Data Analysis

Spring Semester, 2023

January 24, 2023

Introduction and Overview

Welcome to AMS 315. This is a second course with AMS 310 or its equivalent as a prerequisite. The course requires that you attend using a relatively high level of computer and computer network. Each of these points is discussed in more detail in the Syllabus, which is posted on the Class Blackboard.

Examinations and Quizzes

Examinations and quizzes will be given through the Respondus Lock-down Browser. Examination and quiz papers will only be accepted through the Respondus browser.

Evaluation

I will use examinations (quizzes, three midterm examinations, and a comprehensive final), projects (two computer projects), and class participation (as recorded on the Class Blackboard) to evaluate you. I will calculate your total points, which is the sum of the points you earned on your examinations and quizzes, the points that you earned on your projects, and your Blackboard participation points. The maximum total points that you can earn will be about 2,400. Examinations and quizzes will account for about 1700 of the total points possible; projects about 385 points; and Blackboard participation about 300 points.

Resources available to you

The Class Blackboard contains links to the recorded Zoom lectures, the pdf files used in the lectures, material summarizing the text (chapter guides and chapter lecture notes for each chapter), supplemental readings, and pre-recorded Zoom videos discussing the major problem types covered in this course. It will also contain the data sets that you must use in your projects, instructions for accessing online tools and components, and information about examinations and projects. I will hold 3 hours of Zoom contact weekly: on Mondays from 2:30 to 4:00 pm and on Wednesdays from 3:30 to 5:00 pm Stony Brook time. The TAs will also hold 2 hours of Zoom or office contact at the times specified on the Class Blackboard. The Student Accessibility Services Center may also be able to provide you with additional resources.

Free academic support services including one-on-one and small group course-based tutoring, one-on-one skill-based tutoring, peer assisted learning (Supplemental Instruction), and public speaking courses are available for you. Learn more about these services by visiting www.stonybrook.edu/tutoring.

Complete Details

Examinations and Quizzes

Syllabus Quiz

There is a quiz worth 25 points on the class syllabus available for taking now and due before January 30 at 11:59 pm. To get credit for this quiz, you must download the Respondus browser and use it to take this quiz. You will not get any points if you do not use the Respondus browser. You will be penalized additional points if you do not use the Respondus browser for the first midterm. There will be additional penalties if you do not use the Respondus browser in subsequent examinations.

Examinations:

There will be three midterm examinations and a final. You must use the Respondus browser to take the examinations. The total maximum grades on these examinations and quizzes will be about 1,700 points or about 71% of the total points possible. The tentative dates of the midterm examinations are below. All times in this syllabus are Stony Brook times. Midterm examinations will be held during the scheduled class time, which is 4:45 to 6:05 pm, Stony Brook time. If you miss one of the midterms, there will be an 80-minute make-up examination on Friday, May 5 at 1 pm Stony Brook time. This examination will be comprehensive. If you miss two or more midterms, my academic advice is to withdraw from the course. The date and time of the final are set by the University. The final will be on Monday, May 15, from 5:30 pm to 8:00 pm, Stony Brook time.

I will allow you an additional 15 minutes to scan your midterm examination paper and submit it through Blackboard. There will be a 25-point bonus for on-time submission of your midterm examination procedure; that is, a submission before 6:20 will get an extra 25 points added to the examination grade. There will be no bonus or late penalty for a submission between 6:21 and 6:30. Late penalties for a midterm examination paper will begin at 6:31 pm for a midterm exam. The penalty will be a 2-point reduction in your grade for each minute late in your submission time as recorded by Blackboard.

A similar procedure will hold for the final. You will get 20 minutes to scan and submit your final examination paper. The final paper is due at 8:20 pm Stony Brook time on May 15. A final examination paper submitted before 8:20 will receive a 50-point bonus. A paper submitted between 8:21 and 8:30 will not receive a bonus and will not be penalized. A paper submitted after 8:31 will receive a 2-point penalty for each minute after 8:31.

There will be no review classes prior to the examinations, as the lectures are videorecorded with text posted on the Class Blackboard. The make-up examination is given for students who were ill on the day of the examination or who had to miss an examination for a job interview or other professional reason. Please do not miss the opportunity to check your progress by taking each midterm examination as scheduled. Do not give in to the temptation to skip an exam because you do not feel prepared. Only about 33% of eligible students take the make-up.

However poorly prepared you feel you are for a midterm, taking the exam will likely be a net benefit to you.

