STAT2183: Intermediate Statistical Packages

Instructor Information

Instructor: Dalila Benachenhou
Email: dalila_ben@gwu.edu
Meeting Time: TBA

Teacher Assistant: TBA

Course Description

Computers are taking over the world, what are you doing about it? Library of Congress has over 235 Terabytes of data, Facebook stores and analyzes over 30+ Petabytes, Akamai analyzes 75 million events per day to better target advertisement. In fact, the US government thinks that data storage and analysis is so important the Obama administration is investing $200 million in big data research project. While you will be working with small data, the evidence driven methodology, and some statistical technics you will learn in this course are still applicable to big data.

Throughout the course, you will be learning to compose a compelling argument with data. You will be asked to answer such questions as “Is Bernie Madoff’s hedge fund performance too good to be true?” or “Does a gender bias exist for applying to certain programs within a university?” or “Do trading algorithms increase market volatility?” Evidence driven arguments will help you debunk myths, and prejudices. This course will help you use computers in increasing data prevalence to develop a better understanding of the world’s problems, while providing you with beneficial insights in less time. No longer will you be forced to make judgments solely on assumptions and opinions. In fact, making evidence driven arguments will help you improve a company’s internal policies, product quality, government oversight; it will also help create a more appealing product for consumers. The advantage is apparent in many fields from government to physics.

Learning Objectives

By the end of the course, you will know how to make evidence driven decisions to improve your environment. You will hone your writing, critical, and analytical skills, while acquiring experience in both programming, and statistics. By the end of the course you will be able to:
• Integrate statistics, programming, knowledge about your topic, and writing skills to bring value of your work to your boss, clients, community, or society.
• Apply statistical approaches to the appropriate data set.
• Comfortably use statistical tools to help you speed up your analysis
• Use appropriate help resources in the use of statistical tools
• Apply your knowledge to your core field of inquiry
• Compose professional and actionable reports to your boss or clients

Course Prerequisites

One course in statistics: STAT 1051, STAT 1053 or equivalent.
Knowledge of:
- Central limit Theorem and its implication
- Normal distribution
- p-value
- Confidence limit (confidence intervals)

Textbooks

1. Introduction to Statistical Methods and Data Analysis
   Authors: R. Lyman Ott, and Michael Longnecker
   Publisher: Brooks/Colde, 6th Ed
2. SAS Statistics by Example
   Author: Ron Cody
4. R In Action, 2nd Edition,
   Authors: Robert Kabacoff Peter Dalgaard

Assessment of Learning

In-Class
You will be assessed before and after class. You will work on at least 25 case studies using relevant datasets. Before the end of the class, you will be expected to upload your class work to Blackboard. You
will help each other review your reports. You will discuss if a statistical approach is appropriate, and implication of your analysis to your company, your clients, your community, etc. The purpose of in-class exercises is to provide you with a supportive environment to learn evidence driven reasoning, to provide you exposure to various data, and to improve your programming skills for speedy reasoning. Any work done in class has to be uploaded by the night to Blackboard to be counted.

**Before and After Class**
You will keep a Learning Journal. Before class, after watching videos and reading about statistical approaches, and finding the function or procedure in the statistical packages, you will write down basic knowledge needed to perform your evidence driven reasoning. At the beginning of the semester, I will direct you. You will also keep track of observation. You will also keep track of news, advertisements, articles, sayings were evidence driven reasoning could be applied. At the end of the semester, you will write a 1 or 2 pages summary and insight about what you learned.

**Reports**
Throughout the course, you will have to write 3 reports. These reports will have you: set up the inquiry, perform a reconnaissance, analyze, discuss the results and write a conclusion—all important steps to evidence driven reasoning. The report will have professional look and feel, the methodology used will be appropriate, and the conclusion will add value to the reader (boss, clients, government, or community.)

**Midterm Exam**
You will have a midterm covering the process of evidence driven reasoning, hypothesis testing and categorical data analysis. I will direct you in your report-writing problem.

**Final Exam**
You will apply what you learned on a novel data set and communicate your findings in a report. Your report will have an introduction, method and result, a discussion, and a conclusion.

