Institute of Statistical Research and Training
University of Dhaka

Curriculum

M. Phil Program in Applied Statistics and Data Science
Session : 2023–2024

www.isrt.ac.bd/academics/graduate

Institute of Statistical Research and Training

The Institute of Statistical Research and Training (ISRT), University of Dhaka, is the leading institution for training and research in Applied Statistics and Data Science in Bangladesh. It was founded in 1964 by the Late National Professor Dr. Qazi Motahar Husain, an eminent scientist, academician and a leading proponent of the statistical sciences in this country. The Institute offers a 4-year B.S. (Honours) program that has been designed to produce graduates with strong statistical computing skills, sound knowledge of statistical concepts and the versatility to apply these concepts in areas as diverse as medicine, engineering, economics and the social sciences. The 1-year M.S. program consists of specialized courses in areas ranging from environmental statistics to statistical signal processing, statistical machine learning, and causal inference, and has been designed for students with a keen interest in higher studies and research. In addition, the Institute offers Ph.D. and M.Phil. degree programs. Highly experienced faculty members with a minimum educational qualification of a masters degree, most of whom also have Ph.D. degrees from reputed universities across the world, run these programs.

ISRT boasts an academic environment that is highly competitive and conducive to research. Both students and faculty members benefit from the regular seminars and talks given by researchers from home and abroad on topics of current interest. The Institute has a rich library with well over 15,000 books and is equipped with three state-of-the-art computer labs, cloud computing facilities and high-speed internet access for graduate and undergraduate students. The aim is to provide a learning environment that stimulates intellectual curiosity, critical thinking and independent problem-solving skills. The Journal of Statistical Research (JSR), an international journal which is being published bi-annually by ISRT since 1970, serves as a forum for the exchange of research ideas between statisticians in Bangladesh and abroad. Faculty members conduct research in diverse areas such as biostatistics, spatial statistics, statistical pattern recognition, Bayesian analysis and econometrics and regularly publish in peer-reviewed journals.

Among its other activities, the Institute frequently organizes short courses and training programs for non-statisticians working in government and non-government organizations who find themselves using statistics in their work. In doing so, it has played an active role in promoting and creating awareness about the need for sound statistical practices among people from other disciplines so that they may work more efficiently within their organizations. ISRT also maintains close ties with the Bangladesh Bureau of Statistics (BBS) and other organizations responsible for the collection and dissemination of statistical data in Bangladesh, and is frequently called upon to offer its expertise on statistical issues of national interest. Over the years the institute has played a significant role in the country's development by producing world class statisticians for academia and industry in addition to providing statistical expertise on issues of national interest. In addition to that the Institute provides statistical consulting service through StatLab primarily for the students and faculty members of the University of Dhaka, with an aim to strengthen research on campus by assisting graduate students and faculty members of other disciplines.

Vision of the Institute

To take a leading role in producing competent graduates in Applied Statistics and Data Science, conducting cutting-edge research, and creating industrial partnerships to address national and global interests and challenges.

Mission of the Institute

To pursue excellence in Applied Statistics and Data Science education and research and to provide data-driven solutions to industries and stakeholders for the benefit of society.

- M1 To provide quality education in Applied Statistics and Data Science by ensuring an effective learning environment
- M2 To perform original and impactful research in Applied Statistics and Data Science that would enhance knowledge and contribute to the well-being and advancement of society.
- M3 To provide innovative data-driven solutions to the problems and challenges faced by the industries and other stakeholders.

M. Phil Program in Applied Statistics and Data Science

The Master of Philosophy (M. Phil) program in Applied Statistics and Data Science is a TWO academic year program. According to the Dhaka University regulations for M. Phil admission in Applied Statistics and Data Science, candidates who have either four year B.Sc./B.S. honours and one year M.Sc./M.S. degrees or three year B.Sc./B.S. honours and one year M.Sc./M.S. degrees or two year bachelor and two year M.Sc./M.S. degrees **in Applied Statistics and Data Science/Applied Statistics/Statistics** are usually eligible to get admission in M. Phil. Details are given in Dhaka University MPhil policy. Students in MPhil are required to take course work at the 1st year. The syllabus is designed as follows. The candidates are required to take a total of 9 credit hour of theoretical courses from two groups of courses: A and B. The group A consists of the courses related to basic topics and the group B consists of the courses related to advanced topics in Applied Statistics and Data Science. Course from both groups are of 3 credit hours each. Candidates are required to take THREE courses to make a total of 9 credit hours for theoretical courses, maximum TWO courses from each group. The choice of courses will depend on the availability of teaching faculties of the institute. In addition, there will be a 3 credit hours oral comprehensive course, altogether a total of 12 credit hours.

Breakdown of the credit hours

Courses	Credit Hour
Theoretical Courses	9
Oral	3
Total	12

The marks allocation for courses will be as follows:

Theoretical		
Attendance	:	05
In-course exam	:	25
Final exam	:	70

There will be two in-course examinations for each of the courses. Candidate should be present at least 60% of the classes to attend the final exam. The Dhaka University grading policy will be followed for finalizing the result. Candidate must get minimum GPA 2.5 for theoretical courses and oral course, separately. However, no 'F' grade in any theoretical courses will be acceptable. The candidate must have to sit for the examination for the course(s) for which the candidates got 'F' grade. Candidates who have successfully completed first year MPhill with minimum GPA 2.5 in theoretical and oral courses, separately may be transferred to the Ph.D programme within second year on the recommendation(s) of the supervisor(s) certifying satisfactory progress of research work and the Academic Committee of the Institute, Ph.D sub-committee and faculty concerned. The recommendation(s) for transfer will be sent by the Academic Committee of the Institute, Ph.D sub-committee and faculty concerned directly to the Board of Advanced Studies and to the Academic Council , which will accord the final approval.