Examinations will be open book and open notes. You may use any calculator that you wish. The pdf file of statistical tables is available on the class Blackboard in the Assignments section. I recommend that you print it out for use in the examinations. This pdf file will also be attached to your examination file. However, in past semesters some students had trouble accessing this file during an examination.

Students with SASC accommodations will start their mid-term examinations at 4:45 pm Eastern time and must submit their papers to Blackboard within the time allocated by SASC for them. They will start their final examination at 5:30 pm Stony Brook time and will have the time SASC allocated for them. There will be the same bonus point and penalty policy applied to the adjusted end time.

The dates and tentative content of the examinations are:

Feb 23: ***Examination One: Chapters 3, 4, 5, 6, and 7***

Mar 30: ***Examination Two: Chapters 11 and 12, in addition to Chapters 3, 4, 5, 6, and 7.***

Apr 27: ***Examination Three: Chapters 8 and 9, in addition to Chapters 3, 4, 5, 6, 7, 11, and 12.***

May 5: ***Make-up examination for students who missed one mid-term: Chapters 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12.***

May 15: ***5:30 pm-8:00 pm, Stony Brook time. Final examination: Chapters 3-12.***

Projects

There will be two projects with a combined maximum value of 385 points, which will be about 16% of the possible total points in the course. The data for your projects will be posted on the Class Blackboard in a Zip file. The file containing your data set will have the last six digits of your Stony Brook identification number in the name of the file. ***You must analyze your assigned data set. If you do not, your score will be zero.*** If your report is plagiarized from another student, both students will be charged with academic dishonesty—the enabler is more guilty than the plagiarizer.

The data for each project will be generated synthetically according to a model that differs for each student. The models are selected because they occur commonly, as noted in the scientific papers posted on the Class Blackboard. You may use the statistical program of your choice to analyze your data.

The first project will be to analyze two different sets of synthetic data using single predictor linear regression and will be worth a maximum of 100 points. One part of this project will require you to merge two data files; this part focuses on data processing issues and dealing with

missing data and will be worth 40 points. The second part of the project will be to analyze data that may not fully meet the assumptions of Chapter 11 and will be worth 60 points.

The second project will be to analyze synthetic data with multiple predictors and to estimate the model that was used to generate the data. It will be worth a maximum of 200 points. The model generating the data may include transformations of the outcome variable, non-linear predictors, and interactions of predictors.

Each student will submit a report in scientific paper format for each project. The grade of a student who does not submit computer projects will be severely impacted. A minimally satisfactory computer score is a sum of 150 points or more (out of 300) in the grades of project 1 and project 2.

The dates of the projects are below. I will liberally grant extensions for project 1. The project 2 due date can be extended a bit, but not much. Please note that the project 2 due date is near the end of the semester.

Thursday, Mar 9:	Project One data posted.
Tuesday, Apr 11:	Project One report is due at 11:59:00 Stony Brook time. I will liberally allow extensions to this deadline.
Tuesday, Apr 18:	Project Two data posted.
Tuesday, May 2:	Project Two report is due at 11:59:00 pm Stony Brook time.

Course Blackboard

Course materials posted to Blackboard will include:

- The syllabus.
  - An introductory video lecture discussing the syllabus and course; please view this video as soon as possible.
  - Zoom videos of class lectures.
  - Pre-recorded Zoom videos of specific problem types.
  - Supplemental readings.
  - Instructions for accessing online tools and components.
  - Information on assignments and assessments.

Supplemental Readings

Scientific papers and other material will be posted on the Class Blackboard. These papers provide case studies and more detailed discussion and illustration of the application of statistical techniques than your text or I can give. They are an important supplement to your studies. You should study them to enhance your understanding of the applicability of the material in the course and how the material of the course is part of quantitative research.

Tentative Schedule Spring 2023 AMS 315
Lectures are from 4:45 to 6:05 pm Eastern Time

