THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

Topics for STAT3799 Directed Studies in Statistics (6 credits)
(Offered in both 1st and 2nd semesters of 2018 – 2019 for STAT3799)

1. Mortality projection and longevity risk

Life insurance companies face different types of risks. Within the life annuity business we may find what is called 'longevity risk', which refers to the possibility that annuitants live longer than expected according to the life tables used for pricing, determination of benefits and technical reserves.

This situation represents a threat to life annuity business, and therefore we need to rely on projected life tables that account for the improvement in mortality, a fact that has been observed since the second half of the 20th Century in most developed countries.

The student who takes this project is expected to study the most widespread models in the literature for mortality projection in order to mitigate this type of risk.

Requirement: STAT3901 and STAT3909.

Supervisor: **Dr. A. Benchimol**, benchi@hku.hk, Dept of Statistics and Actuarial Science

2. Copulas in risk management

Copulas are functions that join multivariate distribution functions to their one-dimensional marginal distribution functions. The student who takes this project is expected to study the basic theory of copula and some of its applications in risk management. All the related literature will be provided.

Supervisor: **Dr. K.C. Cheung**, kccg@hku.hk, Dept of Statistics and Actuarial Science

3. Data Mining for streaming data

Nowadays, the volume of data is growing in an unprecedented rate. It is getting importance to have a data mining algorithm that can process and analyze data online, in a one pass streaming setting. I would like to have a comprehensive review on the latest data mining techniques adaptive to the online data and their effectiveness in practical applications.

Supervisor: **Dr. Simon K.C. Cheung,** simonkc@hku.hk, Dept of Statistics and Actuarial Science

4. Familial database search on two-person DNA mixtures using peak area information

For crime cases in which no suspect can be identified based on non-DNA evidences such as fingerprints or witness reports, the police force may search a database of DNA profiles from previously convicted criminals or unsolved crime cases. If no offender profile in the database perfectly matches the crime trace, an additional search can be performed, hoping that an individual in the database is a close relative of the perpetrator and can be identified through the search. The role of familial database search as a crime-solving tool has been increasingly recognized by forensic scientists. In this project, student will study the basic knowledge of DNA fingerprinting, DNA database search, methodologies on resolving two-person DNA mixtures based on peak area information; and explore the strategies of forensic investigations from familial database search results.

Requirement: Knowledge of course STAT3608. Strong knowledge in programming language like C++ or R and computational algorithms such as MCMC would be essential.

Supervisor: **Dr. Y.K. Chung**, yukchung@hku.hk, Dept of Statistics and Actuarial Science

5. Test for Parent-of-origin effects on the X-chromosome

Genomic imprinting is an important epigenetic factor in complex traits study, which has generally been examined by testing for parent-of-origin effects of alleles. In this project, we shall work on the detection of parent-of-origin effects on the X-chromosome, using the parental-asymmetry test based on case-parents trios and/or case-parent pairs data. Various other statistical tests would also be considered. The student is expected to write computer programs in R.

Supervisor: **Prof. Tony W.K. Fung**, wingfung@hku.hk, Dept of Statistics and Actuarial Science

6. Measurement Error Problem

The student will be asked to explore parametric, semi-parametric and non-parametric methods when the variables are measured with errors. The students will conduct numerical comparisons among various density estimation and regression methods. Hence, strong computational skill are required. High dimensional measurement error problem will be studied, and non-convex programming will be discussed.

Supervisor: **Dr. F. Jiang**, feijiang@hku.hk, Dept of Statistics and Actuarial Science

7. Latent Class Analysis

Latent class analysis is a method for analyzing the relationships among manifest data when some variables are unobserved. The unobserved variables are categorical, allowing the original dataset to be segmented into a number of exclusive and exhaustive subsets. This project is to explore some latent class analysis methods. It includes the evaluation of some existing algorithm and models as well as the determination of the number of latent classes and variable set.

Requirement: Knowledge of some computer programming languages is essential.

Supervisor: **Dr. C.W. Kwan**, cwkwan@hku.hk, Dept of Statistics and Actuarial Science

8. Analysis of correlated zero-inflated count data

In many medical and public health investigations, the count data encountered often exhibit an excess of zeros, and very frequently this type of data are collected on clusters of subjects or by repeated measurements on each subject. For example, in the analysis of medical expenditure, members in the same family may exhibit some correlation possibly due to housing locality, genetic predisposition, similar dietary and living habit. Ignoring such correlation may lead to misleading statistical inference. This project will survey the models and methods in the literature and apply them to a real data set.

Requirement: Knowledge in programming language like FORTRAN or C++.

Supervisor: **Dr. Eddy K.F. Lam**, hrntlkf@hku.hk, Dept of Statistics and Actuarial Science

9. Applications of extreme value models

Extreme value theory concerns the behaviour of maxima or minima, and has been used extensively in areas such as finance, hydrology, engineering and meteorology where the occurrence of extremes may have catastrophic consequences. In this project, the student will learn the basic modelling techniques for data of extremes and will apply such models to data sets of practical interest. The emphasis is on conceptual understanding of the underlying theory and interpretation of the fitted models.

