DS 5740 | Advanced Statistics

Updated 09.01.2022

This is a *living* document. Updates may occur during the semester.

You will be notified of any changes and receive an amended syllabus when they occur.

I. Course Information

Instructor: Alexander Christensen (he/him/his); alexander.christensen@vanderbilt.edu

Office Hours: Tuesdays 2-3pm (Hobbs Building, Room 221); by appointment

TA: Marisa Blackman; marisa.h.blackman@vanderbilt.edu

Office Hours: Thursday 1-2pm (Sony Building, Room 2062)

Class Times: Monday and Wednesday, 9:00-10:15am

Classroom: Sony Building 2071

Learning Management Systems: GitHub; Brightspace

Communication Expectations: The instructor and TA will strive to acknowledge emails within 24 hours of receipt. Some emails may require thoughtful responses and therefore are not always feasible to respond to in 24 hours. Slack is available to discuss more immediate questions. Even if Slack status appears as "online" for the instructor or TA, we may not be available or able to give thoughtful feedback to your request. Whether over email or Slack, an immediate response should *not* be expected, but a *timely* response is the goal. Your TA *may* cc the instructor on any emails they feel necessary. While I will make myself available on weekends, I will not check emails frequently.

Have some feedback? Submit anonymous feedback about the course here

Description: Predicting future outcomes is fundamental to human activity. Understanding how past events lead to current conditions can enables us to make predictions about the future. This course will focus on how we can predictions about the future based on past evidence. Applied uses include forecasting market outcomes (e.g., stocks, company earnings) to human behavior (e.g., emotions, public health policy).

The majority of the course covers time series forecasting to use past performance patterns to make predictions about the future. Key topics in this component includes regression, decomposition, smoothing, ARIMA models, and dynamic covariation. A second component will focus on applying dimension reduction techniques such as principal component analysis (PCA), exploratory graph analysis (EGA), and k-means clustering to survey data. A third component of this course will focus on A/B testing (also referred to as experimental design). In this component, we'll cover how we can manipulate different features of the stimulus/environment to uncover predictions about behavior.

This course will use R statistical software and packages. All examples in the course will be provided through R code and all modeling can be performed using open-source R packages. Accompanying books are free and based in R. We will not cover all of the material in any one book, but they are valuable resources for you to hone your skills and continue your career path.

Prerequisites: DS 5620

Goals of the Course:

1. Understand how forecasting can be applied and what model is most appropriate
2. Identify when and how to apply data reduction and clustering techniques
3. Know how to construct experimental designs that can identify variables that affect outcomes

Tentative Course Schedule

Content, assignments, and due dates are subject to change based on progress through the materials. We will cover, if only briefly, all content but time constraints may limit coverage of certain topics in favor of other topics.

<u>Time Series Forecasting</u>

Week 1 (08.24): Introduction to course

- **First project assigned**: Pick a stock of your choice (email me your stock ticker)
- **See** Project 1 Rubric

Week 2 (08.29 & 08.31): Time series regression models

- **Assignment #1 (due 09.04)**
    - Complete "Week2-Homework.Rmd"

Week 3 (09.05 & 09.07): Time series decomposition

- **Assignment #2 (due 09.11)**

- Complete “Week3-Homework.Rmd”

Week 4 (09.12 & 09.14): Exponential smoothing models

- **Assignment #3 (due 09.18)**
  - Complete “Week4-Homework.Rmd”

Week 5 (09.19 & 09.21): ARIMA models

- **Assignment #4 (due 09.25)**
  - Exercises in 9.11 of FPP3: 2, 6(a-g), 7(a, c-e)

Week 6 (09.26 & 09.28): Dynamic regression models

- **Assignment #5 (due 10.02)**
  - Fried et al. (2022) pipeline (participant 18; outcome = future)
    - Apply TSLM with ARIMA errors, ARIMA with no lag, and ARIMA with lag = 1
    - Compare models and report fit statistics
    - Discuss which model should be used for forecasting

Week 7 (10.03 & 10.05): Prophet model, vector autoregression, and bootstrapping and bagging models

- **Assignment #6 (due 10.09)**
  - Read Freichel & O’Shea (2022)
    - What time series models are used?
    - What variables are examined?
    - Was there a trend (prophet model)?
    - Was there seasonality (prophet model)?
    - What day of the week was negative mood highest (prophet model)?