Courses in Group A

Course ID	Course Title	Credit Hour
AST601	Applied Bayesian Statistics	3
AST602	Advanced Classical Inference	3
AST603	Advanced Multivariate Techniques	3
AST604	Generalized Linear Models	3

Courses in Group B

Course ID	Course Title	Credit Hour
AST610	Advanced Survival Analysis	3
AST611	Environmental and Spatial Statistics	3
AST612	Advanced Time Series Analysis	3
AST615	Advanced Econometric Methods	3
AST618	Introduction to Causal Inference	3
AST619	Analysis of Longitudinal Data	3
AST622	Statistical Signal Processing	3
AST623	Meta Analysis	3
AST624	Clinical Trials	3
AST625	Statistical Machine Learning	3
AST626	Big Data Analytics	3
AST627	Advanced Statistical Machine Learning	3

Viva

Course ID	Course Title	Credit Hour
AST 640	Oral	3

AST 601: Advanced Bayesian Statistics — Credit 3

Introduction

Bayesian statistics refers to practical inferential methods that use probability models for both observable and unobservable quantities. The flexibility and generality of these methods allow them to address complex real-life problems that are not amenable to other techniques. This course will provide a pragmatic introduction to Bayesian data analysis and its powerful applications.

Objectives

Acquire basic understanding in the principles and techniques of Bayesian data analysis. Apply Bayesian methodology to solve real-life problems. Utilize R for Bayesian computation, visualization, and analysis of data.

Learning Outcomes

Upon completion of the course, students will learn about i) the concept of the Bayesian statistical methods, ii) formulation and derivation of prior and posterior distributions, iii) estimation of the model using different MCMC methods, and iv) application of the Bayesian methods to analyze data and interpret and compare results with the frequentist approach.

Contents

Bayesian thinking: background, benefits and implementations; Bayes theorem, components of Bayes theorem - likelihood, prior and posterior; informative and non-informative priors; proper and improper priors; discrete priors; conjugate priors; semi-conjugate priors; exponential families and conjugate priors; credible interval; Bayesian hypothesis testing; building a predictive model.

Bayesian inference and prediction: single parameter models - binomial model, Poisson model, normal with known variance, normal with known mean; multi-parameter models - concepts of nuisance parameters, normal model with a non-informative, conjugate, and semi-conjugate priors, multinomial model with Dirichlet prior, multivariate normal model; posterior inference for arbitrary functions; methods of prior specification; method of evaluating Bayes estimator.

Summarizing posterior distributions: introduction; approximate methods: numerical integration method, Bayesian central limit theorem; simulation method: direct sampling and rejection sampling, importance sampling; Markov Chain Monte Carlo (MCMC) methods - Gibbs sampler, general properties of the Gibbs sampler, Metropolis algorithm, Metropolis-Hastings (MH) sampling, relationship between Gibbs and MH sampling, MCMC diagnostics - assessing convergence, acceptance rates of the MH algorithm, autocorrelation; evaluating fitted model - sampling from predictive distributions, posterior predictive model checking.

Linear model: introduction, classical and Bayesian inference and prediction in the linear models, hierarchical linear models - Bayesian inference and prediction, empirical Bayes estimation; generalized linear model - Bayesian inference and prediction (logit model, probit model, count data model); model selection - Bayesian model comparison.

Nonparametric and Semiparametric Bayesian models.

Text Books

1. Hoff PD (2009). A First Course in Bayesian Statistical Methods. Springer.

Reference Books

1. Gelman A, Carlin JB and Stern HS, Dunson DB, Vehtari A, and Rubin DB (2013). Bayesian Data Analysis, *3rd edition*. Chapman and Hall.
2. Gill J (2007). Bayesian Methods: A Social and Behavioral Sciences Approach, *2nd edition*. Chapman and Hall.

AST 602: Advanced Classical Inference — **Credit 3**

Introduction

A branch of statistics has been developed to draw conclusion in a short time and cost-effective way regarding the population of interest which is ubiquitously known as statistical inference. It facilitates both parametric and nonparametric approaches under the umbrella of classical and Bayesian paradigms.

Objectives

The course is designed to teach students on advanced methods of statistical inferences including, optimize techniques, parametric and semi-parameric inferential procedures, numerical methods, resampling techniques. Keeping the diversity of demands in current world, this course is designed in such a way that the students can build up their research career in a wide variety of fields such as social science, medical statistics, clinical trials, spatial statistics, multivariate statistics, etc.

Learning Outcomes

After completion of this course, students (i) will be able to draw statistical inference both in classical and Bayesian framework, (ii) will have ample skills in handling data to meet the inferential needs in diverse areas of applications.

Contents

Statistical inference: parametric, nonparametric and semiparametric inference.

Approximate and computationally intensive methods for statistical inference: the general problem of inference; likelihood functions; maximum likelihood estimation; optimization techniques-Newton type methods; EM algorithm-simple form, properties, uses in analysing missing data, fitting mixture models and latent variable model; restricted maximum likelihood (REML) method of estimation; Multi-stage maximization; Efficient maximization via profile likelihood; confidence interval and testing hypothesis in these complex cases; Bayesian method of inference: prior and posterior distribution, different types of prior, credible intervals and testing hypothesis; analytical approximations-asymptotic theory, Laplace approximation; numerical integral methods-Newton-Cotes type methods; Monte carlo methods; simulation methods-Markov chain Monte Carlo.

Resampling techniques: bootstrap-confidence intervals, test, parametric bootstrap, advantages and disadvantages of parametric bootstrap; jackknife-confidence interval, test and permutation test.

Nonparametric inference and robustness: introduction, inference concerning cumulative distribution function (cdf), quantiles and statistical functionals: empirical cdf, quantiles, estimating statistical functionals, influence functions, testing statistical hypothesis-one sample settings, two or more sample settings; tolerance limit; empirical density estimation- histograms, kernel, kernel density estimation.

Text Books

1. Casella G and Berger RL (2003). Statistical Inference, *2nd edition*. Duxbery.

Reference Books

1. Millar RB (2011). Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB. Wiley.
2. Hogg RV, McKean J and Craig AT (2010). Introduction to Mathematical Statistics, *7th edition*. Pearson.

