and non-genetic risk factors are examined. To perform the data analysis concerning the coronary heart disease and myocardial infarction the Lomonosov Moscow State University supercomputer "Chebyshev" was employed.

Keywords - Introduction, Statistical Methods of SNP, Applications

I. INTRODUCTION

Foundations of Statistical Analyses and Applications

The analysis of real data by means of statistical methods with the aid of a software package common in industry and administration will certainly be part of a future professional work of many students in mathematics or mathematical statistics.

Commonly there is no natural place in a traditional curriculum for mathematics or statistics, where a bridge between theory and practice fits into. On the other hand, the demand for an education designed to supplement theoretical training by practial experience has been rapidly increasing.

There exists, consequently, a bit of a dichotomy between theoretical and applied statistics, and this book tries to straddle that gap. It links up the theory of a selection of statistical procedures used in general practice with their application to real world data sets using the statistical software package SAS (Statistical Analysis System). These applications are intended to illustrate the theory and to provide, simultaneously, the ability to use the knowledge effectively and readily in execution.

An introduction to SAS is given in an appendix of the book. Eight chapters present theory, sample data and SAS realization to topics such as regression analysis, categorial data analysis, analysis of variance, discriminant analysis, cluster analysis and principal components.

Statistical analysis is a component of data analytics. In the context of business intelligence (BI), statistical analysis involves collecting and scrutinizing every single data sample in a set of items from which samples can be drawn.

Statistical analysis can be broken down into five discrete steps, as follows:

- Describe the nature of the data to be analyzed.
- Explore the relation of the data to the underlying population.
- Create a model to summarize understanding of how the data relates to the underlying population.
- Prove (or disprove) the validity of the model.
- Employ predictive analytics to run scenarios that will help guide future actions.

The goal of statistical analysis is to identify trends. A retail business, for example, might use statistical analysis to find patterns in unstructured and semi-structured customer data that can be used to create a more positive customer experience and increase sales.

An accessible introduction to performing meta-analysis across various areas of research

The practice of meta-analysis allows researchers to obtain findings from various studies and compile them to verify and form one overall conclusion. Statistical Meta-Analysis with Applications presents the necessary statistical methodologies that allow readers to tackle the four main stages of meta-analysis: problem formulation, data collection, data evaluation, and data analysis and interpretation. Combining the authors' expertise on the topic with a wealth of up-to-date information, this book successfully introduces the essential statistical practices for making thorough and accurate discoveries across a wide array of diverse fields, such as business, public health, biostatistics, and environmental studies.

Two main types of statistical analysis serve as the foundation of the methods and techniques: combining tests of effect size and combining estimates of effect size. Additional topics covered include:

- Meta-analysis regression procedures
- Multiple-endpoint and multiple-treatment studies
- The Bayesian approach to meta-analysis
- Publication bias

Statistical Methods of SNP Data Analysis and Applications

Various statistical methods important for genetic analysis are considered and developed. Namely, we concentrate on the multifactor dimensionality reduction, logic regression, random forests and stochastic gradient boosting. These methods and their new modifications, e.g., the MDR method with ”independent rule”, are used to study the risk of complex diseases such as cardiovascular ones. The roles of certain combinations of single nucleotide polymorphisms and external risk factors are examined. To perform the data analysis concerning the ischemic heart disease and myocardial infarction the supercomputer SKIF ”Chebyshev” of the Lomonosov Moscow State University was employed.We believe that it is appropriate to extend the research tools commonly used in engineering and computer science with those applied in the sciences oriented at studying people. We recognize that the research tools used in psychology and medicine may not be fully applicable and not even completely transferable to the field of software engineering. However, we believe that, in the combination with other research tools, they could help our understanding of the problems associated with software development. In this paper we present one of our first attempts to apply one of such methodologies to the software measures data. We applied the technique of "weighted estimators of common correlation" to 100 Java public domain projects. The results showed that the weighted estimators technique produces high correlation between data where relationship does exist. The technique does not introduce spurious correlation between data that does not have any relationship

Applications of the weighted estimator's technique to software engineering data

We now apply the above approach to a number of samples of software engineering data. We have data for 100 public domain software engineering projects written in Java. These projects have already been analyzed individually using a standard correlation analyses between many
Attributes of interest. We now take three of the relationships that we have studied and present them here as inputs to a meta-analytical study using the weighted estimators of a common correlation approach. The first two relations are meant to demonstrate the accuracy of the meta-analytical method in finding relationships between data that is already known to have a
relationship, thus avoiding a type I error. These relationships are the cyclomatic complexity versus the lines of code and the lines of code versus the weighted method count (WMC) The last relationship is meant to show that the technique also avoids a type II error, that is it does not detect spurious correlations between unrelated data.

CONCLUSION

The results of our experiment in meta-analysis of software measurement data indicates that such techniques may be applicable to the research in software engineering, despite the problem of the data most likely not being homogenous However, research in other fields indicates that is unlikely to be the case. Meta-analytic techniques are widely used statistical tools in other disciplines, especially in psychology and medicine. Due to the peculiar nature of software development, researchers in this field have to deal with the technological and behavioral problems. We believe that the applicability of statistical meta-analysis to empirical software engineering should be further investigated in search for the limits of their application.