Week 8 (10.10): Catch-up/time to work on stock forecasting project

- **First project due (10.16)**

Fall break on 10.12-10.14

Dimension Reduction/Clustering

Week 9 (10.17 & 10.19): Parallel Analysis and Exploratory Graph Analysis

- Second project assigned: Apply dimension reduction or clustering to dataset of your choice (see Open Data)
- **Assignment #7 (due 10.23)**
  - Apply parallel analysis with PCA and EGA to {psych}’s `bfi` dataset
  - Report number of components/factors and communities for PCA/PAF and Walktrap and Louvain (respectively)
  - Compare component and network loadings, comment on differences

Week 10 (10.24 & 10.26): Beyond Exploratory Graph Analysis

- **Assignment #8 (due 10.30)**
  - Apply bootstrap EGA (bootEGA) to the Broad Autism Phenotype Questionnaire (BAPQ) dataset
  - Compute structural consistency and item stability metrics
  - Apply measurement invariance to the mother's and father's in the BAPQ dataset
  - Apply Unique Variable Analysis (UVA) to the BAPQ dataset
  - After UVA, recompute bootEGA, structural consistency, item stability metrics, and measurement invariance for the mother's and father's BAPQ dataset

Week 11 (10.31 & 11.02): Dynamic Exploratory Graph Analysis and Clustering

- **Assignment #9 (due 11.06)**
  - Run through "ESM Ergodicity Script.R"
  - How many dimensions were in the population structure? What was the mean and range of the individual structures?
  - Was the personality dataset ergodic? Interpret whether it was and explain what the interpretation means (e.g., can the aggregate structure generalize to each person in the sample?)
  - When performing the cluster test, how many clusters were identified? What does this result mean? Report the statistics from the single cluster test

Week 12 (11.07 & 11.09): K-means and Hierarchical Clustering

- **Assignment #10 (due 11.13)**
  - Apply K-means and hierarchical clustering to the `iris` dataset
  - For K-means, try different number of centers and iterations, report the most accurate parameters
  - For hierarchical clustering, try different distances and methods, report the most accurate parameters
  - Compare K-means and hierarchical most accurate results, report which method performed the best

<u>A/B Design</u>

Week 13 (11.14 & 11.16): A/B Design (part 1)

- **Assignment #11 (due 11.20)**
  - Analyze group activity from Week 1
  - What type of experimental design was followed?
  - Were there any differences between A and B? Report appropriate *t*-test and effect size
  - Perform permutation on the same data. Report your *p*-value and explain what this means. Do your results differ from the *t*-test?

No courses this week – holiday break (11.21-11.25)

Week 14 (11.28 & 11.30): A/B Design (part 2)

- **Assignment #12 (due 12.04)**
    - Analyze data looking at whether an artist's morality affects whether their artwork is perceived as beautiful
    - What type of experimental design was followed?
    - Were there any differences between "good" and "bad" artists? Report appropriate *t*-test and effect size
    - Perform permutation on the same data. Report your *p*-value and explain what this means. Do your results differ from the *t*-test?

Week 15 (12.05 & 12.07): Catch-up/time to work on dimension reduction/clustering project

- **Second project due (12.11)**

Grading: There will be 12 assignments throughout the semester but only your 10 best grades will count toward your final grade.

