APSTA-GE 2123: Bayesian Inference

NYU Applied Statistics for Social Science Research

Instructor: eric.novik@nyu.edu

21 Jul 2024

1 Course description

This course will introduce students to Bayesian data analysis and Bayesian workflow, a process we follow to develop statistical models, test their efficacy, and make model improvements. This is a deep subject with a rich history, and we can only offer an introductory treatment during a seven-week course. Nonetheless, motivated students should get enough experience from this course to learn how to think like a Bayesian, understand the basics of Bayesian machinery, fit General Linear Models using Bayesian inference, evaluate model quality, and make predictions.

During the course, we will use simulations, the R language, Stan, and the `rstanarm` package.

2 Times, Location, and Office Hours

In-person weekly lectures will be held every Monday from 5 p.m. to 8 p.m. at 194 Mercer St, Room 204, beginning on March 25 and ending on May 06. Office hours will be held online on Thursdays from 5 p.m. to 8 p.m. by appointment.

3 Course prerequisites

Students should be familiar with the R language. Ideally, students will be comfortable with probability and basic differential and integral calculus and have taken a basic regression course. Students don't have to be familiar with Stan or `rstanarm`, and no prior knowledge of Bayes is expected.

4 Course materials

The core text for the course will be *Bayes Rules!* Johnson, Ott, and Dogucu (2022), which is available for free online and as a paper copy from CRC Press.

Students are expected to have access to R and RStudio or another R programming environment. Students should follow the setup instructions from the Getting set up chapter.

For those who are interested in a more in-depth treatment of the subject, you could do worse than the following choices:

- Statistical Rethinking, McElreath (2020)
- A Student's Guide to Bayesian Statistics, Lambert (2018)
- Bayesian Data Analysis, Gelman et al. (2014)
- Stan User's Guide, Stan Development Team (2022)

Aki Vehtari from Aalto University compiled an excellent list of mostly Bayesian texts here.

5 Course Outline

The following tentative schedule assumes seven 3-hour sessions during which we plan to cover the following material.

- **(1) Bayesian Workflow**

  We will briefly discuss the history of Bayesian inference and introduce the key components of the Bayesian workflow — from model development to model testing and decision-making. We will dig into the key components of Bayes's rule and introduce the Binomial model.

  - Bayes Rules!, Chapters 1 and 2
  - Bayesian Workflow Gelman et al. (2020), Chapter 1: Introduction
  - Optional: "Not Only Defended But Also Applied": The Perceived Absurdity of Bayesian Inference, Gelman and Robert (2013)

- **(2) Conjugate models: Beta-Binomial**

  To demonstrate the Bayesian machinery, we will review a few simple models for which analytical solutions are available, particularly the Beta-Binomial. In practice, we seldom use conjugate models, but they have pedagogical value, are very fast, and allow us to validate numerical methods.

  - Bayes Rules!, Chapters 3 and 4
  - Bayesian Workflow Gelman et al. (2020), Chapter 2: Before fitting a model
  - Optional: Statistical Modeling: The Two Cultures, Breiman (2001)

- **(3) Other conjugate models and introduction to posterior sampling**

  This lecture will examine Gamma-Poisson and Normal-Normal models and discuss posterior sampling. We will begin with grid approximations and then show the output from state-of-the-art posterior samplers, such as the NUTS sampler available in Stan. We will introduce the Stan language, demonstrate a few simple models written in Stan, and go over MCMC diagnostics.

  - Bayes Rules!, Chapters 5 and 6
  - Bayesian Workflow Gelman et al. (2020), Chapter 3: Fitting a model
  - Optional: Reflections on Breiman's Two Cultures of Statistical Modeling, Gelman (2021)

- **(4) MCMC, posterior inference, and prediction**

  Most interesting models don't have nice analytic posteriors since most integrals do not have closed-form solutions. Fortunately, Markov Chain Monte Carlo (MCMC) algorithms make it possible to sample from target distributions with arbitrary precision (in infinite time). To understand how MCMC works, we will look under the hood of one of the oldest algorithms, the Metropolis-Hastings algorithm (1953, 1970). We will introduce the posterior predictive distribution.