AST 603: Advanced Multivariate Techniques — **Credit 3**

Introduction

Multivariate analysis skills have been recognized as part of the key requisites for statistical analysts. The complexity of most phenomena in the real world requires an investigator to collect and analyze observations on many different variables instead of a single variable. The desire for statistical techniques to elicit information from multivariate dimensional data thus becomes essential and crucial for data analysts. This course focuses on multivariate

methods based on normal theory. It gives students working knowledge on how to analyze data and solve problems involving measurements of p variables on each of n subjects.

Objectives

The objective of this course is to give students experience with multivariate techniques in the analysis of research data. The aim is to teach students how to select appropriate methods of multivariate data analysis and interpret the results.

Learning Outcomes

Having successfully completed this course, students should be able to (i) know the theoretical concept of advanced multivariate methods, (ii) know about data-dimension concept, necessity and techniques (iii) apply the methods for analysing real life problem and interpret the results.

Contents

Principal components: population principal components, summarizing sample variations by principal components, graphing the principal components, large sample inference.

Factor analysis: the orthogonal factor models, methods of estimation (maximum likelihood estimates and principal factor analysis), selection of loadings and factor (factor rotation, varimax rotation, quartimax rotation, oblimin rotations), factor scores, structural equations models.

Canonical correlation analysis: canonical variates and canonical correlations, sample canonical variates and sample canonical correlations, large sample inference.

Discrimination and classification: separation and classification of two populations, classification of two multivariate normal populations, evaluating classification functions, Fisher's discriminant function, classification with several populations, Fisher's method for discriminating several populations.

Clustering: similarity measures, hierarchical clustering methods, nonhierarchical clustering methods; fuzzy clustering, determination of number of clusters: Gap statistics and its several modifications, several cluster validity indices, cluster's homogeneity test; multidimensional scaling.

Text Books

1. Johnson RA and Wichern DW (2008). Applied Multivariate Statistical Analysis, *6th edition*. Prentice-Hall.

Reference Books

1. Srivastava MS (2002). Methods of Multivariate Statistics. Wiley.

2. T. W. Anderson (2003). An Introduction to Multivariate Statistical Analysis. *3*rd edition. Wiley.

AST 604: Generalized Linear Models — **Credit 3**

Introduction

This course deals with different statistical models for the analysis of quantitative and qualitative data, of the types usually encountered in research.

Objectives

To introduce to the students about the statistical methods including the general linear model for quantitative responses (including multiple regression, analysis of variance and analysis of covariance), binomial regression models for binary data (including logistic regression and probit models), and models for count data (including Poisson regression and negative binomial models). All of these techniques are covered as special cases of the Generalized Linear Model, which provides a central unifying statistical framework for the entire course.

Learning Outcomes

After completing the course students will be familiar with (i) the exponential family of distributions, (ii) the class of generalized linear models (GLM) as regression models with responses from the exponential family of distributions, (ii) the concepts of link functions for modeling the correspondence between the expected value of the responses and covariates and of variance functions for specifying the correspondence between the expected values and variances of the responses, (iii) analyzing data from important special cases of GLMs, in particular logistic regression and Poisson regression, and (iv) extensions of the GLM framework using quasi likelihood based on specied link and variance functions.

Contents

Generalized linear models: exponential family of distributions; estimation: method of maximum likelihood, method of least squares, estimation of generalized linear models; inference: sampling distribution for scores, sampling distribution for maximum likelihood estimators, confidence intervals for model parameters, adequacy of a model, sampling distribution for log-likelihood statistic, log-likelihood ratio statistic (deviance), assessing goodness of fit, hypothesis testing; multiple regression: maximum likelihood estimation, log-likelihood ratio statistic.

Models for binary responses: probability distributions, generalized linear models, dose response models, general logistic regression, maximum likelihood estimation and log-likelihood

ratio statistic, other criteria for goodness of fit, least square methods; multinomial distributions; nominal logistic regression models; ordinal logistic regression models.

Models for count data, Poisson regression and log-linear models: probability distributions, maximum likelihood estimation, hypothesis testing and goodness of fit.

Text Books

1. Dobson, A J and Barnett, A G. (2008). An Introduction to Generalized Linear Models, *3rd edition*. Chapman & Hall.

Reference Books

1. McCullagh, P. and Nelder, J A. (1989). Generalized Linear Models, *2nd edition*. CRC Press.
2. Alan Agresti (2007). An Introduction to Categorical Data Analysis. *2*nd edition. Wiley.

AST 610: Advanced Survival Analysis — Credit 3

Introduction

An introduction to methods of analysing correlated time-to-event data is provided in this course. Some commonly used methods for analysing univariate time-to-event data, e.g. Kaplan-Meier estimate of survivor functions, Cox's proportional hazards models, etc., are reviewed using counting processes notations.

Objectives

The objectives of the course are to teach the theoretical basis of different methods related to analysing correlated time-to-event data and competing risks model and to apply statistical softwares to analyse data using such models.

Learning Outcomes

At the end of the course, students are expected i) to understand the theoretical basis of different methods related to analysing correlated time-to-event data and competing risks model ii) to use a statistical software (e.g. related R packages) to analyse data using such models iii) to interpret the results and write scientific publication.

Contents

Estimating the Survival and Hazard Functions: Introduction and notation, the Nelson-Aalen and Kaplan-Meier estimators, counting process and martingals, properties of Nelson-Aalen estimator.

Semiparametric Multiplicative Hazards Regression Model: Introduction, estimation of parameters, inclusion of strata, handling ties, sample size determinations, counting process form of a Cox model, time-dependent covariates, different types of residuals for Cox models, checking proportionality assumption.

Multiple Modes of Failure: Basic characteristics of model specification, likelihood function formulation, nonparametric methods, parametric methods, semiparametric methods for multiplicative hazards model.

Analysis of Correlated Lifetime Data: Introduction, regression models for correlated lifetime data, representation and estimation of bivariate survivor function.

Text Books

1. Therneau TM and Grambsch PM (2000). Modeling Survival Data: Extending the Cox Model, Springer.

Reference Books

1. Kalbfleisch JD and Prentice RL (2002). The Statistical Analysis of Failure Time Data, *2nd edition*. Wiley.
2. Hougaard P (2000). Analysis of Multivariate Survival Data. Springer.