Assignments: 10 assignments (6 points each)

Projects: 2 projects (20 points each)

Total: 100 points

Letters and signs:

A = 93-100; A- = 90-92

B+ = 87-89; B = 83-86; B- 80-82

C+ = 77-79; C = 73-76; C- = 70-72

D+ = 67-69; D = 63-66; D- = 60-62

F = 0-59

Late work policy: A second after midnight on the Sunday assignments are due (i.e., 12:00:00am on Monday) your grade on the assignment is worth 80% of your overall grade for the assignment. There is only one deadline to turn in all late assignments: the **Sunday** *two weeks* **before** the official last day of the semester (11.20.2022).

Grade appeal policy: The instructor makes all final decisions on grades. The instructor is willing to revisit a grade only in the instance of suspected error. Do not request an appeal based on a disagreement with the instructor's judgment. If you disagree with a grade that you've received on an assignment or project, then you are welcome to submit an appeal to Dr. Christensen. The appeal should be submitted via email and must include: (1) clear and specific

reference to the part of the assignment or project in question and (2) justification for why more credit is earned, citing specific material or evidence. The appeal should be no later than 7 days after the grade has been posted on Brightspace. If, after the appeal you still have concerns, you should raise them with the DGS (Dr. Blocher; jesse.blocher@vanderbilt.edu or Associate DGS (Dr. Kang; h.kang@vumc.org) .

Vanderbilt Honor Code: In accordance with the undergraduate Vanderbilt University Honor Code:

I pledge to pursue all academic endeavors with honor and integrity. I understand the principles of the Honor System, and I promise to uphold these standards by adhering to the Honor Code in order to preserve the integrity of Vanderbilt University and its individual members.

In addition, I abide by responsibility detailed in the Vanderbilt's Faculty Guide to the Honor System. In general, this course is intended to be collaborative and therefore students may and should work together on assignments and projects. Each assignment and project, however, must be a student's own work and submitted separately. All projects must be completed on separate stocks (project 1) and survey datasets (project 2). Each assignment and project should be signed with the shortened version of the Honor Code that you are submitting (either at the very beginning or very end of the materials turned in):

I pledge on my honor that have neither given nor received unauthorized aid on this examination.

Classroom Accommodations

Vanderbilt University and Dr. Christensen are committed to equal opportunity for students with disabilities. If you need course accommodations due to a disability, please contact VU Student Access Services to initiate the process. After SAS has notified me of relevant accommodations, we will discuss how these accommodations may best be approached in this class, and I will facilitate the accommodations.

If emergencies or extenuating circumstances keep you from class, please get notes and announcements from a classmate. You're welcome to arrange a meeting with Dr. Christensen or the teaching assistant to ask questions about the missed material (also check the class's Brightspace and GitHub pages).

Mental Health & Wellness

If you are experiencing undue stress that may be interfering with your ability to perform academically, Vanderbilt's Student Care Network offers a range of support services. The Office of Student Care Coordination (OSCC) is the central and first point of contact to help you navigate and connect to appropriate resources. You can schedule an appointment with the OSCC

at https://www.vanderbilt.edu/carecoordination/ or call 615-343-WELL. You can find a calendar of services at https://www.vanderbilt.edu/studentcarenetwork/satellite-services/. If you or someone you know needs to speak with a professional counselor immediately, the University Counseling Center offers Urgent Care Counseling. Students should call the UCC at (615) 322-2571 during office hours to speak with an urgent care clinician. You can also reach an on-call counselor after hours or on the weekends by calling (615) 322-2571 and pressing option 2 at any time. You can find additional information at https://www.vanderbilt.edu//ucc/.

Names and Pronouns

If you would like to use a different name or different pronouns than those provided through YES, please let me know at any time prior to or during the semester. Information is available through the LGBTQI Life offices about how to change either or both in YES.

Religious Holidays

If you will miss class to observe a religious holiday, please email Dr. Christensen in advance of your planned absence. You may make up (without penalty) any work missed due to the observance of a religious holiday. Please let Dr. Christensen know two weeks prior to the observance so he can make proper arrangements for you. Dr. Christensen has done his best to organize his class schedule around observances outlined here but please let him know if any observances are not listed that apply to you.