Requirement: The student should be competent in computer programming. Knowledge in or willingness to learn the R programming language is essential.

Supervisor: **Dr. David Lee**, leedav@hku.hk, Dept of Statistics and Actuarial Science

10. Inference about ordered binomial probabilities

Consider the following hypothetical example. Two groups of candidates apply for graduate studies at a renowned university. The two groups are fairly homogeneous except that the first group have a professional qualification but the second group do not. We are interested in the success rates of the two groups in getting offers from the university. At first sight this seems like a rather simple two-sample problem concerning two binomial probabilities, one for each group. However, common sense tells us that the first group should have a success rate at least not smaller than that of the second group, so that the two binomial probabilities are actually "ordered" in a known manner. With such prior knowledge, standard statistical inference about binomial probabilities suddenly becomes not so standard. This project investigates the issues involved in the above problem setting, and studies different plausible solutions to the problem.

Supervisor: **Prof. Stephen M.S. Lee**, smslee@hku.hk, Dept of Statistics and Actuarial Science

11. Security design in blockchain architecture

In this project we begin with a review of the basic architecture for blockchain in Python. This includes state transition rules, method for creating blocks, mechanisms for checking the validity of transactions, blocks, and the full chain. Next, we will create new blocks from data, validate the new blocks and add them to the existing blockchain.

Security is of the utmost importance in any blockchain architecture, in this project we will discuss 3 popular verification methods: public key cryptography, digital signature algorithm and trusted time-stamping. Despite the advanced level of technical sophistication, we will construct practical examples wherever possible.

Supervisor: **Dr. Eric A.L. Li**, ericli11@hku.hk, Dept of Statistics and Actuarial Science

12. Bootstrap approximation in time series modeling

The traditional time series modeling and further inference are based on the normality assumption or large enough sample size. In the real applications, the normality may be broken and the results may not be accurate for the moderate or small sample sizes. The bootstrap is a computer-intensive method, and the information in the real data is repeatedly used. Hence it may provide more accurate results. This project hopefully can train students for some bootstrap methods to dependent data, and some knowledge of computer languages such as FORTRAN or C is required since a little more computation will be involved.

Supervisor: **Dr. G.D. Li**, gdli@hku.hk, Dept of Statistics and Actuarial Science

13. DNA Sequencing Data

DNA sequencing data holds the promise of identifying causal rare variants associated with human traits and diseases with genetic background. In this project, we shall work on the detection of rare variants using next generation DNA sequencing data. The students are expected to have basic knowledge of kernel machine regression and statistical genetics, and should be strong in programming.

Supervisor: **Dr. Z. Liu**, zhhliu@hku.hk, Dept of Statistics and Actuarial Science

14. Sentiment Analysis of Corpora with Mixed Languages

Sentiments of movie reviews, product reviews and customer comments provide important information to business leaders for their decision making. Traditionally, the sentiments are classified by the counting of terms with positive and negative polarities. This lexicon approach does not take the semantic structure of textual data into account of sentiment analytics. In this project, students are required to analyze the sentiments of opinion data which could be in English language or any other languages. Consequently, students can understand how the semantic structure of textual data can play an important role in sentiment classification.

Requirement: Knowledge of R/python programming is required for this project.

Supervisor: **Dr. Gilbert C.S. Lui,** csglui@hku.hk, Dept of Statistics and Actuarial Science

15. Cointegration in financial analysis

The goal of this project is to test cointegration in financial time series. Students are required to have basic understanding of cointegration and some knowledge of computer programming.

Supervisor: **Dr. C. Wang**, stacw@hku.hk, Dept of Statistics and Actuarial Science

16. A Statistical Study on Financial Market Anomalies

Market anomalies can refer to strange patterns in financial data which violate the efficient market hypothesis (EMH). Some famous market anomalies include weekend effect, January effect and size effect. Traders using technical analysis and trading strategies may earn abnormal profits from market inefficiency. This project aims to study various market anomalies based on statistical analysis. Investigations should be made on the existence or significance of the effects of any market anomalies in various financial markets.

Students taking this project are expected to study the relevant literature and conduct statistical tests using real market data. Elementary programming skills may be required to process large amount of

data.

Supervisor: **Dr. K.P. Wat**, watkp@hku.hk, Dept of Statistics and Actuarial Science

17. Investigation of Non-normality in a Simple Errors-in-variables Model

In a classical linear regression model, it is usually assumed that the predictive variable is not subject to any kind of random error. However, it is not always true in many applications. In addition, it is also a common practice to assume that the error in the regression model is normally distributed. Unfortunately, we may often find that most real data sets do not really exhibit such nice properties. In this project, student will investigate the non-normality situation where the errors in a regression model exist. Computer programming skill is required.