**Grading**
The weight of each activity is in the following table.

<table>
<thead>
<tr>
<th>Activities</th>
<th># Activities</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Class Work/Quiz</td>
<td>22</td>
<td>15%</td>
</tr>
<tr>
<td>Paper</td>
<td>1</td>
<td>15%</td>
</tr>
<tr>
<td>Reports</td>
<td>3</td>
<td>30%</td>
</tr>
<tr>
<td>Midterm</td>
<td>1</td>
<td>20%</td>
</tr>
<tr>
<td>Final</td>
<td>1</td>
<td>20%</td>
</tr>
</tbody>
</table>
Your final grade will be based on the following grade scale. The final grade scale reflects how well you met the learning objective.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>97-100</td>
</tr>
<tr>
<td>B+</td>
<td>87-89</td>
</tr>
<tr>
<td>C+</td>
<td>77-79</td>
</tr>
<tr>
<td>D+</td>
<td>67-69</td>
</tr>
<tr>
<td>F</td>
<td>0-59</td>
</tr>
<tr>
<td>A</td>
<td>93-96</td>
</tr>
<tr>
<td>B</td>
<td>83-86</td>
</tr>
<tr>
<td>C</td>
<td>73-76</td>
</tr>
<tr>
<td>D</td>
<td>63-66</td>
</tr>
<tr>
<td>A-</td>
<td>90-92</td>
</tr>
<tr>
<td>B-</td>
<td>80-82</td>
</tr>
<tr>
<td>C-</td>
<td>70-72</td>
</tr>
<tr>
<td>D-</td>
<td>60-62</td>
</tr>
</tbody>
</table>

**Schedule**

Full schedule will be posted in Blackboard. First day of school August 29th, 2016

**Note:**

You will work on 5 themes. Each theme is based on a specific statistical topic. Each theme will last between 2 and 3 weeks. In themes 1 and 2 you will code exclusively in SAS. In themes 4 and 5 you will code exclusively in R. In theme 3, you will code in both SAS and R. Each theme covers a new statistical topic, and short cuts in SAS or R for speedy analysis. They will all build your skills on evidence driven reasoning given the statistical topic.

Before class period, you will read books and watch videos about statistical approaches, and explore how to perform your analysis using SAS or R. You will also keep a learning journal to deepen you understanding of evidence driven reasoning, effective use of tools, and statistical approaches.

**Unit 1: Hypothesis Testing for One Continuous Variable  (January 12th)**

*Why I cannot find my shoe size?*

**Dataset:** One numerical variable

**Comparison:** compare its population $\mu$, or $\sigma$ to some expected value

**Possible test:** t-test, Sign Test

**Plots:** boxplot, qq-plot

**Outcome:**

- Know the steps to Hypothesis Testing
- Confidence in applying parametric and non-parametric statistical approaches to given dataset
• Reinforce the understanding of distribution function, statistics, confidence limit, p-value, and differentiating between population and sample statistics (mean and standard deviation)
• Statistics approaches: parametric: t-test, non-parametric: sign test
• Understand the importance of plots: boxplot, histogram
• Overview of how to write a meaningful conclusion
• SAS: data step, import, proc means, proc ttest, proc univariate, ODS

Unit 2: Hypothesis Testing: Comparison of 2 Continuous Variables
You suspect that a riverside community is releasing semi-treated sewage into a river and this, as a consequence, is changing the level of dissolved oxygen of the river. What do you do?
Dataset: One numerical and one nominal variable with 2 values
Comparisons: compare their $\mu$, or $\sigma$
tests: pooled t-test, unequal-variance ttest, Wilcoxon Rank Sum Test, Equal Variances test
Plots: boxplot, qqplot
SAS: proc transpose, proc ttest, proc npar1way, proc boxplot, ODS, data step
Outcome:
• Know how to build an evidence base reasoning in this particular case.
• Understand the difference and similarity between this theme and theme 1.
• Apply parametric approaches when appropriate
• Plots: boxplot, qqplot, and histogram

Unit 3: AOV, Kruskal Wallis, Tukey
Are the tobacco companies truthful when stating that they have no control over the nicotine level?
Datasets: 2 variables, one numeric, and one nominal or ordinal.
Plots: boxplot, and qqplot
Test: Is there differences in the means among these t populations?
• SAS: PROC ANOVA, PROC NPAR1WAY, PROC GLM
• R: anova
Outcome:
• See the similarity between this and unit 2, and 3
• Expand comparison of means beyond 2 populations
• Expand SAS familiarity

