BIOL 343 – Syllabus

Advanced Data Analysis for Biologists

Course Information

Fall Semester, 2022, September 6 to November 29
3.0 Credits
Hybrid (In-Person and Online)
Pre-requisites: BIOL 243
Lectures Tuesdays, 8:30-10:30, JEFF 126
Tutorials Fridays, 10:30-11:30, MacCory D201

Instructor

Name: Dr. Robert I. Colautti (he/him)
Office Address: 4325 BioSciences, 116 Barrie St.
Office Hours: Tuesdays, 10:30-11:30
Phone: 613-533-2353; Please use only during office hours.
Email: robert.colautti@queensu.ca

About Me: My teaching takes a student-centered approach that is supportive of diverse learners. I set high expectations and provide extensive resources so that student learning extends beyond the timeline course. I try to emphasize a growth mindset that is focused on effort, personal development, and quality of work, rather than mastery or excellence. I try to teach students how to recover and learn from failure, which I believe is essential for a successful career. My teaching philosophy draws on two decades of research and mentorship experience, resulting in primary research in top journals (e.g. Science, PNAS, PRSB) and dozens of former students employed in the public and private sectors or continuing to advanced degrees. I have a broader perspective than many biologists, drawing from an MSc in aquatic ecology from the University of Windsor and a PhD in quantitative genetics at the University of Toronto, followed by bioinformatics and computational biology research at Duke University (North Carolina), the University of British Columbia, and the University of Tuebingen (Germany). I am involved in a wide range of research projects, from COVID-19 and Lyme disease to plant ecology and rapid evolution. As a faculty member in the Biology Department at Queen's University, my goal has been to develop courses that cover everything I wish I had known when I was starting out. The result of this effort is the course material for BIOL 343, BIOL 432, BIOL 860 and BIOL 812.

Website: https://ecoevogeno.org/
Twitter: @ColauttiLab

Important University Dates

Key dates (first day of class, tuition due date, last day to add/drop courses) are important to your academic success. Please find them at Important Dates.

Welcome to BIOL 343!

This course is designed for students who want to develop analytical skills for a career in data science or to analyze data for advanced courses or thesis/dissertation research. These skills are taught primarily through tutorials that have been written specifically for this class, building on the foundations of BIOL 243.

In this course, we focus on the R coding environment and learn how to gain biological insights through data visualization and formal statistical analysis, and then present findings in a formal report. We emphasize a professional mindset, taught through an applied, hands-on approach that will require regular practice and assessments, supplemented with short, conceptual topics in lecture format.

This course covers fundamentals of statistics through application, beginning with frequency distributions, central moments, and summary statistics, followed by linear models and model selection as a basis for exploring more advanced models (Generalized Linear Mixed Models, Generalized Additive Models, and basic Machine Learning).

In contrast to many biology courses, coding and quantitative skills learned in this course must be developed through extensive practice, trial, and failure. The philosophy of this course is that you won't learn to code and analyze data by reading and memorization – only through extensive application and practice. There is no final exam. Some fundamental concepts are explored in the course textbook, which is tested in in class. Regular assignments and quizzes are designed to reinforce skill development and problem-solving. In addition, you are **STRONGLY ENCOURAGED** to find opportunities to practice coding wherever possible. It will take longer at first, but it will save a lot of time in the long run. Finally, be prepared to get frustrated – you will make many errors and most of your coding time will be debugging and searching for answers on the internet. It is important to know that **this is COMPLETELY NORMAL** and self-directed research to solve coding problems is perhaps the most crucial skill you will learn in this course.

Quotes from previous students:

I took this course because I need a statistics prerequisite to apply for veterinary school, and I did not put in the effort required in intro stats to get an adequate grade. I put more effort into this class than any of my other courses this semester, and it did pay off.

We were being tested on week 1 material in week 8.

I found myself referring to the self tutorials online more than the lectures because they were easier to understand and more useful.

A lot of times I had to rewatch videos 2-3 times to understand… and sometimes I just couldn't get it and had to turn to the internet for explanation

The use of self tutorials as well as the live coding lectures were helpful to my learning.

The website of self-tutorial are good, it showd every useful information with a well-organized form, which is better than the zoom lecture. I think prof. Colautti sometimes spend too much time on finding the bug of code.

In some cases the lack of background doing this kind of assigments required much more effort and time from my side which was sometimes quite frustrating even though the answer was not really complicated.

The assignments sometimes were very quick and brief and sometimes took a long time.

The weekly assignments were also extremely time consuming, often taking up to 7-10 hours for each.

