UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

Topics for STAT3819 Project in Statistics and Actuarial Science (6 credits)
(2013 – 2014)

1. Introduction to the Gerber-Shiu function in ruin theory

The Gerber-Shiu function introduced by Gerber and Shiu (1998) unifies and generalizes the study of various ruin-related quantities in ruin theory, including the time of ruin, surplus prior to ruin and the deficit at ruin of the insurance company. In this project, the student will learn the nice analytic properties of the Gerber-Shiu function as well as its applications. Recent advances and generalizations of Gerber-Shiu function will also be discussed. Strong knowledge in stochastic processes and computational skills will be required. The student is assumed to be familiar with software packages such as Mathematica, Maple or Matlab.

Supervisor: **Dr. Eric C.K. Cheung**, Department of Statistics and Actuarial Science (MW502J)
(eckc@hku.hk)

2. Dependence structures in multiple life insurances and annuities

The price of a multiple-life insurance/annuity product depends not only on the marginal distributions of the underlying future lifetimes, but also on their dependence structure. In this project, the effect of dependence structure on the actuarial present values will be studied. In the course of the research, the student will learn some basic theory of dependence structures.

Supervisor: **Dr. K.C. Cheung**, Department of Statistics and Actuarial Science (MW522)
(kccg@hkucc.hku.hk)

3. H shares and A shares

Many Chinese companies in China are dual-listed in Hong Kong and China (Shanghai or Shenzhen) by issuing H shares and A shares respectively, with price discrepancies having been found between them. The student who takes this project is expected to study the relationship between the movements of H shares and A shares, taking into account various economic factors.

Requirement: Knowledge of financial markets and SAS programming.

Supervisor: **Dr. K.S. Chong**, Department of Statistics and Actuarial Science (MW504)
(kschong@hku.hk)

4. Analysis of Card Counting Strategies in Casino Games

Card counting is the process of tracking the live cards played in a card game in order to determine when the deck is in favor of the player and hence when the player should increase the bet size. Since its development in the 1950's, the card counting strategy has been known to be capable of effectively increasing the player's edge in some casino games, especially blackjack. In this project, students will be asked to perform mathematical analysis on some selected casino games and explore different existing card counting strategies using Monte Carlo simulations. Students are also encouraged to design new card counting systems; or suggest modifications to the game rules to prevent card counters from profiting.

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

Supervisor: **Dr. Y.K. Chung**, Department of Statistics and Actuarial Science (MW504)
(yukchung@hku.hk)

5. Analysis of population census data

The Hong Kong Population Census was conducted in 2011. The census data provide a lot of information on the social and economic situation in Hong Kong. In this project, the student will analyse the 5% sample data set for the 2011 Population Census.

Requirement: Initiative and good knowledge about the social and economical situation in Hong Kong.

Supervisor: **Prof. W.K. Fung**, Department of Statistics and Actuarial Science (MW523)
(wingfung@hku.hk)

6. EM Algorithms for ML Factor Analysis

The general theory of EM algorithms proves that each iteration of EM increases the likelihood and also that if an instance of the algorithm converges, it converges to a (local) maximum of the likelihood. An advantage of EM algorithms, such as those for factor analysis, is that each iteration is simple to program and computationally inexpensive. This project is to explore the application of EM algorithms for maximum likelihood factor analysis.

Requirement: Knowledge of programming in SAS/IML or other programming languages is essential.

Supervisor: **Dr. C.W. Kwan**, Department of Statistics and Actuarial Science (MW508)
(cwkwan@hku.hk)

7. 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**, Department of Statistics and Actuarial Science (MW519)
(hrntlkf@hku.hk)

8. Statistical inference for local estimating equations

A problem in nutritional epidemiology had motivated Carroll, Ruppert and Welsh (1998, Journal of the American Statistical Association) to develop a general class of nonparametric estimators obtained by a set of what they term local estimating equations. Generality of their approach is exemplified by the fact that it covers as special cases methods like Nadaraya-Watson kernel regression, local polynomial regression, local likelihood estimation and varying coefficient generalised linear models. This project explores different applications of local estimating equations and studies problems of statistical inference based on local estimating equations.

Supervisor: **Prof. Stephen M.S. Lee**, Department of Statistics and Actuarial Science (MW528)
(smslee@hku.hk)

9. 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. Li**, Department of Statistics and Actuarial Science (MW502H)
(gdli@hku.hk)

10. Nonlinear Modelling of Environmental Time Series

The changes in environmental factors, for example, rainfall, wind speed, wind direction, affect our daily life in recent years. Due to the rapid changes in the atmospheric conditions, the modeling and hence prediction of these highly fluctuated factors becomes a challenging task since the conventional linear approach may be inappropriate. In this study, the nonlinearity and the non-Gaussianity of some environmental time series of Hong Kong will be explored by modern statistical techniques.