AST 611: Environmental and Spatial Statistics — Credit 3

Introduction

Spatial statistics encompasses diversified statistical methods for analyzing data obtained from stochastic process indexed by the space. This branch is enrich enough to gain insight from data exploiting the dependence over space. Its myriad applications caught profound attention of people from both academia and practitioners.

Objectives

Technology is indispensable for modern life, and its advances in different aspects of our life made several things possible. Now a days data have been collected along with extensive additional information. Spatial data is one of such examples. In recent years, analysis of spatial data receives great attention over the world. As a result, several theories have been developed for different types of spatial data analysis. This course is designed to introduce

the graduate student with few of such theories so that they can develop their skill in spatial data analysis. To comprehend this course, students need a sound knowledge of Mathematical statistics, particularly the concepts of stochastic process. It is expected that the student will be able to analyze different spatial data from diverse fields after successful completion of the course.

Learning Outcomes

After completing this course students are expected to have knowledge about i) spatial and non spatial data, ii) geostatistical data and analysis, iii) spatial interpolation, iv) apply auto regressive model to areal data, v) point pattern data analysis.

Contents

Review of non-spatial statistics and stochastic process, overview of different types of spatial data; random field and spatial process - geostatistical/point reference process, areal/lattice process and point process; spatial data concern.

Geostatistical data: real data examples, measure of spatial dependence- variogram and covariance, stationarity and isotropic, variograms and covariance functions, fitting the variograms functions; Kriging, linear geostatistical model - formulation, simulation, estimation and prediction, generalized linear geostatistical model - formulation, simulations, estimation and prediction. Areal data: neighborhoods, testing for spatial association, autoregressive models (CAR, SAR), estimation/inference; grids and image analysis, disease mapping. Point pattern data: locations of events versus counts of events, types of spatial patterns, CSR and tests - quadrat and nearest neighbor methods, K-functions and L-functions, point process models- estimation and inference, health event clustering.

Special topics in spatial modeling: Hierarchical models, Bayesian methods for spatial statistics, Bayesian disease mapping, Spatio-temporal modeling, more on stationarity. Use of R and GIS software to give emphasis on analysis of real data from the environmental, geological and agricultural sciences.

Text Books

1. Cressie N (1993). Statistics for Spatial Data, *Revised edition*. Wiley.
2. Banerjee S, Carlin BP, and Gelfand AE (2014). Hierarchical Modelling and Analysis for Spatial Data, *2nd edition*. Chapman and Hall.

Reference Books

1. Cressie N and Wikle CK (2011). Statistics for Spatio-Temporal Data. Wiley.
2. Illian J, Penttine A, Stoyan H and Stoyan D (2008). Statistical Analysis and Modelling of Spatial Point Patterns. Wiley.

AST 612: Advanced Time Series Analysis — **Credit 3**

Introduction

This is an introductory course of time series theory. The objective of this course is to equip students with various classical time series models, deriving their properties, inference methods and forecasting techniques for analyzing time series data. From computational point of view, it aims to demonstrate the theory with real datasets. Conclusions and proofs are given for some basic formulas and models; these enable the students to understand the principles of time series theory.

Objectives

This course is designed to make student familiar with time series data and methods for analysing such data and to use them in forecasting.

Learning Outcomes

On successful completion of this course, the students are expected to (i) identify the main components of the time series and apply a suitable exponential smoothing technique to forecast with a variety of time series models such as additive or multiplicative seasonal models by updating the components, (ii) formulate time series regression models for the series with trend and seasonal components and make forecasts from these models, (iii) Identify ARIMA models tentatively from ACF and PACF, (iv) estimate parameters of the ARIMA models by the method of moments, least squares method and maximum likelihood method, and (v) check the adequacy of the model and make forecasts.

Contents

Introduction and examples of time series; simple descriptive techniques: time series plots, trend, seasonal effects, sample autocorrelation, correlogram, filtering.

Probability models: stochastic processes, stationarity, second-order stationarity, white noise model, random walks, moving average (MA) processes, autoregressive (AR) processes, ARMA processes, seasonal ARMA processes, the general linear process; properties, estimation and model building, diagnostic checking.

Forecasting: naive procedures, exponential smoothing, Holt-Winters, Box-Jenkins forecasting, linear prediction, forecasting from probability models.

Non-stationary time series: non-stationarity in variance-logarithmic and power transformations; non-stationarity in mean; deterministic trends; integrated time series; ARIMA and seasonal ARIMA models; modelling seasonality and trend with ARIMA models.

Stationary processes in the frequency domain: the spectral density function, the periodogram, spectral analysis.

Concept of state-space models: dynamic linear models and the Kalman filter.

Text Books

1. Jonathan DC and Kung-Sik C (2008). Time Series Analysis - With Applications in R. Springer.
2. Spyros M, Steven W and Rob H (1997). Forecasting – Methods and Applications, *3rd edition*. Wiley.

Reference Books

1. Chatfield C (2003). The Analysis of Time Series, *6th edition*. Chapman & Hall.
2. Shumway RH and Stoffer DS (2011). Time Series Analysis and Its Applications: With R Examples. Springer.
3. Brockwell PJ and Davis RA (2002). Introduction to Time Series and Forecasting. *3rd edition*. Springer.

AST 615: Advanced Econometric Methods — **Credit 3**

Introduction

This course covers a range of econometric methods required to conduct empirical economic research and understand applied econometric results. Topics include models for panel data, simultaneous equations models, models with lagged variables, and limited dependent variables.

Objectives

To introduce students to the basic principles of econometric analysis. To gain theoretical understanding of the methods needed for econometric research including their underlying assumptions, advantages and limitations. To understand how to use different econometric tools in real-world economic problems and interpret findings

Learning Outcomes

On successful completion of this course the students should be able to (i) understand the basic principles of econometric analysis and econometric model building, (ii) gain theoretical understanding of the methods/models needed for econometric research including their underlying assumptions, advantages and limitations and (iii) understand how to use different econometric tools in real-world economic problems and interpret findings.

