DEPARTMENT OF THE ARMY
UNITED STATES MILITARY ACADEMY
West Point, NY 10996

MADN-MATH 29 May 2024

MEMORANDUM FOR COL Mike Scioletti, Head, Department of Mathematical Sciences, United States Military Academy, West Point, New York 10996

SUBJECT: MA478, Generalized Linear Models, Course End Report, AY24-2

1. Purpose:
The intent of this memorandum is to consolidate course information for MA478, Mathematical Statistics, and to review and assess how the course, as it was conducted in AY24-2, met program and course objectives. The following summary and assessments are based on instructor and cadet feedback.

2. Course Summary.

Substantial changes to MA478 in AY24-2 are highlighted below in yellow.

MA478 introduces statistical modeling beyond that gained in MA376. Students learn statistical models for analyzing quantitative and qualitative data. Methods are generally taught in the generalized linear model (GLM) framework and include binomial, multinomial, count regression, robust regression, time series models and spatial models. In addition, students are also exposed to Bayesian inferential methods. Students are taught how to handle issues such as missing data, outliers, and influential observations. Students focus on understanding advanced modeling techniques and parameter interpretability. The course primarily uses the R statistical programming language.

For most lessons, students were provided a guided lecture that included a sparse note-taking framework for definitions and theorems along with recommended practice problems. Students were also given a daily take home quiz to ensure they retained the critical components of the lesson.

Course grades were based on three graded homework assignments, a midterm exam, a final comprehensive project (group and individual component) and a Term End Exam. In addition, the students were given up to 50 points on their daily take home quizzes. The students got to select their best 10 quizzes to be used for those 50 points.

Fourteen students were enrolled in the course, 13 Applied Statistics and Data Science Majors, and one Systems Engineering major.

The course was divided into 30 lessons that covered the following topics:

Lesson	Date	Topic

1	9-Jan-24	Introduction / Foundations
2	11-Jan	Linear Models: Estimation and Inference
3	17-Jan	Linear Models: Model Diagnostics
4	19-Jan	GLMs and the Exponential Family
5	23-Jan	GLMs: Link Functions and Model Fitting
6	25-Jan	GLMs and Inference
7	30-Jan	GLMs and Variable Selection
8	1-Feb	Binary Data I
9	6-Feb	Binary Data II
10	8-Feb	Binomial and Proportion Response
11	13-Feb	Multinomial Data
12	15-Feb	Multinomial Data: Contingency Tables
13	21-Feb	Count Regression I: Poisson
14	23-Feb	Count Regression II: Neg Binomial
15	27-Feb	Quasi-Likelihoods
16	29 Feb	Overflow / Review
17	5-Mar	Midterm Exam (WPR)
18	7-Mar	Mixed-Models: Random Effects
19	12-Mar	Mixed-Models: Normal Linear
21	14-Mar	GLMM
23	21-Mar	GLMM Part 2 – Temporal and Spatial Models
20	3-Apr	Bayesian Inference: Basics and MCMC
22	5-Apr	Final Project Proposal Briefing
20	9-Apr	Bayesian Inference: INLA
24	11-Apr	GEE Models
25	16-Apr	Generalized Additive Models
26	18-Apr	Project Comp: DROP
28	23-Apr	Final Project Briefings
29	30-Apr	Final Project Briefings
30	7-May	Course Review / TEE Discussion

3. ASDS Student Learning Outcomes Addressed.

MA476 supports two ASDS SLOs and assessed both SLOs for AY24

SLO#1

Develop and conduct experiments or test hypotheses, analyze and interpret data, and use scientific judgment to draw conclusions.