  - Bayes Rules!, Chapters 7 and 8
  - Bayesian Workflow Gelman et al. (2020), Chapter 4: Using constructed data to find and understand problems

- **(5) Bayesian linear regression and model evaluation**

  Modern MCMC algorithms, such as Hamiltonian Monte Carlo (Homan and Gelman (2014)), allow us to sample from complicated posteriors efficiently. The Stan language implements this method, and we will use it for the remainder of the course to draw samples from posterior distributions. We will also build a linear regression model in the `rstanarm` package, which exposes many Stan models via a familiar R language formula interface. We will discuss what makes a good model.

  - Bayes Rules!, Chapters 9 and 10
  - Bayesian Workflow Gelman et al. (2020), Chapter 6: Evaluating and using a fitted model

- **(6) Expanding the linear model and modeling counts**

  We will discuss linear model expansion and compare model performance using leave-one-out cross-validation (LOO). We will also introduce models for analyzing count data, such as Poisson and Negative Binomial.

  - Bayes Rules!, Chapters 11 and 12
  - Bayesian Workflow Gelman et al. (2020), Chapter 7: Modifying a model

- **(7) Logistic regression and introduction to hierarchical models**

  In the final lecture, we will discuss logistic regression and introduce hierarchical models – the workhorse of Bayesian analysis. We will fit and evaluate hierarchical regression and compare the results to complete pooling and no-pooling estimates.

  - Bayes Rules!, Chapters 13 and 15
  - Bayesian Workflow Gelman et al. (2020), Chapter 5: Addressing computational problems

6 Grading

There will be six homework assignments, an analysis project that can be completed in groups of up to three people, and two in-class quizzes. All assignments must be completed in R Markdown, Quarto, or similar systems suitable for scientific publication. During the final, each team will present the analysis to the class. The presentation should contain at most five slides and take 10 minutes. We will distribute the presentation template and some guidelines for the analysis during the first week of class.

- Six homework assignments: 60%
- Two quizzes: 15%
- Group project and presentation: 20%
- Class participation: 5%

The homework will be assigned at the end of each lecture and due by the following Monday at 5 pm. Since we may be reviewing some homework questions during class, we cannot accept late submissions. The students can discuss the homework problems with each other, but they should not collaborate on writing up the solutions.

The quizzes are designed to test your understanding of the material. They will be administered on two random days during the semester, and a well-prepared student should be able to complete a quiz in about 20 minutes.

The group project is a collaborative exercise, as are most real-world projects. Please refer to the project guidelines. Each team must present their work to get full credit for the analysis project, with 90% weight given to the project and 10% to the presentation.

To earn class participation credit, students must post in the online lecture discussions on Brightspace under the Class Participation topic. For each lecture, please ask a question about the material or the homework, ask a general question about the topic, or respond to the question asked by another student.

7 ChatGPT and Large Language Models

Most of you are familiar with LLMs. They are extremely helpful tools, but they can be helpful and harmful during learning. Please do not use LLMs for your homework (it may give you wrong answers, and you may not know why), but when researching project topics and checking and improving your project code, feel free to use LLMs. When writing up the results of your project, please use your own words — do not use LLMs to create narrative text for you, but if you do, you have to cite it as being produced by the LLM. If you have more questions regarding "fair use," feel free to post in the discussion room on Brightspace.

8 Q&A and After-Class Discussion

We will set up several discussion threads on Brightspace at the beginning of class, including homework Q&A and Project Q&A. All questions regarding class material, homework, and the project should be posted on Brightspace so other students can benefit from the answers. Please don't email the instructor or the Grader about the course material.

9 Participation and Attendance Policies

Students are expected to come to class. The students will participate in live surveys during class, discuss concepts with their peers, and take short quizzes.