Unit 4: Categorical Data Analysis
Your friend tells you that she doesn't see the importance of seatbelt. In fact, she just saw in the news a person who survived because he wasn't wearing his seatbelt. Does it mean that the state should remove seat belt laws?
Dataset: One or many nominal or ordinal variables
Comparisons: frequencies, and percentages
Possible outcomes: follow a certain distribution, 2 variables are independent, one outcome is more likely than another outcome (odd ratio)
Tests: goodness of fit test, Chi-square test of independence or homogeneity
Plots: pie chart, barchart, mosaic
SAS: Proc Format, Proc Freq, proc gchart, put, and input, if statement, data step
R: attach and load packages, tapply, apply, mosaicplot, basic matrix manipulation, Boolean variables
Outcome:
- Know difference between Theme 3 and Theme 2 and 1
- Know that the variables are all categorical (nominal and ordinal), and you only have frequency as a number.
- Plot: mosaic, bar chart

Unit 5: Regression Model: Linear Regression
Can I predict sales of my product by the amount I spend on advertising on TV, radio, or newspaper?
- Datasets: Except for the response variable which has to be numerical, the other variables can be either numerical, nominal, or categorical
- Plots: qq-plot, matrix scatterplot, scatterplot, influence plot, etc...
- Tests: Anova to test if one of the parameter is different from 0, Anova to compare s complete and a sub model, t test to find if parameters are significantly different from
- R: data.frame, lm, Anova, plot, ...
- Outcome:
  - Know when to design a linear regression model
  - How to diagnose if the model is appropriate
  - How to use the model to make inference or make future predictions
  - Can see the difference between such model and Theme 3.
  - Can differentiate between outliers, high leverage, and influential observations and know how to deal with them.
  - Understand the consequence of collinearity on the performance of the model
  - Can diagnosis and deal with collinearity
  - Understand when a tighter model is more appropriate than a large model
  - Can interpret the parameters in the model

Unit 6: Classification Models: Logistic Regression and K Nearest Neighbors
Does the aggravated level of the crime and the race of a victim has an influence if one gets death penalty?
Outcome:
- Compare models to choose the best model
- Confusion matrix
- Get the probabilities of success
- Derive odd ratios, and odds
- Know how to interpret parameters values
Attendance
To get full value of the course, attendance is required. If you are in class you will:

- Learn to see the big picture on how you can help your company, clients, community, or world.
- Have hands on experience with vary diverse datasets.
- Work on developing your evidence driven reasoning skills.
- Most likely pass the course. The number one reason for failure in the course has been lack of attendance.
- Have a supportive environment where you can try and try until you succeed.

I will not give exams on religious holidays. Please inform me, at least one week in advance, of your absence.

You need to communicate as soon as you know if you intend to be absent from class.

If you decide to be absent for no apparent reason, you will still have to do the case on your own but your grade can be only half of the total grade.

Rules
- All cell phones need to be turned off, and stored.
- We will not accept late homework work.
- Late project will get 1 point penalty for each late day.
- Midterm is 1:15min; if you don’t upload your exam within this time we will not correct it.

Academic Integrity
I personally support the GW Code of Academic Integrity. It states: “Academic dishonesty is defined as cheating of any kind, including misrepresenting one’s own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information.” For the remainder of the code, see: http://www.gwu.edu/~ntegrity/code.html

Support for Students outside the classroom

DISABILITY SUPPORT SERVICES (DSS)
Any student who may need an accommodation based on the potential impact of a disability should contact the Disability Support Services office at 202-994-8250 in the Marvin Center, Suite 242, to establish eligibility and to coordinate reasonable accommodations. For additional information please refer to: http://gwired.gwu.edu/dss/
UNIVERSITY COUNSELING CENTER (UCC) 202-994-5300

The University Counseling Center (UCC) offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems. Services for students include:
- crisis and emergency mental health consultations
- confidential assessment, counseling services (individual and small group), and referrals

http://gwired.gwu.edu/counsel/CounselingServices/AcademicSupportServices

Security
In the case of an emergency, if at all possible, the class should shelter in place. If the building that the class is in is affected, follow the evacuation procedures for the building. After evacuation, seek shelter at a predetermined rendezvous location.

Calendar

<table>
<thead>
<tr>
<th>Events</th>
<th>Due Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report 1 will cover Theme 2</td>
<td>September 23rd</td>
</tr>
<tr>
<td>Report 2 will cover Theme 3</td>
<td>October 25th</td>
</tr>
<tr>
<td>Mid-term</td>
<td>October 20th</td>
</tr>
<tr>
<td>Report 3 will cover Theme 4</td>
<td>November 17th</td>
</tr>
<tr>
<td>Journal Report</td>
<td>December 1st</td>
</tr>
<tr>
<td>Last day of classes</td>
<td>December 12th</td>
</tr>
<tr>
<td>Reading day</td>
<td>December 13th</td>
</tr>
<tr>
<td>Final Exam</td>
<td>TBA</td>
</tr>
</tbody>
</table>