Contents

Econometric modeling, data and methodology; specification analysis and model building: bias caused by omission of relevant variables, pretest estimation, inclusion of irrelevant variables, model building; testing non-nested hypotheses, encompassing model, comprehensive approach-J test, Cox test; model selection criteria.

Models for panel data: fixed effects: testing significance of group effects, within- and between-groups estimators, fixed time and group effects, unbalanced panels and fixed effects; random effects: GLS, FGLS, testing for random effects, Hausman's specification test.

Simultaneous equations models: illustrative systems of equations, endogeneity and causality; problem of identification: rank and order conditions for identification; limited information estimation methods: OLS, estimation by instrumental variables (IV), Two-Stage Least Squares (2SLS), GMM Estimation, limited information maximum likelihood and the k class of estimators, 2SLS in nonlinear models; system methods of estimation: Three-Stage Least Squares (3SLS). full-information maximum likelihood, GMM estimation, recursive systems and exactly identified equations; comparison of methods-Klein's Model I; specification tests; properties of dynamic models: dynamic models and their multipliers.

Models with lagged variables: lagged effects in a dynamic model, lag and difference operators; simple distributed lag models: finite distributed lag models, infinite lag model: geometric lag model; Autoregressive Distributed Lag (ARDL) models: estimation of the ARDL model, computation of the lag weights in the ARDL model, stability of a dynamic equation, forecasting; Vector Autoregressions (VAR): model forms, estimation, testing procedures, exogeneity, testing for Granger causality, impulse response functions, structural VARs, application: policy analysis with a VAR.

Limited dependent variable: truncated distributions, moments of truncated distributions, truncated regression model; censored data: censored normal distribution, censored regression (Tobit) model, estimation, issues in specification; censoring and truncation in models for counts, application: censoring in the Tobit and Poisson regression models.

Text Books

1. Greene WH (2011). Econometric Analysis, *7th edition*. Prentice Hall.

Reference Books

1. Gujarati DN (2010). Basic Econometrics, *5th edition*. McGraw-Hill.
2. Wooldridge JM (2010). Introductory Econometrics: A Modern Approach, *5th edition*. Cengage Learning.

AST 618: Introduction to Causal Inference | **Credit 3**

Introduction

The course provides an introduction to causal inference with a cohesive presentation of concepts of, and methods for, causal inference.

Objectives

The aim of the course is to facilitate students to define causation in biomedical research, describe methods to make causal inferences in epidemiology and health services research, and demonstrate the practical application of these methods.

Learning Outcomes

Upon completion of the course, students are expected i) to be familiar with causation in biomedical research ii) to learn methods and models for estimating causual inference iii) to apply the methods to real data and interpret the results.

Contents

Causal effects: individual causal effects, average causal effects, causation versus association.

Randomized experiments: randomization, conditional randomization, standardization, and inverse probability weighting.

Observational studies: identifiability conditions, exchangeability, positivity, and consistency.

Effect modification: stratification, matching, and adjustment methods.

Interaction: identifying interaction, counterfactual response type and interaction.

Directed acyclic graphs (DAGs): complete and incomplete DAGs, statistical DAGs, DAGs and models, paths, chains and forks, colliders, d-seperation.

Unconfounded treatment assignment: balancing scores and the propensity score; Estimating propensity scores: selecting covariates and interactions, constructing propensity score strata, assessing balance conditional on estimated propensity score; Assessing overlap in covariate distributions; Matching to improve balance in covariate distributions: selecting subsample of controls to improve balance, theoretical properties of matching procedures; Subclassification on propensity scores: weighting estimators and subclassification; Matching estimators: matching estimators of ATE; A general method for estimating sampling variances for standard estimators for average causal effects.

Longitudinal causal inference: g-formula and marginal structural models.

Mediation analysis: traditional approaches (direct and indirect effects), counterfactual definitions of direct and indirect effects, regression for causal mediation analysis, sensitivity analysis

Text Books

1. Hernan MA and Robins JM (2019). Causal inference. Boca Raton: Chapman & Hall/CRC
2. Imbens GW and Rubin DB (2015). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press.

Reference Books

1. Morgan SL and Winship C (2014). Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge.
2. VanderWeele T (2015). Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford.

AST 619: Analysis of Longitudinal Data — Credit 3

Introduction

Longitudinal data arise when multiple measurements of a response are collected over time for each individual in the study and hence are likely to be correlated, which presents substantial challenge in analyzing such data. This course covers topics related to statistical methods and models for drawing scientific inferences from longitudinal data.

Objectives

The objectives of the course are to teach students to understand the unique features of and the methodological implication of analyzing the data from longitudinal studies, as compared to the data from traditional studies, to understand statistical methods/models, particularly linear/generalized linear mixed models, GEE approaches for analyzing longitudinal data. It is also expected that students will be familiar with the proper implementation and interpretation of the statistical methods/models for analyzing longitudinal data and software packages analyzing such data.

Learning Outcomes

Upon completion of the course, students will achieve skills i) to understand the nature of longitudinal/clustered data ii) to understand the models and methods for analysing longitudinal/clustered data iii) to analyse such data and interpret the results iv) to understand and interpret research findings of the published reports and articles.

Contents

Longitudinal data: Concepts, examples, objectives of analysis, problems related to one sample and multiple samples, sources of correlation in longitudinal data, exploring longitudinal data.

Linear model for longitudinal data: Introduction, notation and distributional assumptions, simple descriptive methods of analysis, modelling the mean, modelling the covariance, estimation and statistical inference.

ANOVA for longitudinal data: Fundamental model, one sample model, sphericity condition; multiple samples models.

Linear mixed effects models: Introduction, random effects covariance structure, prediction of random effects, residual analysis and diagnostics.

Extension of GLM for longitudinal data: Review of univariate generalized linear models, quasi-likelihood, marginal models, random effects models, transition models, comparison between these approaches; the GEE methods: methodology, hypothesis tests using wald statistics, assessing model adequacy; GEE1 and GEE2.