10 Other Resources

The following resources may be helpful to those who need a refresher on the prerequisites.

- R for Data Science, Wickham and Grolemund (2016)
- Calculus Made Easy, Thompson (1980)
- Calculus, Herman, Strang, and OpenStax (2016)
- YouTube: Essence of Calculus, Sanderson (2018)
- Intoduction to Probability, Blitzstein and Hwang (2019)
- Introduction to Probability Cheatsheet v2, Chen (2015)

11 Academic Integrity

All students are responsible for understanding and complying with the NYU Steinhardt Policies and Academic Integrity.

12 Students with Disabilities Statement

Students with physical or learning disabilities are required to register with the Moses Center for Students with Disabilities at 726 Broadway, 2nd Floor, (212-998-4980) and are required to present a letter from the Center to the instructor at the start of the semester in order to be considered for appropriate accommodation.

13 Mental Health Statement

If you are experiencing undue personal and/or academic stress during the semester that may be interfering with your ability to perform academically, the NYU Wellness Exchange (212 443 9999) offers a range of services to assist and support you. I am available to speak with you about stresses related to your work in my course, and I can assist you in connecting with the Wellness Exchange. Additionally, if you anticipate any challenges with completing the assignments, readings, exams and other work required in this course, I encourage you to register with the Moses Center (212 998 4980) in advance so that you may be granted the proper academic accommodations.

References

Blitzstein, Joseph K., and Jessica Hwang. 2019. *Introduction to Probability*. Second edition. Boca Raton: crc Press/Taylor & Francis Group.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Gelman, Andrew. 2021. "Reflections on Breiman's Two Cultures of Statistical Modeling." *Observational Studies* 7 (1): 95–98. https://doi.org/10.1353/obs.2021.0025.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2014. *Bayesian Data Analysis, Third Edition*. https://doi.org/10.1007/s13398-014-0173-7.2.

Gelman, Andrew, and Christian P. Robert. 2013. ""Not Only Defended But Also Applied": The Perceived Absurdity of Bayesian Inference." *The American Statistician* 67 (1): 1–5. https://doi.org/10.1080/00031305.2013.760987.

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. "Bayesian Workflow." *arXiv:2011.01808 [Stat]*, November. http://arxiv.org/abs/2011.01808.

Herman, Edwin, Gilbert Strang, and OpenStax. 2016. *Calculus Volume 1*. https://d3bxy9euw4e147.cloudfront.net/oscms-prodcms/media/documents/CalculusVolume1-OP.pdf.

Homan, Matthew D., and Andrew Gelman. 2014. "The No-u-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *The Journal of Machine Learning Research* 15 (1): 15931623.

Johnson, Alicia A., Miles Q. Ott, and Mine Dogucu. 2022. "Bayes' Rule." In, 17–48. Chapman; Hall/CRC. https://doi.org/10.1201/9780429288340-2.

Lambert, Ben. 2018. *A Student's Guide to Bayesian Statistics*. 1st edition. Los Angeles: SAGE Publications Ltd.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in r and STAN*. 2nd edition. Boca Raton: Chapman; Hall/CRC.

Sanderson, Grant. 2018. "Essence of Calculus," November. https://www.youtube.com/playlist?list=PLZHQObOWTQDMsr9K-rj53DwVRMYO3t5Yr.

Stan Development Team. 2022. *Stan Modeling Language Users Guide and Reference Manual*. 2.30 ed. https://mc-stan.org/docs/stan-users-guide/index.html.

Thompson, Silvanus P. 1980. *Calculus Made Easy: Being a Very-Simplest Introduction to Those Beautiful Methods of Reckoning Which Are Generally Called by the Terrifying Names of the Differential Calculus and the Integral Calculus*. 3d ed. New York: St. Martin's Press.

Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. First edition. Sebastopol, CA: O'Reilly. https://r4ds.had.co.nz/.