Requirement: Basic knowledge of linear time series models is preferred and the knowledge of R programming is required for this project.

Supervisor: **Dr. Gilbert C.S. Lui**, Department of Statistics and Actuarial Science (MW506)
(csglui@hku.hk)

11. Modeling of Risk Factors for Solvency Calculations of Life Insurance Companies

In order to fulfill the requirements of Solvency II, European life insurance companies have to develop internal models for calculations of the asset and liability values. However, these models are usually based on a complete valuation of the entire portfolio of the company and can be very time consuming when the company wants to test the impact of various risk factors and control variables on the solvency position. In this project, the student will analyze the risk factors affecting the solvency position of a typical insurance company and develop appropriate simple models for the solvency calculations. The student is expected to collect information relating to solvency calculations and experience data of various risk factors. The end result should be a workable model that can produce solvency calculations for a typical insurance company to perform analysis of the different levels of risk factors and resulting solvency requirements under various scenarios.

Supervisor: **Dr. Louis F. K. Ng**, Department of Statistics and Actuarial Science (MW505)
(flouisng@hku.hk)

12. Bayesian inference using MCMC sampling

In this project, the student shall learn the Bayesian inference using MCMC sampling methods, with some computer software, WINBUGS, SAS or R code.

Supervisor: **Prof. K.W. Ng**, Department of Statistics and Actuarial Science, (MW525)
(kaing@hku.hk)

13. Sample surveys with sensitive questions by non-randomized response techniques

In epidemiological, medical, public health, psychological and sociological surveys, investigators may require to solicit information on sensitive or highly personal questions such as sex, abortion, illegitimate birth, AIDs, illegal betting, shoplifting, drug-taking, tax evasion, annual income, family violence, students' cheating behavior and so on. When sensitive questions are asked directly, some respondents may refuse to answer; even worse, they might provide wrong answers to maintain privacy. To overcome the aforementioned difficulty, Warner (1965, JASA) proposed a so-called randomized response (RR) technique that allows investigators to obtain sensitive information while protecting privacy of respondents. Since the introduction of the Warner model, voluminous work related to the RR technique has been developed in various directions over the past 45 years. Despite these advances, all RR models have the following limitations: (I) a lack of reproducibility (i.e., the same respondent is expected to give the same answer by the design if the survey is conducted again); (II) a lack of trust from the interviewees because randomization device (RD) is controlled by interviewer; and (III) a higher cost due to the use of RDs.

In the past several years, to address some of these issues, without using any RDs some researchers have developed several non-randomized response (NRR) techniques including the crosswise design, the triangular design, and the parallel design for a single sensitive question with a binary outcome. In this project, the admitted candidates are expected to (i) collect real data in Hong Kong for a specific sensitive topic using the three models, (ii) to compare their performances, and (iii) to analyze these data by using likelihood-based method and Bayesian method via R program.

Requirement: Knowledge of LATEX technique, Bayesian methods, and R programming.

Supervisor: **Dr. Gary G.L. Tian**, Department of Statistics and Actuarial Science (MW520)
(gltian@hku.hk)

14. Comparison of Several Free Statistical Packages

This project aims to compare various free or even open-source statistical software packages, thereby providing students with opportunities to perform statistical computations with different resources. In this project, students should gain extensive experience in programming with several free statistical packages such as R, Sage, Scilab, SciPy, etc. Some common statistical analyses and simulation studies are to be carried out by using those packages and students are required to learn the way to code the programs for several statistical problems. Students taking this project are expected to be self-motivated in learning programming skills.

Supervisor: **Dr. K.P. Wat**, Department of Statistics and Actuarial Science (MW529)
(watkp@hku.hk)

15. 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**, Department of Statistics & Actuarial Science (MW511) (rwong@hku.hk)

16. Risk theoretic applications with a class of mixed Erlangs

It is known that a variety of distributions are of mixed Erlang type, in which case computational formula exists for many quantities of interests in risk theory. In this project, the student studies distributional properties of the class of Erlang mixtures as well as various risk theoretic applications including analysis of insurer's surplus process and discounted aggregate claims. In particular, the student is assumed to possess strong computational skills such as Mathematica, Matlab or Maple, also statistical background related to the EM algorithm.