Mandatory Reporter Obligations

All University faculty and administrators are mandatory reporters. What this means is that all faculty, including me, must report allegations of sexual misconduct and intimate partner violence to the Title IX Coordinator. In addition, all faculty are obligated to report any allegations of discrimination. I am willing to discuss with you such incidents but can only do so in the context of us both understanding my reporting obligations. If you want to talk with someone in confidence, officials in the Student Health Center, the University Counseling Center, and the Office of the Chaplain and Religious Life (when acting as clergy) can maintain confidentiality. In addition, officials in the Project Safe Center have limited confidentiality, in that they must report the incidents but can do so without providing identifying information. The Project Safe Center serves as the central resource for those impacted by sexual misconduct and intimate partner violence and can assist with navigating all facets of the University's resource and support network and other processes.

Class Technology Policy

Please turn the volume off of ALL electronic devices—cell phones, laptops, tablets, etc.—before you come to class. It can be quite distracting to fellow students and to your instructor if your device starts to play music in the middle of class. A computer is a fundamental part of the course and will be required as part of in-class code examples.

In-class Policy

In alignment with Vanderbilt's academic mission, on-campus academic programs were designed for in-person instruction because it offers an important opportunity for learning, leadership, and the scholarly exchange of ideas. As such, university policy generally requires in-person teaching for undergraduate, professional, and graduate courses (excluding our online programs designed for distance learning). Dr. Christensen will not offer remote or online options unless otherwise mandated by Vanderbilt.

Books (all available for free):

Time Series Forecasting

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (3rd edition). Melbourne, AUS: OTexts. https://otexts.com/fpp3/

Dimension Reduction and Clustering

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd edition). New York, NY: Spring. https://hastie.su.domains/ElemStatLearn/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (2nd edition). New York, NY: Springer. https://www.statlearning.com/

A/B testing

Taback, N. (2022). *Design and analysis of experiments and observational and studies using R* (1st edition). New York, NY: Chapman and Hall/CRC. http://designexptr.org/index.html

Software

R: https://cran.r-project.org/

Rtools (Windows only): https://cran.r-project.org/bin/windows/Rtools/rtools42/rtools.html

RStudio: https://www.rstudio.com/products/rstudio/download/#download

Open Data

Forecasting

Emotions: https://emotedatabase.com/datasets/

Fried et al. (2022): https://doi.org/10.17605/OSF.IO/MVDPE

Suicidality and mood: https://osf.io/dah89/?view_only=656b08a4d8d74c4dad04da3935526601

Dimension Reduction and Clustering

Open Psychometrics: https://openpsychometrics.org/_rawdata/

General

Journal of Open Psychology Data: https://openpsychologydata.metajnl.com/

Open Science Framework: https://osf.io/search/

UCI Machine Learning Repository: https://archive-beta.ics.uci.edu/ml/datasets

Additional Resources

Augmented Dynamic Adaptive Model

Svetunkov, I. (2022). *Forecasting and analytics with ADAM*. Monograph. OpenForecast. Lancaster, UK. https://www.openforecast.org/adam/

Statistics (more on the math than applied side)

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. The MIT Press: Cambridge, MA. https://probml.github.io/pml-book/book1.html

Acknowledgements

All errors in course materials are mine and mine only

Dr. Hyndman and Dr. Athanasopoulos for their forecasting book and materials

Dr. Friedman, Dr. Hastie, Dr. James, Dr. Tibshirani, and Dr. Witten for their statistical learning books

Dr. Taback for their experimental design book and materials

Fried et al. (2022) for making their data open

Beck & Jackson (2022) for making their data open

Freichel & O'Shea (2022) for making their preprint and data open

Dr. Chatterjee and ChatLab at UPenn for making their data open

Marisa Blackman for her support throughout the course