Requirement: Strong knowledge in computer programming and statistical simulation technique is a must.

Supervisor: **Dr. Raymond W.L. Wong**, rwong@hku.hk, Dept of Statistics & Actuarial Science

18. Deep Learning with Time-To-Event Data

Students will study the state of the art machine learning methods for time-to-event data and apply them to real financial and biomedical datasets. To appreciate the power and impact of deep learning in real applications in Finance and Medicine, it is of great interest to implement various machine learning methods in building up predictive models and assessing their practical performance. Students will survey an extensive literature and conduct extensive numerical studies in the investigation. The passion for programming and dedication to making impact on practice will be greatly appreciated.

Requirement: R (and Phython) programming.

Supervisor: **Dr. J.F. Xu**, xujf@hku.hk, Dept of Statistics & Actuarial Science

19. Change Measure: Survey and Applications

Change measure is a useful and powerful tool in a number of areas, including mathematical finance, actuarial science and probability theory. In this course, we will study various change measure techniques, such as Girsanov theorem, Esscher transform, and applications in option pricing, premium calculation. The student needs to have some probability background in order to take this course. An advanced probability course (with some measure theory included) is preferred.

Supervisor: **Prof. H.L. Yang**, hlyang@hku.hk, Dept of Statistics & Actuarial Science

20. Wishart matrix, eigenvalue distribution and the Marchenko-Pastur law

Wishart matrix is a matrix model for sample covariance matrix from a multivariate normal distribution. It has a long history and many interesting results exist for its eigenvalues and eigenvectors. When the dimension increases to infinity, the empirical distribution of the eigenvalues converges to the celebrated Marchenko-Pastur law.

In this project, students will learn some basis theory on Wishart matrix and multivariate normal distributions. Some techniques from random matrix theory will be needed to derive the Marchenko-Pastur distribution. A good command of multivariate analysis and matrix algebra is required.

Supervisor: **Prof. Jeff J.F. Yao**, jeffyao@hku.hk, Dept of Statistics & Actuarial Science

21. Deep learning with application in artificial intelligence

This project will focus on extracting useful information from structured and unstructured data and formulating statistical models for inference and prediction. In particular, we will develop deep learning, including deep neural networks for imaging analysis and computer vision and natural language processing for text data analysis. Extensive computation will be needed and real data will be used for analysis and illustration.

Requirement: The student needs to have experience with Python and R programming.

Supervisor: **Prof. G.S. Yin**, gyin@hku.hk, Dept of Statistics and Actuarial Science

22. Boosting for Ranking Predictions

Ranking data are often encountered in practice when individuals are asked to rank a set of items. We see examples in university rankings, journal rankings and gene rankings just to name a few. A typical task of studying ranking data is to predict the ranking of the items based on their profiles. In this project, we will study some boosting algorithms for ranking data and apply them to a few real ranking data sets.

Supervisor: **Dr. Philip L.H. Yu**, plhyu@hku.hk, Dept of Statistics and Actuarial Science

23. Insurance Risk Models with Dependent Risks

In classical risk theory, the assumption of independence in the study of the surplus process of an insurance company plays an important role. Since this assumption is rather restrictive and unrealistic, insurance risk models with dependent risks have been studied extensively in the past few decades. In this project, a number of these models will be discussed. In particular, for each

of these models, numerical and simulation studies will be carried out to assess the impact of the dependence structure on some actuarial quantities related to ruin.

Supervisor: **Prof. K.C. Yuen**, kcyuen@hku.hk, Dept of Statistics and Actuarial Science

24. Statistical Analysis of Large-scale Educational Assessment Datasets

In this project, we will explore international large-scale assessment datasets such as PISA, TIMSS and PIRLS with data manipulation and visualization techniques, then perform statistical analysis and predictive modeling. In particular, machine learning approaches will be investigated and compared to traditional modeling techniques such as item response theory and latent variable models. Proficient programming skills with R and/or Python are required.

Supervisor: **Dr. A.J. Zhang**, ajzhang@hku.hk, Dept of Statistics and Actuarial Science

25. Forecasting time series: with application to stocks trading

This project aims to forecast forward behavior of stock prices using either of the following techniques: threshold time series models, neural networks, random decision forests, support vector. Simulated trading strategies based on the forecast results are also required.

Requirement: Knowledge of course STAT3612 or STAT8017, and skills in statistical programming using either SAS, R, or C++.

Supervisor: **Dr. Z.Q. Zhang**, zhangz08@hku.hk, Dept of Statistics and Actuarial Science

26. Non-linear time series analysis

Non-linear time series models have achieved a great success in real applications. This project aims to give a study on the modelling and statistical inference of many non-linear time series models such as threshold AR, GARCH, and their variants. Students are expected to use these methodologies to analyze real data sets.

Requirement: Statistics and Matlab.

Supervisor: **Dr. K. Zhu**, mazhuke@hku.hk, Dept of Statistics and Actuarial Science

********** END **********