… t*he lecture content did not often get entirely covered within the lecture time.*

Equity, Diversity and Inclusivity Statement

We recognize equity and diversity are central to our educational mission and standards of excellence. We are working to dismantle direct, indirect, and systemic discrimination that still exists within our institutional structures, policies and practices -- and in our community. These take many forms and work to differentially advantage and disadvantage persons across social identities such as race, ethnicity, disability, gender identity, sexual orientation, faith and socioeconomic status, among other examples. As students and educators, we have important roles to play to identify and address systemic discrimination for the benefit of all.

Land Acknowledgement

Queen's is situated on traditional Anishinaabe and Haudenosaunee territory. We are so privileged and grateful to be able to be live, learn and play on these lands. Therefore, let us be mindful of the what we learn while on these lands, and how we might apply our newfound skills and knowledge for the benefit of all.

Expectations

For Instructors & Teaching Assistants

As course instructor, I am responsible for developing and editing the course material, which was written specifically for this course and for you, the student interested in data science. I will ensure that all relevant course material is available online and released on a weekly basis so that you know where you should focus your limited time and attention. The course content is a work in progress, so I welcome any feedback on this material, from small spelling/grammar errors to general suggestions for improvement.

To accommodate differences in learning styles, I will make the main content available in complementary forms including an online textbook and pre-recorded videos with annotated scripts. The textbook will lean heavily on a tutorial style, with step-by-step instructions that are reiterated in the online videos. The videos and textbook are designed to be complementary, with overlap emphasizing important skills and techniques. I will also be responsible for recording and posting in-person lectures for those who are not able to attend, or wish to review, the synchronous sessions.

I will make mistakes in these tutorials. Everyone makes mistakes, and coding is particularly prone to error, especially when there are distractions. I will use these opportunities to demonstrate how to troubleshoot errors by carefully reading the warning messages and running smaller subsets of code to identify where the problems lie. **Learning how to troubleshoot mistakes is one of the most important skills you can learn in this course.**

The entire teaching team (instructor + teaching assistants) is committed to maintaining a healthy and inclusive learning environment. Just as we make mistakes, we recognize that mistakes and errors are an important part of your learning process too. We respect and value students who are not afraid to take risks or try things that might be 'wrong'. Above all, we value students who are not afraid to fail. We will use frequent assignments and testing to limit the impact of mistakes on your final grade. We will provide timely feedback – usually within two weeks – so that you can improve your grade on future assignments.

We will communicate twice per week during lecture and tutorial. Lectures will cover only part of the scheduled time allotment, so that there will be ample time available for questions or assistance.

For Students

It is expected that you will attend weekly lectures and tutorials, though we understand this may not be possible for everyone, all the time, particularly in the post-COVID era. Therefore, everything you need to succeed in this course will also be available online.

You are expected bring a laptop capable running Windows, MacOS, or Linux programs, and you must be able to access Queen's wireless network during lecture and tutorial sessions. When working in class or following lectures at home, **it is very important that you code along in real time**. The only way to effectively learn to program is to practice, and you are expected to practice as much as possible!

You are expected to complete quizzes online, which are administered at the beginning of lecture or tutorial. Assignments are also submitted online and generally due with 48-72 hours of being posted online. You will be prepared for these short deadlines, which are essential to reinforce what you learn in the lectures and online self-tutorials. You must complete quizzes and assignments alone, without communicating with other students. **Any attempts to communicate about quiz or assignment answers will be treated as a breach of academic integrity**. Plan to devote 3-5 hours to learning the lecture material and 10 hours to complete the assignments.

You are expected to check the course website regularly (or use alerts) to keep track of deadlines. **Late assignments are scored as zero** (but see below regarding accommodations).

Any questions or concerns about the course should be raised in lecture or tutorial, or privately during weekly office hours (no appointment needed). Email is generally not an effective tool for course material, and questions that can be addressed in person will not receive an email response. However, email is encouraged for urgent issues (e.g., broken website links or other time-sensitive errors).

For Interactions

You will have regular interactions with the teaching team (TAs, instructor) and with your classmates. In all interactions, you are expected to be respectful and always behave with integrity, both in face-to-face interactions and when engaging online.

This course will also involve group-based activities that will require communication outside of the classroom. You are responsible for maintaining contact and collaborating with all members of your group in a respectful and timely manner. Remember that other members of your group may not have the same resources or privileges and may need some flexibility or accommodation. **Developing skills to collaborate effectively within a diverse group of peers is an important learning outcome of this course**.