Jan 24:	Introduction, Chapter Three, Data Description
Jan 26:	Chapter Four, Probability and Probability Distributions (simulation issues)
Jan 31:	Chapter Five, Inferences about Population Central Values
Feb 2:	Chapter Five, Inferences about Population Central Values
Feb 7:	Chapter Six, Inferences Comparing Two Population Central Values
Feb 9:	Chapter Six, Inferences Comparing Two Population Central Values
Feb 14:	Chapter Seven, Inferences about Population Variances
Feb 16:	Chapter Seven, Inferences about Population Variances
Feb 21:	Chapter Eleven, Linear Regression and Correlation
Feb 23:	***Examination One: Chapters 3, 4, 5, 6, and 7***
Feb 28:	Chapter Eleven, Linear Regression and Correlation
Mar 2:	Chapter Eleven, Linear Regression and Correlation
Mar 7:	Chapter Eleven, Linear Regression and Correlation
Mar 9:	Chapter Twelve, Multiple Regression and the General Linear Model *Project 1 Data Posted*
Mar 14:	Spring recess, no class
Mar 16:	Spring recess, no class
Mar 21:	Chapter Twelve, Multiple Regression and the General Linear Model
Mar 23:	Chapter Twelve, Multiple Regression and the General Linear Model
Mar 28:	Chapter Twelve, Multiple Regression and the General Linear Model
Mar 30:	***Examination Two: Chapters 11 and 12, in addition to Chapters 3, 4, 5, 6, and 7.***
Apr 4:	Chapter Eight, Inferences about More than Two Population Central Values
Apr 6:	Chapter Eight, Inferences about More than Two Population Central Values
Apr 11:	Chapter Eight, Inferences about More than Two Population Central Values *Project 1 Due at 11:59:00 pm Eastern U.S. Time*
Apr 13:	Chapter Nine, Multiple Comparisons
Apr 18:	Chapter Nine, Multiple Comparisons *Project 2 Data Posted*
Apr 20:	Chapter Nine, Multiple Comparisons
Apr 25:	Chapter Nine, Multiple Comparisons
Apr 27:	***Examination Three: Chapters 8, and 9 in addition to Chapters 3, 4, 5, 6, 7, and 12.***
May 2:	Chapter Ten, Categorical Data *Project 2 Due at 11:59:00 pm Eastern U.S. Time*
May 4:	Chapter Ten, Categorical Data
May 5:	***Make-up examination for students who missed one mid-term, 1:00 pm.***
May 15:	***5:30 pm-8:00 pm Eastern US Time. Final examination: Chapters 3-12.***

Course Issues

Prerequisites

The prerequisite for this course is AMS 310 or equivalent. For example, a student who has completed AMS 102, AMS 110 or PSY 201 and received a grade of B or better should be reasonably well prepared for this course. Specifically, you should have some knowledge of probability theory, a basic understanding of the central limit theorem, knowledge of the single sample t-test, and an understanding of the concepts of testing hypotheses and confidence intervals. I will also use basic matrix calculations (such as matrix multiplication and inversion) in the second half of the course and some calculus (differentiation, optimization of a function, and Taylor series approximations). I do expect that you have some familiarity with computing and that you wish to expand your expertise to being able to work with statistical computing packages.

Recommended text:

Ott, R.L., Longnecker, M. (2016). *Statistical Methods and Data Analysis, latest edition*. Pacific Grove, CA: Duxbury. The tables that you will need in your examinations are available on the class Blackboard. You may use earlier editions of this text or another text if you prefer. If you use another text, please make sure that it has the statistical tables that you will need.

Assignments

There are no homework assignments. Past examination problems are in the study guides for each chapter that are posted on the class Blackboard. Past examinations, and their solutions are also posted. You may turn in your solutions for comments from the TAs or me. The TAs will read your solutions and offer suggestions and instruction. They are instructed ***not*** to solve the problem for you. Because of the rampant sharing of homework solutions, homework papers will not be graded or included in the grading process. Example problems will be discussed in lectures. Zoom videos solving specific examination problems are posted on the class blackboard. In lieu of homework, your usage of the class Blackboard will be tracked and is a component in your evaluation.

Zoom Contact Hours (Office Hours)

My Zoom contact hours are on Mondays from 2:30 to 4:00 pm and on Wednesdays from 2:30 to 4:00 pm Stony Brook time. There will be extra office hours the week before each examination. These will be announced on the class Blackboard. My e-mail is Stephen.Finch@stonybrook.edu. The TAs and I will hold office hours using Zoom. If you have confidential issues that you wish to discuss with me, please be aware that I will not have a waiting room for my contact hours. Please arrange an individual Zoom meeting with me to discuss private issues.

Incompletes

Professor Arkin, who is the undergraduate program director in the Applied Mathematics Department, is the only person who can authorize an incomplete grade for an upper division

AMS course. If Professor Arkin has not authorized an incomplete, absence from the final may result in failure of this class.

Since this class can be offered as part of the training requirements of the Society of Actuaries, academic integrity issues are fundamental. Academic integrity standards will be enforced to the best of our ability.