We tested the applicability of the "weighted estimator of a common correlation" technique to software measurement data. The data came from the public domain and not necessarily from the same development environment. Our results show that this technique avoids both, type I and type II errors. In the light of the published research on the applicability of meta-analysis to software measurement data we conclude that further analysis is needed to provide guidance and determine the limitations in applying meta-analytic techniques to empirical software engineering research.

REFERENCES

[1] Y. Fujikoshi, R. Shimizu and V. V. Ulyanov, “Multivariate Statistics: High-Dimensional and Large-Sample Approximations,” Wiley, Hoboken, 2010.

[2] S. Szymczak, J. Biernacka, H. Cordell, O. González- Recio, I. K?nig, H. Zhang and Y. Sun, “Machine Learning in Genome-Wide Association Studies,” Genetic Epidemiology, Vol. 33, No. S1, 2009, pp. 51-57. doi:10.1002/gepi.20473

[3] D. Brinza, M. Schultz, G. Tesler and V. Bafna, “RAPID Detection of Gene-Gene Interaction in Genome-Wide Association Studies”, Bioinformatics, Vol. 26, No. 22, 2010, pp. 2856-2862. doi:10.1093/bioinformatics/btq529

[4] K. Wang, S. P. Dickson, C. A. Stolle, I. D. Krantz, D. B. Goldstein and H. Hakonarson, “Interpretation of Association Signals and Identification of Causal Variants from Genome-Wide Association Studies,” The American Journal of Human Genetics, Vol. 86, No. 5, 2010, pp. 730-742. doi:10.1016/j.ajhg.2010.04.003

[5] Y. Liang and A. Kelemen. “Statistical Advances and Challenges for Analyzing Correlated High Dimensional SNP Data in Genomic Study for Complex Diseases,” Statistics Surveys, Vol. 2, No. 1, 2008, pp. 43-60. doi: 10.1214/07-SS026

[6] H. Schwender and I. Ruczinski, "Testing SNPs and Sets of SNPs for Importance in Association Studies," Bio- statistics, Vol. 12, No. 1, 2011, pp. 18-32. doi: 10.1093/biostatistics/kxq042

[7] M. Ritchie, L. Hahn, N. Roodi, R. Bailey, W. Dupont, F. Parl and J. Moore, "Multifactor-Dimensionality Red- uction Reveals High-Order Interactions Among Estrogen- Metabolism Genes in Sporadic Breast Cancer," The American Journal of Human Genetics, Vol. 69, No. 1, 2001, pp. 138-147. doi:10.1086/321276

[8] D. Velez, B. White, A. Motsinger, W. Bush, M. Ritchie, S. Williams and J. Moore, "A Balanced Accuracy Fun- ction for Epistasis Modeling in Imbalanced Datasets Using Multifactor Dimensionality Reduction," Genetic Epidemiology, Vol. 31, No. 4, 2007, pp. 306-315. doi: 10.1002/gepi.20211

[9] I. Ruczinski, C. Kooperberg and M. LeBlanc, "Logic Regression," Journal of Computational and Graphical Statiststics, Vol. 12, No. 3, 2003, pp. 475-511. doi: 10.1198/1061860032238

[10] H. Schwender and K. Ickstadt, "Identification of SNP Interactions Using Logic Regression," Biostatistics, Vol. 9, No. 1, 2008, pp. 187-198. doi: 10.1093/biostatistics/kxm024

[11] L. Breiman, "Random Forests," Machine Learning, Vol. 45, No. 1, 2001, pp. 5-32. doi:10.1023/A:1010933404324

[12] J. Friedman, "Stochastic Gradient Boosting," Computational Statistics & Data analysis, Vol. 38, No. 4, 2002, pp. 367-378.

[13] X. Wan, C. Yang, Q. Yang, H. Xue, N. Tang and W. Yu, "Mega SNP Hunter: A Learning Approach to Detect Disease Predisposition SNPs and High Level Interactions in Genome Wide Association Study," BMC Bioinformatics, Vol. 10, 2009, p. 13. doi:10.1186/1471-2105-10-13

[14] A. Bulinski, O. Butkovsky, A. Shashkin, P. Yaskov, M. Atroshchenko and A. Khaplanov, "Statistical Methods of SNPs Analysis," Technical Report, 2010, pp. 1-159 (in Russian).

[15] G. Bradley-Smith, S. Hope, H. V. Firth and J. A. Hurst, "Oxford Handbook of Genetics," Oxford University Press, New York, 2010.

[16] S. Winham, A. Slater and A. Motsinger-Reif, "A Comparison of Internal Validation Techniques for Multi- factor Dimensionality Reduction," BMC Bioinformatics, Vol. 11, 2010, p. 394. doi:10.1186/1471-2105-11-394

[17] A. Arlot and A. Celisse, "A Survey of Cross-validation Procedures for Model Selection," Statistics Surveys, Vol. 4, No. 1, 2010, pp. 40-79. doi:10.1214/09-SS054

[18] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2nd Edition, Springer, New York, 2009.

[19] R. L. Taylor and T.-C. Hu, "Strong Laws of Large Numbers for Arrays of Rowwise Independent Random Elements," International Journal of Mathematics and Mathematical Sciences, Vol. 10, No. 4, 1987, pp. 805-814.

[20] E. Lehmann and J. Romano, "Testing Statistical Hypo- theses," Springer, New York, 2005.