Course Learning Outcomes

Students completing this course shall be able to:

1. Produce professional graphs and figures to visualize and explore biological data
2. Translate real-world observations into data appropriate for analysis
3. Apply and interpret statistical and machine learning models to test scientific hypotheses
4. Develop a robust strategy for quality assurance and quality control
5. Contrast the use of fixed vs. random effects in linear vs generalized linear models
6. Write clean and coherent reports that **OPEN** and **REPRODUCIBLE**

Course Materials

The following texts are required for this course:

Essential Biostatistics. A Nonmathematical Approach by Harvey Motulsky (Oxford University Press). This book is available from the Queen’s Campus bookstore and online retailers. It is also available through Queen’s Course Reserves. This is a short book that covers conceptual issues in statistical analysis. Suggested readings are assigned each week and the content is tested on two in-class exams throughout the semester.

R Crash Course for Biologists by Robert I. Colautti. This book is available for free via the course website and forms the basis for some of the early lectures (e.g., data visualizations, data wrangling, reports with R Markdown). This book adopts a self-tutorial style, and it is important that you take the time to physically type out the commands in your computer. The simple act of typing is critical to develop coding skills. Sometimes you will not get the same output, and that’s a good opportunity to learn how to troubleshoot typos and other errors.

R Stats for Biologists by Robert I. Colautti. This book is also available for free via the course website and forms the basis for later stats lectures. As with the R Crash Course, this book adopts a self-tutorial style and you should physically code along on your computer.

In addition to the above required reading, I highly recommend *R for Data Science by Hadley Wickham*. This book is available for free online (https://r4ds.had.co.nz/), and provides further detail on data management techniques in R.

Course Timeline

The following is the planned timeline for the course, however an updated version is available on the course website. Note that there is a **quiz each week**, which is administered at the beginning of class or tutorial. There is also an **assignment due each week**, which is typically posted at the end of lecture and due within 48-72 hours. Two midterms covering chapters in Motulsky are scheduled for Week 04 (Chapters 1-13) and Week 08 (Chapters to 1-25).

	Topic
Week 01	Intro to statistical thinking and data management in R
Week 02	Intro to data visualizations with `qplot()` and `ggplot()`
Week 03	Statistical Inference: the population, the sample and statistical distributions
Week 04	Basic linear models with QA/QC; In-class Test #1
Week 05	Advanced linear models
Week 06	Likelihood, information criteria, and model selection
Week 07	Generalized Linear Models (GLM) and experimental design
Week 08	Linear Mixed Models (LMM); In-class Test #2
Week 09	Generalized Additive Models (GAM)
Week 10	Intro to machine learning and Principal Components Analysis (PCA)
Week 11	Discriminant Analysis
Week 12	Support vector models and decision trees

Suggested Time Commitment

Each week, you should commit 1 to 3 hours reviewing lecture videos, 3 to 6 hours working through the tutorials, 1 to 3 hours reading chapters from the Motulsky book, and 3 to 12 hours completing assignments. Note the wide range of expected time to complete the assignments, which depends on an unpredictable amount of time needed for troubleshooting. It is strongly recommended that you budget 12 hours or more for each assignment, to ensure that you do not miss the submission deadline.

Assessment

50%	Weekly Assignments
20%	Weekly Quizzes
15%	In-class test 1 (Motulsky, Chapters 1-13)
15%	In-class test 2 (Motulsky, Chapters 1-25)

Essential Requirements and Flexibility to Succeed

The workload in this course is demanding because frequent and regular practice is essential to develop competence as a data scientist. However, we also understand that it may not be possible for every student to commit 20+ hours every week to this course. Therefore, we have incorporated a “Universal Design for Learning” that adds flexibility for students who require accommodations, without the need to register with QSAS or request permission from the teaching team.

Instead of traditional accommodations, we account for unforeseen circumstances in our course design. To do this, we will remove the two lowest weekly quiz scores and the two lowest assignment scores before calculating the final grade. This means that students can miss up to two weekly quizzes and two weekly assignments without penalty. Additionally, students may take up to 1 additional day to complete a quiz and 3 additional days to complete an assignment.