Grading

The target grade distribution is roughly 25% A, 25% B, and 25% C+, and the remainder C or lower. That is, the target course GPA is 2.75. An examination score in the upper quartile is roughly an A grade; a score in the second quartile is roughly a B grade; a score in the third quartile is roughly a C+ grade; and a score in the lowest quartile is a C or lower grade.

Your final grade will be determined using your final examination grade and your total point score. Your total point score is based on three components. One is your computer project component, which is the sum of the scores received for the two projects. The second is the examination component, *E*, which will be calculated by the sum of your examination and quiz scores:

$$E = E_1 + E_2 + E_3 + E_F + Q.$$

Each in-class examination will have about 6 questions, each worth about 50 points. Each midterm examination will be worth approximately 300 points. The final examination will have about 14 problems and will be worth approximately 600 points.

The third component *P*, class participation, is based upon your usage of the class Blackboard. The class participation component is the sum of four components:

$$P = P_1 + P_2 + P_3 + P_{Cumulative},$$

where P_i is your class participation score for the *i*th examination and $P_{Cumulative}$ is your class participation for the whole semester. The relative weight of the participation scores are $0 \leq P_i \leq 60$ and $0 \leq P_{Cumulative} \leq 120$.

Your total point score is the sum of your examination scores, project scores, and class participation score:

$$TP = E_1 + E_2 + E_3 + E_F + Q + CP_1 + CP_2 + P.$$

Total point boundaries for an *A*, *B*, *C+*, *C*, *D*, and *F* grades will be set. Similarly, final examination boundaries for *A*, *B*, *C+*, *C*, *D*, and *F* grades will be set. Each student will have a total point grade and a final examination grade. A student with satisfactory computer projects will have a course grade that is the higher of the final examination grade and the total point grade. The grade of a student without satisfactory computer projects scores will be the total point grade.

That is, if you have at least minimally satisfactory work on the computer reports, a strong performance on the final will be a major factor in the final grading decision. A student with at least minimally satisfactory computer performance who gets an *A* on the final gets an *A* on the course; a student with satisfactory computer performance who gets a *B* in the final gets at least a *B* in the course; and so on.

AMS 315
Data Analysis
Spring Semester, 2023
Lecture 1

Human Subjects Research

- Must have institutional approval (IRB approval, Institutional Review Board)
- Participants must give informed consent; the researchers must explain to the participant the goals of the research and the risks.
- Participants may withdraw from the study at any time.

Five Most Important Contributions of Statistics

1. Randomized Experiment (Clinical Trial)
2. Genetic Statistics
3. Methods to Analyze Observational Studies
4. Quality Control
5. Opinion Surveys Using Random Samples

Let me discuss each of these in turn, give example studies that use each of these contributions, and describe what chapters in the text address these contributions.

1. Randomized Experiment:

In a randomized experiment, a difference in response between the group given an experimental treatment and the group given a control treatment is ***CAUSED*** by the experimental treatment or is a chance event.

What is randomization of assignment of participant to treatment group?

The two groups are roughly balanced on all independent variables, whether important or not.

Examples of Randomize Experiments

- ❖ Covid vaccine trials: 40,000 participants in one study; vaccine protected against Covid; vaccine did not cause side-effects.
- ❖ Scandinavian Simvastatin Survival Study
  - Tobert, Nature Reviews, 2003 (Class Blackboard).
  - 4,444 participants with coronary heart disease with a range of cholesterol levels on a lipid lowering diet in a 5-year study.

- 2,221 randomly selected received Simvastatin; 2,223 received placebo in double blind study.
- Simvastatin participants had significantly lowered coronary deaths and any cause deaths.

❖ Fuller discussion in Chapters 6, 8, and 9.

2. Genetic Statistics

Not covered in this course.

❖ Example 1: statistics identified the gene in the mechanism for familial hypercholesterolemia, which led to the development of the statin class of medicines. See Tobert review paper and Brown and Goldstein Nobel Prize speech.

❖ Example 2: Researchers have found genetic associations with susceptibility to Covid and risk of serious outcomes.

3. Methods to Analyze Observational Studies

Chapters 10, 11, and 12

In an observational study, correlation shows association, not necessarily causation.

Example of Observational Study

❖ Reported association of prior mononucleosis infection and subsequent contraction of multiple sclerosis.

4. Quality Control

Chapters 5 and 6

5. Opinion Surveys Using Random Samples

Chapter 10.

Please follow a news source (New York Times, Wall Street Journal, Newsday, CNN, …) regularly to identify stories and issues that are relevant to this course.