Introduction to the concept of conditional models, joint models, their applications to bivariate binary and count data. Estimation, inference and test of independence.

Generalized Linear Mixed Models (GLMM): Introduction, estimation procedures–Laplace transformation, penalized quasi-likelihood (PQL), marginal quasi-likelihood (MQL).

Numerical integration: Gaussian quadrature, adaptive gaussian quadrature, Monte Carlo integration; markov chain Monte Carlo sampling; comparison between these methods.

Statistical analysis with missing data: Missing data, missing data pattern, missing data mechanism, imputation procedures, mean imputation, hot deck imputation. estimation of sampling variance in the presence of non-response, likelihood based estimation and tests for both complete and incomplete cases, regression models with missing covariate values, applications for longitudinal data.

Text Books

1. Verbeke G and Molenberghs G (2000). Linear Mixed Model for Longitudinal Data. Springer.
2. Molenberghs G and Verbeke G (2005). Models for Discrete Longitudinal Data. New York: Springer-Verlag.

Reference Books

1. Islam MA and Chowdhury RI (2017). Analysis of Repeated Measures Data. Springer.
2. Diggle PJ, Heagerty P, Liang K-Y, and Zeger SL (2002). Analysis of Longitudinal Data, *2nd edition*. Oxford.

AST 622: Statistical Signal Processing — Credit 3

Introduction

Much of the information around us can be described as signals. Statistical signal processing uses stochastic processes, statistical inference and mathematical techniques to describe, transform, and analyze signals in order to extract information from them.

Objectives

This course is designed to provide statistics graduate students with an overview of different types of signals, their representations and the use of statistical methods, such as estimation and hypothesis testing, to extract information from signals. Its objective is to introduce students to real life applications demonstrating the use of statistics in signal analysis.

Learning Outcomes

At the end of this course a student should be able to : (i) understand basic concepts of signals, signal properties and their representations in time and frequency domains (ii) Apply well-known statistical estimation techniques to estimate signal parameters from noisy signal measurements (iii) Apply well-known statistical decision theory methods to detect signals in Gaussian noise and assess the performance of these methods (iv) Understand and appreciate the importance of statistics in solving real life problems occurring in the domain of signal processing and communication.

Contents

Introduction to signals: Signals and their classification; real world analog signals: audio, video, biomedical (EEG, ECG, MRI, PET, CT, US), SAR, microarray, etc; digital representation of analog signals; role of transformation in signal processing. Orthogonal representation of signals. Review of exponential Fourier series and its properties.

Signal estimation theory: Estimation of signal parameters using ML, EM algorithm, minimum variance unbiased estimators (Rao-Blackwell theorem, CRLB, BLUE), Bayesian estimators (MAP, MMSE, MAE), linear Bayesian estimators.

Signal detection theory: Detection of DC signals in Gaussian noise: detection criteria (Bayes risk, Probability of error, Neyman-Pearson), LRT; detection of known signals in Gaussian noise: matched filter and its performance, minimum distance receiver; detection of random signals in Gaussian noise: the estimator correlator.

Applications: Scalar quantization, image compression, pattern recognition, histogram equalization, segmentation, application of signal estimation and detection theory to signal communication, signal recovery from various types of linear and nonlinear degradations, copyright protection, enhancement, etc.

Text Books

1. Kay S.M. (1993). Fundamentals of Statistical Signal Processing: Estimation Theory. Prentice Hall.
2. Kay SM (1998). Fundamentals of Statistical Signal Processing: Detection Theory. Prentice Hall.
3. Gonzales RC and Woods RE (2017). Digital Image Processing. *4*th edition, Pearson.

Reference Books

1. Gonzalez RC and Woods RE (2008). Digital Image Processing, *3rd edition*. Pearson Education, Inc.
2. Rahman SMM, Howlader T, Hatzinakos D (2019). Orthogonal Image Moments for Human-Centric Visual Pattern Recognition. Springer.
3. Soliman SS and Mandyam DS (1998). Continuous and Discrete Signals and Systems, *2nd edition*. Prentice-Hall.

AST 623: Meta Analysis — Credit 3

Introduction

Meta-analysis refers to the quantitative analysis of study outcomes. Meta-analysis consists of a collection of techniques that attempt to analyze and integrate results that accrue from research studies. This course provides an overview of systematic review and meta-analysis from a statistician's point of view.

Objectives

The main objectives are to introduce students with the merits of meta-analysis and how it can form an important and informative part of a systematic review, with the most common statistical methods for conducting a meta-analysis, and with how to analyze and interpret the results.

Learning Outcomes

At the end of the course, students should be familiar with i) the research synthesis, ii) systematic review of the existing research iii) data extraction from systematic review, iv) models and methods for analyzing meta-data for new findings and publication.

Contents

Introduction to systematic review and meta analysis: Motivation, strengths and weakness of meta-analysis, problem formulation (why study meta analysis), systematic review process.

Types of results to summarize; overview of effect size; effect size calculation for both continuous and discrete data.

Combining effect size from multiple studies; fixed effect and random effects models and their estimation; heterogeneity between studies and its estimation techniques; test of homogeneity in meta analysis; prediction intervals; subgroup analysis, Meta regression: random effect meta regression, baseline risk regression.

Publication bias in meta analysis; Power analysis for meta analysis; effect size rather than P-values; Meta analysis based on direction and P-values, reporting the results of meta analysis.

Introduction to Bayesian approach to meta analysis; Meta analysis for multivariate/longitudinal data; network meta analysis.

Text Books

1. Borenstein M, Hedges LV, Higgins JPT and Rothstein HR (2009). Introduction to Meta-Analysis, John Wiley & Sons, UK.
2. Hartung J and Knapp G and Sinha BK (2011). Statistical Meta-Analysis with Applications. John Wiley & Sons, UK.

