Statistics 3301
Statistical Modeling for Discovery I
3-semester-hour course

Autumn 2018 Syllabus

Instructor:	Dr. Christopher Hans	**When:**	MWF 1:50-2:45
Email:	hans@stat.osu.edu	**Where:**	PO 150
Office:	PO 118 (Pomerene Hall)	**Office Hours:**	Wed. 5:00-6:00
Website:	Carmen		Thur. 1:30-2:30

Grader:	Shuyi Wang	**Office hours:**	See below
Email:	wang.7649@osu.edu	**Location:**	PO 151

Prerequisite: C- or better in Stat 3202. Prereq or concur: Math 2568.

Class format: Three 55-minute lectures per week

Course Description and Learning Outcomes

Statistical models for data analysis and discovery in big-data settings, with primary focus on linear regression models. The challenges of building meaningful models for data are explored, and emphasis is placed on model building and the use of numerical and graphical diagnostics for assessing model fit. Interpretation and communication of the results of analyses is emphasized.

Upon successful completion of the course, students will be able to

1. Formulate regression models that describe relationships between variables and understand the models' statistical foundations

2. Perform a complete regression analysis and communicate the results in both statistical and problem-specific terms

3. Use linear regression methods to build models for large data sets and use the results of the analysis to recommend actions

4. Evaluate and compare different regression models using formal statistical methods and graphical techniques

5. Understand the challenges of regression modeling for data collected over time

Textbook and Other Course Materials

We will use the textbook *Applied Linear Regression, Fourth Edition* (2014) by Sanford Weisberg. An electronic version of the book can be accessed for free through The Ohio State University Libraries at http://bit.ly/1Q4xbLB (although there may only be limited copies available at any given time). Reading and homework will be assigned from the book throughout the semester. Be sure to use the fourth edition (red/orange cover) and not the third edition (green cover). The material in the textbook will be supplemented with additional course notes as necessary throughout the semester. More information about the textbook can be found at http://users.stat.umn.edu/~sandy/alr4ed/.

Students are required to use the R software environment for statistical computing and graphics. R can be downloaded for free at http://www.r-project.org. Instructions for using the software will be given in class. Many students prefer to use RStudio, an IDE designed for use with R. RStudio is available for free at http://www.rstudio.com. Once R has been downloaded and installed, students will also need to install the (free) R package **alr4,** which contains the data sets discussed in the textbook.

We will occasionally use a graphics package called **ggplot2**. Examples of how to use this package to create statistical graphics will be given in class. The book "ggplot2: Elegant Graphics for Data Analysis" (Second Edition) by Hadley Wickham provides a detailed description of the software. If you are on campus, you should be able to download the book for free at the following link: http://www.springer.com/us/book/9783319242750. I recommend reading the book if you are interested in developing your ggplot2 skills, but you are not required to read the book for Stat 3301.

Another useful resource is the book "R for Data Science" by Garrett Grolemond and Hadley Wickham. The book can be accessed for free at http://r4ds.had.co.nz/. This reference many come in handy if you run into trouble working with data in R.

Lecture notes will be posted on Carmen along with the R Markdown source used to generate the notes. You will likely find it helpful to print out these notes ahead of time and bring them with you to class so you can annotate them as we discuss the material. The R Markdown source will contain all of the R code used to generate plots and analyses shown in the notes. You can refer to this R code when doing the homework assignments and studying the course material. The source code will be helpful if you are interested in learning how to use R Markdown (not required for this class).

Data Analytics Learning Center: GTA / Grader Office Hours

Graduate teaching assistants (GTAs) for Stat 3201, 3202, 3301, 3302, 3303 and 4620 will hold their office hours in the Data Analytics Learning Center (DALC) in Pomerene 151. You can meet with the GTA for our course in the DALC during his or her office hours to discuss questions you have about the course material, homework assignments, R, etc.

You are welcome to stop by the DALC when it is open even if it is not currently being staffed by the GTA for our course, e.g. if you are looking for a place to study or work on an assignment for one of the supported courses. If the DALC is staffed by a GTA for another Statistics course when you stop by, he or she will help you if possible, but may not be able to answer all of your questions.