Grading Scheme and Grading Method

All components of this course will receive numerical marks, weighted by the percentage shown in the "Assessment" section, above. The final grade you receive for the course will be derived by converting your numerical course average to a letter grade according to Queen's Official Grade Conversion Scale:

Queen's Official Grade Conversion Scale

Grade	Numerical Course Average (Range)
A+	90-100
A	85-89
A-	80-84
B+	77-79
B	73-76
B-	70-72
C+	67-69
C	63-66
C-	60-62
D+	57-59
D	53-56
D-	50-52
F	49 and below

Questions about the course and contacting the teaching team

Coding involves a lot of trial-and-error that can be frustrating for students new to the discipline. First, know that this is completely normal, even to seasoned data scientists. A very common and effective approach to solving errors or other problems is to search Google or Stack Overflow. Often, simply copying and pasting an error into an online search will produce a helpful link. Very often you can just type 'How do I X in R' (or the R package name like ggplot2, dplyr) into Google and look for links to similar questions answered on the Stack Overflow website. In addition, we will generally leave ample time at the end of lectures and tutorials. You are strongly encouraged to ask the question in lecture or tutorial so that all students can benefit from the answer. Any private questions or issues can be discussed during office hours (no appointment necessary). Email is not an effective mode of communication in this course, except to notify the teaching time of potential errors in the online material.

Course Announcements

News and general announcements are posted on the home page of the course website, and new content is released to the course website every week (e.g., lecture topics, assignments). I strongly encourage you to check the website after each lecture.

Course Feedback

Course feedback is solicited throughout the semester. Your feedback is very important to the teaching team and may be used to adjust the course both in the future and the present. In addition to the feedback that we ask for during lectures and tutorials, we welcome any suggestions for improvement that you would be willing to provide (email is a good format for this).

Academic Accommodations and Considerations

As noted above, this course uses a Universal Design for Learning philosophy, adding flexibility in deadlines and grading for students who need accommodations. Any additional accommodations should be handled through Queen's Student Accessibility Services (QSAS): https://www.queensu.ca/studentwellness/accessibility-services

Academic Integrity

Queen's students, faculty, administrators and staff all have responsibilities for upholding the fundamental values of academic integrity; honesty, trust, fairness, respect, responsibility and courage. These values are central to the building, nurturing, and sustaining of an academic community in which all members of the community will thrive. Adherence to the values expressed through academic integrity forms a foundation for the "freedom of inquiry and exchange of ideas" essential to the intellectual life of the University (see the Senate Report on Principles and Priorities).

Students are responsible for familiarizing themselves with the regulations concerning academic integrity and for ensuring that their assignments and their behaviour conform to the principles of academic integrity. Information on academic integrity is available in the Arts and Science Calendar (see Academic Regulation 1), on the Arts and Science website, and from the instructor of this course. Departures from academic integrity include plagiarism, use of unauthorized materials, facilitation, forgery, use of forged materials, contract cheating, unauthorized use of intellectual property, unauthorized collaboration, failure to abide by academic rules, departure from the core values of academic integrity, and falsification, and are antithetical to the development of an academic community at Queen's. Given the seriousness of these matters, actions which contravene the regulation on academic integrity carry sanctions appropriate to the severity of the departure that can range from a warning or the loss of grades on an assignment to the failure of a course to a requirement to withdraw from the university.

Plagiarism is a form of cheating and includes copying code written by students. There is no 'right answer' for the assignments in this class – there are often many potential coding solutions. You will also develop your own coding style, which will make it obvious when code has been copied. To avoid potential for plagiarism, ALWAYS COMPLETE ASSIGNMENTS ON YOUR OWN. As a bonus, you will learn to code better. On the other hand, it is completely fine to ask others to help you troubleshoot an error message or help you figure out why your code isn't working properly. If you become aware of anyone trying to share or solicit code for the assignments, please point them to this passage and inform the teaching team immediately.

Copyright of Course Materials

Course materials created by the course instructor, including the textbooks, online tutorials, slides, presentations, quizzes, assignments, and other similar course materials, are the instructor's intellectual property. It is a departure from academic integrity to distribute, publicly post, sell or otherwise disseminate an instructor's course materials or to provide an instructor's course materials to anyone else for distribution (including note sharing sites), posting, sale or other means of dissemination without the instructor's express consent. A student who engages in such conduct may be subject to penalty for a departure from academic integrity and may also face adverse legal consequences for infringement of intellectual property rights.

Technology Requirement

You must have a laptop computer with internet access to participate in this course. Before attending the first lecture, you should install the following software:

- The R programming environment (free): https://www.r-project.org/
- R Studio Desktop (Open Source Edition, free): https://www.rstudio.com/products/rstudio/#rstudio-desktop
- Open R Studio and run the following lines in the terminal and press enter after each. NOTE: this will install some of the R packages that we use in the course. It may take several minutes to install each one. Be sure to type each line EXACTLY:
    - `install.packages("ggplot2")`
    - `install.packages("tidyverse")`