Course Title: APSTA-GE 2122: Applied Statistical Modeling and Inference

Number of Credits: 3

Meeting Pattern: 3 hours total split across two meetings per week.

Course Description (~250 words or less):

This is a course in intermediate and advanced statistical inference techniques in the context of applied research questions in data science. Assuming some prior exposure to probability and statistics, this course will first cover topics such as exploratory data analysis and visualization, principles of estimation and hypothesis testing and the general and generalized linear models, including scientific computation (e.g., E-M Algorithm, Newton-Raphson, and Monte Carlo techniques). These topics are followed by recent developments in model selection and Bayesian modeling. The student will be expected to understand the mathematical theory, implement related statistical algorithms in statistical programming language such as R, and interpret models and parameters in the context of applied statistical analysis of real data.

Course Prerequisites/Expectations:

- REQUIRED: an introductory course in probability and statistics, and basic programming skills.
- NOTE: this course covers material at a rapid pace and has significant outside of class assignments.

Learning Objectives:

By the end of the course, students will be able to:

1. Understand various aspects of applied statistical inference and a wide range of statistical models , emphasizing computational implementation using the statistical programming language R.
2. Analyze data and fit models using statistical packages in R and interpret findings from the perspective of an applied statistician collaborating with a substantive researcher.
3. Gain deeper understanding of the statistical theory, specifically likelihood-based inference that underlies statistical practice, primarily through software implementation and simulation approaches.

Course Format: (Lecture, lab, seminar, recitation or combination)

One 2 hour lecture and one 1 hour required lab session per week; Spring offering

Course Outline (list of lectures/topics each session)

Week	Topic	Knowledge blocks	Programming skills	Textbooks
1	Data Exploration	Various central	Data	

	(univariate, multivariate)	tendency/spread measures, robustness. Data mining techniques such as curve smoothing, PCA, multidimensional Scaling, density estimation.	manipulation/visualization	
2,3	Principles of estimation	sufficient statistic, likelihood principle, Maximum Likelihood estimation and information matrix	Newton Raphson type	Casella and Berger Chapters 3, 6, 7
4, 5	Principles of hypothesis test (bivariate, examples of two sample t test, etc.) and simulation	Test statistics, finite sample (exact test, bootstrap), CLT, null distribution. type I and type II error rate, power, multiple HT, FDR, shrinkage.	Simulation/permutation Exact test, bootstrap Construct a program demonstrating power analysis, and false discovery rate	Casella/Berger Ch 8
6	Linear regression module	Least squares, simple linear regression, multiple linear regression, predictive vs explanatory models, collinearity, modeling assumptions	Simple R programming	Gelman + Hill Ch 3+4 And/or Dobson ch6
	Midterm exam (possibly in 1 hour lab)			
7	Linear model with interactions, connection with multi-factor ANOVA, ANCOVA models,	Interpretation, diagnostics, implicit assumptions such as homogeneity of regression coefs. Overfitting.	R package	Gelman/Hill
8,9	Generalized linear models: Logistic regression (and multinomial) and Poisson regression	GLM theory, models for categorical outcomes, OR and RR interpretations	Short Project: Program a ZIP model	Gelman + Hill Ch 5 + 6 And/or Dobson Ch 7 - 9
10	Generalized Estimating Equations	MLE as a GEE, sandwich estimator		Dobson Ch 11 + Liang and Zeger (1986)
11,12	Principles of model selection	Model misspecification; K-L Div. Information criterion, cross-validation, penalty based variable selection procedures and properties (such as LASSO) Graphical LASSO for covariance	R package and some coding.	Burnham and Anderson (see below) Ch 1 – 3 ----------

		matrix selection (useful tool for network explorations)		
13	Probit model, Gaussian mixture model	Data Augmentation; Talk about the Missing data formulation of probit model, Data Augmentation	FINAL PROJECT using: E-M	Van Dyk and Meng (2001)
14	Bayesian Modeling	Bayes' Theorem, Conjugate Priors; Gibbs Sampling; MCMC	Part of Final Project: MCMC	Bayesian Data Analysis Ch 1,2,3,11 Or Gelman+Hill Ch 18

Course Requirements

The grade for this course will be determined as follows:

- 6 problem sets (for a total of 25%), midterm exam (20%) short project (20%), final project (35%)

The problem sets are designed to understand the concepts introduced in class and sharpen computational statistical skills

The projects are designed to help students to conduct a more complicated and complete data analysis that involves models selection, model fitting, parameter estimation and interpretation.

Students are expected to work independently on the projects. They are encouraged to work together on homework to improve computational programming skills, however, each homework needs to be written independently.

Required Readings and/or Text (a partial reading list is acceptable)

Casella and Berger (2001), *Statistical Inference*, Cengage Learning

Kenneth P. Burnham David R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer.

Andrew Gelman, et al. (2013), *Bayesian Data Analysis* (3rd Edition). Chapman and Hall/CRC

Andrew Gelman and Jennifer Hill (20067), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge

Annette Dobson, *An Introduction to Generalized Linear Models*, second edition, Chapman and Hall/CRC

Kyee Liang and Scott Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73 (1): 13-22.

David van Dyk and Xiaoli Meng (2001). The art of data augmentation. Journal of Computational and Graphical Statistics, Volume 10, Number 1,

Academic Integrity:

All students are responsible for understanding and complying with the NYU Steinhardt Statement on Academic Integrity. A copy is available at: http://steinhardt.nyu.edu/policies/academic_integrity.

Students with Disabilities:

Students with physical or learning disabilities are required to register with the Moses Center for Students with Disabilities, 726 Broadway, 2nd Floor, (212-998-4980 and online at http://www.nyu.edu/csd) and are required to present a letter from the Center to the instructor at the start of the semester in order to be considered for appropriate accommodation.