The hours during which the GTA/grader for our course will hold office hours in PO 151 can be found at https://data-analytics.osu.edu/dalc. A complete list of hours during which the DALC will be staffed by GTAs for Statistics Department courses can also be found at that link.

In rare situations due to last minute emergencies, the GTA assigned to the DALC may not be able to attend his or her office hours. If the DALC is closed when the schedule indicates it should be open, we recommend waiting for a few minutes. If no one shows up in a reasonable amount of time, please email your instructor to let us know about the problem. You can also contact your GTA to see about arranging a make-up time to meet.

Assignments

Homework will be assigned regularly throughout the semester, will be due on the dates announced in class, and will be graded. Assignments will consist of a mix of technical questions to assess students' understanding of the statistical models, and questions asking students to perform analyses of data sets. The grade for the data analysis portion of each assignment will be based on both the accurateness and appropriateness of the analysis, as well as the clarity of the description of the analysis and results.

Graphics Lab: There will be one take-home "lab" exercise that can be turned in any time before Fall Break (by Wednesday, October 10). This will be a guided exercise that is intended to introduce you to some of the functionality of the graphics package ggplot2. You will have several weeks to work on the lab assignment, however try to avoid leaving it until the last minute. The lab will be graded as "complete"" or "incomplete". An "incomplete" lab is one where your final output does not match the output specified in the instructions. The lab will count toward 5% of your final grade for the course; a "complete" lab receives full credit, while an incomplete lab receives no credit. No late labs will receive credit. If you upload your materials by the end of the day on Wednesday, October 3 (one week before the due date), I will check your lab to see if it is complete. If it is not complete, you can revise it before the final submission deadline.

Data Analysis Project: There will be an individual data analysis project that will be completed in parts throughout the semester. All students will work on the same data set and analysis. The data analysis project will have three components, with due dates spread throughout the semester:

- Exploratory data analysis (due Monday, September 17)
- Model building and analysis (due near the end of October; date TBA)

- Model checking, refinement and comparison (due near the end of November; date TBA)

Information about each of the three components will be given during the semester.

Exams

There will be two in-class midterms that cover material from lecture, the assigned readings and homework. The first midterm will be held on **Friday, September 28** and the second will be held on **Friday, November 2**.

A cumulative final examination will be given during the university's examination period on **Wednesday, December 12 from 2:00-3:45** in our usual classroom. No early or makeup exams will be given except in cases of documented emergencies (serious illness, etc.). Please make sure you arrange your after-semester December travel plans so that you are able to attend the final exam.

Course Grade Information

The final course grade will be based on homework assignments, the labs, the end-of-term data analysis, the two midterms and a comprehensive final examination. The weights for each component of the grade are:

Homework	Graphics Lab	Midterm 1	Midterm 2	Data Analysis	Final Exam
15%	5%	20%	20%	15%	25%

Statement on Academic Misconduct

It is the responsibility of the Committee on Academic Misconduct to investigate or establish procedures for the investigation of all reported cases of student academic misconduct. The term "academic misconduct" includes all forms of student academic misconduct wherever committed; illustrated by, but not limited to, cases of plagiarism and dishonest practices in connection with examinations. Instructors shall report all instances of alleged academic misconduct to the committee (Faculty Rule 3335-5-487). For additional information, see the Code of Student Conduct http://studentlife.osu.edu/csc/.

Special Accommodations

The University strives to make all learning experiences as accessible as possible. If you anticipate or experience academic barriers based on your disability (including mental health, chronic or temporary medical conditions), please let me

know immediately so that we can privately discuss options. To establish reasonable accommodations, I may request that you register with Student Life Disability Services. After registration, make arrangements with me as soon as possible to discuss your accommodations so that they may be implemented in a timely fashion. SLDS contact information: slds@osu.edu; 614-292-3307; slds.osu.edu; 098 Baker Hall, 113 W. 12th Avenue.

Syllabus Version

8/21/18:   Original version