Performance Indicators:

1. Students are able to recall the basic statistical concepts of data analysis, data collection, modeling, and inference.
2. Students are able to design a study, employ statistical inference, and draw conclusions using formal modeling

SLO#3

Formulate or design a system, process, procedure, or program to meet desired needs

Performance Indicators:

1. Students are able to define clear requirements to a problem
2. Students are able to leverage existing packages and tools to solve computational problems.

a. See Student Outcome Analysis form for an in-depth assessment of the above student outcomes. Assessments were made using homework assignments, a midterm examination, and a term end exam. Details of each follow:

- Homework Assignments: There were three individual homework assignments re-enforcing MA478 course concepts via implementation in R along with technical writing. For each assignment, students drafted a professionally written report detailing data exploration, data preparation, model building, and model selection performed. In the first homework assignment, students were also asked theoretical questions to ensure comprehension of basic linear regression concepts. In homework 1, students were asked to examine data from West Point's Indoor Obstacle Course Test and provide a clear and concise understanding of the impact of height on IOCT times. In the second homework assignment, students were given an auto insurance dataset and needed to build appropriate GLMs to determine the probability that a person will crash their car and the amount of money it will cost if the person does crash their car. In the third homework, students were given a dataset containing information on commercially available wines and needed to build appropriate GLMs to predict the number of cases of wine sold.

- Midterm Examination (WPR): The midterm was an individual assignment focused on interpreting and writing out statistical models. The average grade on the Midterm was an 88%, however this was elevated by allowing the students to make corrections on their work. Originally, the midterm average was closer to a 75%. The corrections, though, seemed to be beneficial for most students knowledge of the material. We asked the students to explain what their answer was, why it was wrong, and then explain the correct process for answering the question.

- Term End (EXAM): The Term End Exam this year focused more on model interpretation rather than predictive performance measures. We retained the Kaggle competition, however the predictive performance was only 10/200 points. The students were also asked to produce a hastily written report (no more than 3 pages) explaining their models and interpreting the model parameters. The students struggled on this task due to coding difficulties and time management.

b. Course Objectives. MA478 course objectives supported these student outcomes:

- *Understand generalized linear models and other advanced regression methods*. Students develop a good understanding and ability to use computational methods for GLMs and related statistical techniques while taking this course as demonstrated by their increasing application.

- *Be able to independently learn and appropriately apply advanced statistical techniques*. Students learn the theoretical foundations for a variety of GLMs including ones with Normal, Binary, Binomial and Proportion, Multinomial, and Count responses, along with additional statistical methods including mixed-models, robust models, and panel models. These concepts are reviewed and coded from first principals to enable both understanding and execution. This was demonstrated by their increasingly sophisticated algorithmic development and understanding as the course proceeded. Students leave the course with a broader view of statistical problem-solving techniques and a way to approach problems.

- *Be able to critically read and interpret data science literature that applies advanced statistical techniques and understand ethical issues in statistical models*. Students completed textbook readings for each lesson throughout the course, which was demonstrated through their participation in class, homework assignments, and the group final project. Ethical issues were addressed in the context of model assessment methods along with social concerns that were discussed during the analysis of the Chicago Crime Data.

- *Be able to successfully use R statistical programming software to analyze data.* Students wrestle with the most efficient and effective way to structure, visualize, use and analyze data. Communicating their work, both for reference by an external audience and their “future self”, are major ongoing themes in the course. In addition to code

functionality, they are assessed on this ability to well-document and communicate their work.

- When faced with a real-world problem, be able to select and ethically execute appropriate statistical modeling techniques to gain insight into the problem to help solve it.
    - The course exposes students to multiple real-world problems via the homework assignments, final project, and term end analysis competition. The students wrangle and munge messy data sets and make decisions for data exploration, data preparation, model building, and model selection. They demonstrate persistence and grit while following a data science process to drive data-driven decision making. Students faced ethical dilemmas in analyzing Chicago burglary rates and examining socio-economic covariates that may impact the number of crimes in a region. Students were asked to address this during their report.

Recommendations.

a. *Retain daily quizzes or graded events.* Students stated that having multiple low-stress graded events helped them better retain knowledge throughout the semester.

b. Retain the time series analysis and spatial modeling overview. Offering these techniques helped to reinforce the purpose of Generalized Linear Mixed Models.

c. *Consider a new TEE format.* Currently, while the TEE competition is fun for the students, it is difficult to assess a students ability to build a meaningful inferential model. It ended up testing more the students' ability to manage their time, which is a useful task, but it comes at a cost of being able to assess their ability to build and interpret a meaningful GLM.

The point of contact for this memorandum is the undersigned at nicholas.clark@westpoint.edu.

NICHOLAS J. CLARK
COL, US Army
ASDS Program Director