DALHOUSIE UNIVERSITY

Faculty of Science Course Syllabus
Department of Mathematics and Statistics
STAT 4620/5620 Data Analysis
Winter 2020

Instructor: Dr. Joanna Mills Flemming Joanna.Flemming@Dal.Ca

Lectures: MF 11:35pm-12:55pm LSC-Oceanography 03652

Office Hours: W 1:05pm-2:25pm Chase Building 103

Course Description

At the beginning of this course students are required to select a dataset of relevance to their field of study (or interest). Each student must analyze their dataset and prepare a report (due at the end of the term) describing their data, analyses, and findings. Graduate students are required to orally present both their proposal and final report.

This course begins with a thorough description of the multi-disciplinary field of **data science**, making clear the role of **statistics** therein. Issues surrounding **data ethics** and **reproducibility** will then be discussed followed by an extensive review of tools for exploratory data analysis (**EDA**). **Statistical models** will be described commencing with **linear models** (**LMs**) and **generalized linear models** (**GLMs**). Next, **additive** and **generalized additive models** (GAMs) will be introduced followed by their **mixed model** extensions. **Tree-based methods**, **longitudinal models** and **spatial statistics** will be demonstrated with a view to completing their *statistical toolbox*. Emphasis will be placed on understanding model assumptions and method implementation. Real and relevant data sets will be used throughout the course to demonstrate best practices for data analysis. The **R programming language** will be used exclusively.

Course Prerequisites

STAT 3340, STAT 3460, or the instructors consent.

Course Objectives/Learning Outcomes

This course aims to provide (upper level undergraduate and graduate) students with an awareness of important considerations when undertaking data analysis along with working knowledge of a range of statistical methodologies. Students will develop the confidence to perform appropriate data analyses in order to answer scientific questions of interest.

Specific learning outcomes:

- Capacity to recognize important features of data (e.g., heterogeneity, dependence).
- Understanding of zero-inflation, zero-truncation and over/under-dispersion.
- Proficiency with fitting GLMs, GAMs and their extensions.
- Knowledge of hierarchical modelling frameworks and interpretation of random effects.
- Understanding of tree-based methods and longitudinal models.
- Appreciation for the field of Spatial Statistics.
- Working knowledge of the R language and environment for statistical computing and graphics.

Suggested Reference Texts

Generalized Additive Models: An Introduction with R, Second Edition. Simon N. Wood.
Core Statistics. Simon N. Wood.

DALHOUSIE UNIVERSITY

Course Assessment

Component	*Weight (% of final grade)*	*Due Date(s)*
Project Proposal	0% Undergrad / 5% Grad	Feb 3
Presentation	0% Undergrad / 5% Grad	Mar 30
Project	30% Undergrad / 30% Grad	Apr 10
Final exam	30%	Apr 6
Assignments (5)	40% Undergrad / 30% Grad	Jan 17, Jan 31, Feb 14, Mar 6, Mar 20

Conversion of numerical grades to Final Letter Grades follows the Dalhousie Common Grade Scale

A+ (90-100)	B+ (77-79)	C+ (65-69)	D	(50-54)
A (85-89)	B (73-76)	C (60-64)	F	(<50)
A- (80-84)	B- (70-72)	C- (55-59)		

Course Policies

There will be *five* assignments. These will provide students with the opportunity to review the statistical theory and methods discussed in class and apply these techniques to analyze real and relevant datasets. These assignments *must* be completed using R (https://www.r-project.org). Late assignments will not be accepted.

Cell phones and other electronic devices should be SILENCED before class begins.

Course Content

***Indicates that an assignment is due*.**

Week 1	Jan 6th What is Data Science?	Jan 10th Data Ethics and Reproducibility
Week 2	Jan 13th Tools for Exploratory Data Analysis	Jan 17th Tools for EDA*
Week 3	Jan 20th Zero-Inflated and Zero-Truncated Data	Jan 24th Over- and Under-dispersion
Week 4	Jan 27th Linear Models	Jan 31st LMs in R*
Week 5	Feb 3rd *10-MINUTE PROPOSALS*	
Week 6	Feb 10th Generalized Linear Models	Feb 14th GLMs in R*
STUDY BREAK		
Week 7	Feb 24th Additive Models	Feb 28th AMs in R
Week 8	Mar 2nd Linear Mixed Models	Mar 6th LMMs in R*
Week 9	Mar 9th GLMMs and GAMMs	Mar 13th GLMMs and GAMMs in R
Week 10	Mar 16th Spatial Statistics	Mar 20th Tree Based Methods*
Week 11	Mar 23rd Longitudinal Models	Mar 27th Your Statistical Toolbox
Week 12	Mar 30th *15-MIN PRESENTATIONS*	Apr 3rd *15-MIN PRESENTATIONS*
Week 13	Apr 6th *FINAL EXAM*	Apr 10th *FINAL REPORT*