Supervisor: **Dr. J.K. Woo**, Department of Statistics and Actuarial Science (MW530) (jkwoo@hku.hk)

17. Time Series Analysis by Generalized Threshold Model

Generalized Threshold Model (GTM, Samia and Chan, 2011) is a non-linear time series model extending the Threshold Autoregressive Model (TAR, Lim and Tong, 1980) to data of the entire exponential family. The popularity of the threshold models is based on the piecewise linearity for which the theory of autoregressive time series model can be easily applied. Students of this project are expected to use GTM for analyzing real time series data sets and compare the result with linear time series model like ARMA.

Requirement: Knowledge of R programming.

Supervisor: **Dr. Karl K.Y. Wu**, Department of Statistics & Actuarial Science (MW508) (karlwu@hku.hk)

18. Valuing Guaranteed Minimum Death Benefits

Motivated by the Guaranteed Minimum Death Benefits in various deferred annuities, this project investigates the calculation of the expected discounted value of a payment at the time of death. The payment depends on the price of a stock at that time and possibly also on the history of the stock price. The time-until-death random variable is assumed to be independent of the stock price process which is assume to follow a compound Poisson process.

I will provide all the related literature and ideas to the student who takes this project. The student should have taken some courses on option pricing. The student needs to do some theoretical work and numerical examples.

Supervisor: **Prof. H. Yang**, Department of Statistics & Actuarial Science (MW526)
(hlyang@hku.hk)

19. Analysis of large data sets: new tools from random matrix theory

Large data sets refer to data where the number of variables, or data dimension say p, is large compared to the sample size, say n. Modern statistical problems involve frequently such large data sets from various fields like genomic data analysis, financial portfolio optimization or design of wireless communication networks. For example in a genomic micro-array, p is several thousands and n several hundreds. Classical multivariate statistical tools dramatically fail to analyse these large data sets: either there are not applicable any more or lack efficiency.

New tools have emerged recently from the theory of random matrices. Most of them are based on the distribution of eigenvalues of sample covariance matrices which are computable from the data. Classical tools like Hotelling T^2 (tests on the mean), testing of generalized linear hypothesis (for regression or MANOVA) have been corrected or adapted to cope with large data sets. The theory behind is appealing and applications to large-dimensional data analysis are significant.

In this project, the student will i) learn some fundamental theorems from the theory of random matrices; ii) learn some new statistical tools developed recently; iii) start some own thinking about unsolved problems or perform some simulation experiments in order to get a deeper understanding of these results. I have included below an expository paper on the subject for a first introduction.

First reading: Z.D. BAI (2005). High dimensional data analysis. *COSMOS*, Vol. 1, No. 1, 17–27. (downloadable from: http://web.hku.hk/~jeffyao/Bai-cosmos-05.pdf)

Supervisor: **Dr. Jeff J. Yao**, Department of Statistics & Actuarial Science (MW502G)
(jeffyao@hku.hk)

20. Bayesian Hierarchical Modeling and Dose Finding with Longitudinal Data

Longitudinal data are common in clinical trials for Alzheimer's disease. In a phase II trial, the goal is to identify one or several doses that may have disease modification effects, which would be moved forward to phase III clinical trials for confirmative testing. Students will develop Bayesian hierarchical modeling for longitudinal measurements, and incorporate the slopes of disease deterioration into decision making. In the Bayesian adaptive framework, any arm that shows futility would be terminated earlier, so that the rest of patients would be allocated to the remaining arms. Students will conduct extensive simulation studies to demonstrate the performance of the Bayesian hierarchical modeling and dose finding method.

Requirement: R programming

Supervisor: **Dr. G. Yin**, Department of Statistics and Actuarial Science (MW502E)
(gyin@hku.hk)

21. Visualizing Big Ranking Data

Ranking data arise from a group of judges ranking a set of objects according to some preference criterion. Effective visualization of ranking data can reveal important statistical properties of the population of judges. In this project we develop an intuitive, easy to use, and computationally efficient framework for the visualization of ranking data with a large number of judges. The visualization techniques will be applied to big ranking data such as APA voting data and movie preferences. The knowledge of R is needed.

Supervisor: **Dr. Philip L.H. Yu**, Department of Statistics and Actuarial Science (MW521)
(plhyu@hku.hk)

22. What has driven up CPI in Hong Kong?

Students registered in this project are going to collect various of variables that might be related to dynamic of CPI in Hong Kong, and build proper regression model to identify any possible/potential reasons/resources for the driving up of CPI in Hong Kong these years.

Supervisor: **Dr. Z. Zhang**, Department of Statistics and Actuarial Science (MW511)
(zhangz08@hku.hk)

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