Reference Books

1. Mathias Harrer, Pim Cuijpers, Toshi Furukawa, David Ebert (2019). Doing Meta-Analysis with R. CRC Press.
2. Ding-Geng Chen, Karl E. Peace (2019): Applied Meta-Analysis with R and Stata. Chapman & Hall

---

AST 624: Clinical Trials — Credit 3

Introduction

The clinical trial is "the most definitive tool for evaluation of the applicability of clinical research". It represents "a key research activity with the potential to improve the quality of health care and control costs through careful comparison of alternative treatments". The course is designed to give an overall idea of clinical trial studies. It will provide an introduction to the statistical and ethical aspects of clinical trials research.

Objectives

The main objectve o fthe course is to teach students on the topics include design, implementation, and analysis of trials, including first-in- human studies, phase II and phase III studies. The course will enable applying existing methodologies in designing clinical trials and will also foster research in this area.

Learning Outcomes

Upon completion of the course, students will achieve skills i) to understand, design a trial for assess the effectiveness of a drug, and ii) to implement and analysis data from such trials and interpret the results.

Contents

Statistical approaches for clinical trials: Introduction, comparison between Bayesian and frequentist approaches and adaptivity in clinical trials. Phases of clinical trials, pharmacokinetics (PK) and pharmacodynamics (PD) of a drug, dose-concentration-effect relationship and compartmental models in pharmacokinetic studies.

Phase I studies: Determining the starting dose from preclinical studies. Rule-based designs: 3+3 design, Storer's up-and-down designs, pharmacologically-guided dose escalation and design using isotonic regression. Model-based designs: continual reassessment method and its variations, escalation with overdose control and PK guided designs.

Phase II studies: Gehan and Simon's two-stage designs. Seamless phase I/II clinical trials: TriCRM, EffTox and penalised D-optimum designs for optimum dose selection.

Phase III studies: Randomised controlled clinical trial, group sequential design and multi-arm multi-stage trials in connection with confirmatory studies.

Text Books

1. Berry SM, Carlin BP, Lee JJ, and Muller P (2010). Bayesian Adaptive Methods for Clinical Trials. CRC press.
2. Rosenbaum SE (2012). Basic Pharmacokinetics and Pharmacodynamics: An Integrated Textbook and Computer Simulations. John Wiley & Sons.

Reference Books

1. Shein-Chung Chow, and Jen-Pei Liu. (2013). Design and Analysis of Clinical Trials: Concepts and Methodologies, 3rd Edition. Wiley.
2. Tom Brody (2016). Clinical Trials: Study Design, Endpoints and Biomarkers, Drug Safety, and FDA and ICH Guidelines. Elsevier.

AST 625: Statistical Machine Learning — Credit 3

Introduction

The course provides a broad but thorough introduction to the methods and practice of statistical machine learning and its core models and algorithms.

Objectives

The aim of the course is to provide students of statistics with detailed knowledge of how Machine Learning methods work and how statistical models can be brought to bear in computer systems not only to analyze large data sets, but also to let computers perform tasks, that traditional methods of computer science are unable to address.

Learning Outcomes

After completing the course, students will have the knowledge and skills to: i) Describe a number of models for supervised, unsupervised, and reinforcement machine learning, ii) Assess the strength and weakness of each of these models, iii) Know the underlying mathematical relationships within and across statistical learning algorithms, iv) Identify appropriate statistical tools for a data analysis problems in the real world based on reasoned arguments, v) Develop and implement optimisation methods for training of statistical models, vi) Design decision and optimal control problems to improve performance of statistical learning algorithms, vii) Design and implement various statistical machine learning algorithms in real-world applications, viii) Evaluate the performance of various statistical machine learning algorithms, ix) Demonstrate a working knowledge of dimension reduction techniques. Identify and implement advanced computational methods in machine learning.

Contents

Statistical learning: Statistical learning and regression, curse of dimensionality and parametric models, assessing model accuracy and bias-variance trade-off, classification problems and K-nearest neighbors.

Linear regression: Model selection and qualitative predictors, interactions and nonlinearity.

Classification: Introduction to classification, logistic regression and maximum likelihood, multivariate logistic regression and confounding, case-control sampling and multiclass logistic regressionl, linear discriminant analysis and Bayes theorem, univariate linear discriminant analysis, multivariate linear discriminant analysis and ROC curves, quadratic discriminant analysis and naive bayes.

Resampling methods: Estimating prediction error and validation set approach, k-fold cross-validation, cross-validation- the right and wrong ways, the bootstrap, more on the bootstrap.

Linear model selection and regularization: Linear model selection and best subset selection, forward stepwise selection, backward stepwise selection, estimating test error using mallow's Cp, AIC, BIC, adjusted R-squared, estimating test error using cross-validation, shrinkage

methods and ridge regression, the Lasso, the elastic net, tuning parameter selection for ridge regression and lasso, dimension reduction, principal components regression and partial least squares.

Moving beyond linearity: Polynomial regression and step functions, piecewise polynomials and splines, smoothing splines, local regression and generalized additive models.

Tree-based methods: Decision trees, pruning a decision tree, classification trees and comparison with linear models, bootstrap aggregation (Bagging) and random forests, boosting and variable importance.

Support vector machines: Maximal margin classifier, support vector classifier, kernels and support vector machines, example and comparison with logistic regression.

Text Books

1. James G, Witten D, Hastie T and Tibshirani R (2013). An Introduction to Statistical Learning: with Applications in R, *1st edition*. Springer.
2. Hastie T, Tibshirani R and Friedman J (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction, *2nd edition*. Springer.

Reference Books

1. Masashi Sugiyama (2016). Introduction to Statistical Machine Learning. Elsevier Inc.

AST626: Big Data Analytics — Credit 3

Introduction

This course is designed to introduce key computational concepts, tools and techniques for curating, managing, and analyzing data of large volume, various types and different frequencies. This course assumes basic exposure to the concepts of artificial intelligence, machine learning algorithms and computer programming.

Objectives

The principal aim of this course is to introduce Big Data and its characteristics and challenges to students, and to teach them appropriate tools for managing and analyzing such large-scale data. Further objectives include helping students understand applications of big data in different fields and also ethical issues related to Big Data.

