Introduction to Business Statistics

Instructor and TA

Robert Stine	444 Huntsman Hall	stine@wharton
Justin Khim	F96 Huntsman Hall	jkhim@wharton

Professor Stine's in-person office hours are on Tuesday afternoon, from 3:30-5:30pm. Evening office hours will be held online using BlueJeans; BlueJeans is similar to Skype and available to students. Mr. Khim will have office hours on Tuesday and Thursday from 4:30–6:00pm. In general, course related questions should be posed using Piazza rather than via regular email; you can register to use Piazza via the course Canvas page.

Overview

This course focuses on the statistical technique known as regression analysis. Regression analysis is used in business, economics, health, politics, and the sciences. Regression analysis is perhaps the most important and heavily used methdology in statistics. Its popularity arises from its ability to find patterns, provide interpretable insights, and make predictions. The ideas that underlie regression analysis form the basis for recent methods used in predictive modeling, particularly neural networks that underlie deep learning.

The essential prerequisite for this course is familiarity with the contents of Statistics 101. Topics from Stat 101 include data displays (boxplots, histograms, quantile plots, and scatterplots), summary statistics (such as the mean, standard deviation, and correlation), and properties of statistical estimates and tests (including standard error, confidence intervals, p-values). The first two weeks of Stat 102 quickly review these concepts, focusing on applications and assumptions. If you need to refresh your knowledge, consult materials from Stat 101 and the textbook.

At each class, I expect that you will have read assigned portions of the textbook. Classes focus on a four-step paradigm for answering questions from data:

Frame the problem (motivation)
    What's the question I'm trying to answer? How can data help?

Pick the appropriate statistics (method)
    Which statistics should I compute? Do I have a model for the problem?
    Are assumptions well-matched to the problem?

Perform the data analysis (mechanics)
    Software does the heavy lifting, but you need to know how to use it correctly
    and interpret of the resulting output.

Interpret the results (message)
    If you cannot explain the results to someone who understands the question, but
    doesn't know statistics, then no one is going to pay attention to you.

Software does the computations, but the key role of a data scientist is to identify an appropriate approach and to interpret its results within the context of the problem. The software itself is not a main focus of the course. I will use JMP for classroom examples. You're welcome to use something else (such as Excel, StatCrunch or R), but neither I nor the course TA can be expected to offer help with that choice.

Computers in the classroom

Occasional attendance quizzes require you to access Canvas from your device. A smartphone, tablet, or laptop works fine. Otherwise, you are allowed to use a laptop only while sitting in the back row of the classroom. Tablets for note taking are allowed in any class location.

Classroom behavior

Coming and going from class during lecture is disruptive. Please avoid doing so. During exams, you must stay in the room for the duration of the exam; you may leave early when you finish. Hats are not allowed during exams.

Grading

Your grade for this course is composed of graded homework assignments (MyStatLab), attendance in class, and three in-class midterm exams.

Assignments	20%
Attendance	10%
Midterm exam 1	20%
Midterm exam 2	25%
Midterm exam 3	25%

Assignments are due roughly weekly and are graded on-line; the schedule and problems of of assignments are available through the MyLab and Mastering link on Canvas. Note that there is no credit for late assignments; they must be completed by the shown date. In general, each question allows 3 attempted answers.

Attendance will be sampled through occasional interactive class activities; you will need to have a Canvas-linked device with you at class. All questions about grades on exams must be resolved within 1 week of posting grades.

Midterm exams will be held during the scheduled classroom times during the semester. The course calendar has the specific dates. If your current plans conflict with those dates, consider changing your plans. You will not be excused from an exam due to a conflict with a job interview or other personal events. Makeup exams will be available in the event of a documented medical emergency; in such cases, you must let me know *prior to* your scheduled exam that you will not be able to take the exam.

Textbook and Software

Stine and Foster, *Statistics for Business: Decision Making and Analysis, Third Edition*. Pearson. You will need MyLab access as offered in Stat 101.

JMP, Version 12 or newer

Planned schedule of lectures and exams

Date	Topic	Reading
Jan 10	Overview	Ch 3-6, 11-12
Jan 15	Martin Luther King Jr. Holiday	
Jan 17	Standard error and sampling distributions	Ch 13-14
Jan 22	Confidence intervals and tests	Ch 15-16
Jan 24	Two-sample comparison	Ch 17
Jan 29	Chi-squared, dependence, and Simpson's paradox	Ch 5, 18 (-18.4)
Jan 31	Linear patterns	Ch 19
Feb 5	Residual analysis	Ch 19
Feb 7	Data transformations	Ch 20
Feb 12	Review	Midterm 1
Feb 14	**Mid-term Exam 1**	
Feb 19	Finding patterns in data	Ch 20, 27
Feb 21	Simple Regression Model (SRM)	Ch 21
Feb 26	Inference and prediction in the SRM	Ch 21
Feb 28	Regression diagnostics	Ch 22
Mar 5	Spring Break	
Mar 7	Spring Break	
Mar 12	Multiple Regression Model (MRM)	Ch 23
Mar 14	Inference in multiple regression	Ch 23
Mar 19	Collinearity in multiple regression	Ch 24
Mar 21	Building regression models	Ch 24
Mar 26	Review	Midterm 2
Mar 28	**Mid-term Exam 2**	
Apr 2	Categorical variables in regression	Ch 25
Apr 4	Interactions in regression	Ch 25
Apr 9	More complex categorical features	Ch 25
Apr 11	Trends in time series	Ch 27
Apr 16	Models for time series	Ch 27
Apr 18	Classifiers and regression	Handout
Apr 23	Review	
Apr 25	**Mid-term Exam 3**	