Learning Outcomes

The principal aim of this course is to (i) introduce Big Data and its characteristics and challenges to students, (ii) teach them appropriate tools for managing and analyzing such

large-scale data, (iii) help students understand applications of big data in different fields and also ethical issues related to Big Data.

Contents

Introduction to Big Data: definition, characteristics and applications of Big Data in various fields.

Data pre-processing: data collection and extraction – scraping data, data cleaning- handling missing values, noisy data and outliers; redundancy and correlation analysis, tuple duplication, conflict detection and resolution. Structured and unstructured data and databases - relational and NoSQL databases. Data reduction– overview, Wavelet transformation‘, Attribute Subset Selection, Data Cube Aggregation; Data Transformation and Discretization.

Introduction to Big Data Analytics: techniques to address data analysis issues related to data volume (Scalable and Distributed analysis), data velocity (High-Speed Data Streams), Data Variety (Complex, Heterogeneous, or Unstructured data), and Data Veracity (Data Uncertainty).

Database management essentials for Big Data organization and manipulation: Introduction to data organization (lists, queues, priority queues, trees, graphs, hash). Basic graph models and algorithms for searching, shortest path algorithms, flow networks, matching. Processing and streaming Big Data, introduction to data architecture software including MapReduce, Hadoop distributed file system, Spark, Terradata, and how these tools work.

Data Analysis and Visualization Techniques: Descriptive statistics, probabilistic modeling of Big Data (e.g., graphical models, latent variable models, hidden Markov models.) Bayesian Inference (e.g., variational inference, expectation propagation, sampling.) Bayesian Machine Learning (e.g., Bayesian linear regression). Fundamentals of data visualization, Infographics, layered grammar of graphics. Introduction to Modern, mosaic plots, parallel coordinate plots, introduction to GGobi data visualization system, linked plots, brushing, dynamic graphics, model visualization.

Big Data Ethics and Privacy: Ethical considerations in data collection and analysis, privacy and security concerns in Big Data, legal and regulatory frameworks for Big Data.

Textbooks

1. Balusamy B, Nandhini AR, Kadry S, and Gandomi AH (2021). Big Data: Concepts, Technology, and Architecture. Wiley.

Reference Books

1. Li K-C, Jiang H, Yang LT, and Cuzzocrea A (2015). Big Data: Algorithms, Analytics, and Applications. Chapman & Hall/CRC.
2. Erl T, Khattak W, and Buhler P (2016). Big Data Fundamentals: Concepts, Drivers & Techniques. The Prentice Hall.

AST627: Advanced Statistical Machine Learning — Credit 3

Introduction

The course is typically designed for students who have a basic understanding of mathematics, programming, and machine learning concepts. The course provides a broad but thorough introduction to the methods and practice of advanced statistical machine learning and its core methods, models, and algorithms.

Objectives

The aim of the course is to provide students of Applied Statistics and Data Science with detailed knowledge of how advanced Machine Learning methods work and how statistical models can be brought to bear in computer systems not only to analyze large, high-dimensional, unstructured, and big data sets but also to let computers perform tasks efficiently, that traditional methods of statistical and computer science are unable to address. Students will understand the underlying theory and perform assignments that involve a variety of real-world datasets from a variety of domains. They will learn recent statistical techniques based on a synthesis of resampling techniques, and neural networks that have achieved remarkable progress and led to a great deal of commercial and academic interest.

Learning Outcomes

After successful completion of the course, students are expected to (i) Describe a number of advanced machine learning techniques including deep learning, neural network and reinforcement machine learning, (ii) Assess the strength and weaknesses of each of these models, (iii) Know the underlying mathematical relationships within and across statistical learning algorithms, iv) Identify appropriate statistical tools for a data analysis problems in the real world based on reasoned arguments, v) Develop and implement optimisation algorithms for advanced models, (vi) Design decision and optimal control problems to improve performance of statistical learning algorithms, (vii) Design and implement various advanced statistical machine learning algorithms in real-world applications, (viii) Have an understanding of how to choose a model to describe a particular type of data, ix) Evaluate the performance of various advanced statistical machine learning algorithms.

Contents

Overview of the techniques of Artificial Intelligence (AI), advanced statistical machine learning, and data mining. Overview of supervised and unsupervised learning techniques, and sources of big data (such as social media, sensor data, and geospatial data).

Unsupervised Learning: clustering techniques (k-means, hierarchical clustering, DBSCAN) and dimensionality reduction methods (principal component analysis, t-SNE), multidimensional scaling.

Moving beyond linearity: Polynomial regression and step functions, piecewise polynomials and splines, smoothing splines, local regression, and generalized additive models.

Ensemble methods: Decision trees, pruning a decision tree, classification trees and comparison with linear models, bootstrap aggregation (Bagging) and random forests, boosting and variable importance, AdaBoost, XGBoost, and lightGBM.

Support vector machines: Maximal margin classifier, support vector classifier, kernels, and support vector machines, example and comparison with logistic regression.

Deep Learning: deep neural networks, feedforward neural networks, convolutional neural networks (CNNs), recurrent neural networks (RNNs), and deep reinforcement learning, TensorFlow or PyTorch.

Applications of machine learning in natural language processing: recurrent neural networks, backpropagation through time, long short-term memory, attention networks, memory networks.

Reinforcement Learning: Overview of the basics of reinforcement learning algorithms.

Probabilistic Graphical Models.

Textbooks

1. James G, Witten D, Hastie T and Tibshirani R (2013). An Introduction to Statistical Learning: with Applications in R, *1st edition*. Springer.
2. Hastie T, Tibshirani R and Friedman J (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction, *2nd edition*. Springer.

Reference Books

1. Erl T, Khattak W, and Buhler P (2016). Big Data Fundamentals: Concepts, Drivers & Techniques, Prentice Hall.
2. Goodfellow I, Bengio Y and Courville A (2016). Deep Learning. MIT Press.

AST 640: Oral — Credit 3

Each student must be examined orally by a committee of selected members at the end of the academic year.
