OCN 682 Introduction to Programming and Statistics in R
Syllabus (subject to change)

Instructor: Anna B. Neuheimer, office MSB 614, annabn@hawaii.edu, 956-2613, abneuheimer.org

Course Motivation and Goals

Many scientists are choosing to learn a programming language to handle all aspects of data analysis (exploring, summarizing, analyzing, visualizing) as well as statistical and theoretical modeling tasks. Programming languages offer many benefits over "point and click" options, allowing work to be explicit & documented, promoting experimentation and exploration, and recording work-flow from start to finish. The latter aspect allows for quality control (as everything is documented, mistakes are easier to find and research is reproducible) and the ability to simply rerun analyzes when data are updated. Moreover, the recording of processes in a programming language allows one to automate sequences of tasks that are often repeated - basic syntax and techniques can be applied to many different problems. In this way, data analysis is accelerated when a programming language is learned.

Learning one programming language helps you learn others. But where do you start? A great place is with the R programming language (http://www.r-project.org/), where benefits include:

- R's free and open-source: There is no need to keep track of expensive licenses, it's available to all (e.g. Mac, Windows, Linux), and the source code is explorable and editable.
- The R community is large and strong: This provides a great community for learning, and also means new analytical tools are constantly being developed and shared as packages (http://cran.r-project.org/web/packages/). Many discipline-specific packages are available (e.g. for climate research, oceanography, fisheries) and learning R means leveraging the expertise of these experts.
- R's great for statistical modeling: R's foundation is one of statistical analysis. Designed by statisticians, the resulting language is intuitive for statistical analysis and the flexible statistical analysis toolkit is large.
- R's great for graphics: Graphics are fully customizable and a number of packages are available that allow for the creation of publishable quality graphics.
- R can interface with other languages (e.g. Python, C/C++) when beneficial e.g. to increase speed.

In this course, you will **learn how to use R for effective data analysis and statistical modeling**. The course will begin with an introduction to programming languages and programming in R. You will learn basic syntax, coding grammar and "etiquette", and a range of vocabulary to aid in data analysis. The path to success in learning any programming language requires i) resources, ii) practice and iii) a community of fellow users accessible for trouble-shooting. This class will provide you with all three. Next, we will learn **how to define a research hypothesis and formulate a corresponding statistical model to test the hypothesis** (including how to translate that model into coding syntax). We will explore strategies in picking a starting model, assessing model fit, selecting the "best specified" model for the data and reporting (and visualizing) statistical results for publication. We will walk through a number of examples geared towards research problems encountered in biology and oceanography, though those in other disciplines will find the methods transfer to their areas as well. You will also have a chance to apply these methods directly to your own research through a term project that will

explore a research hypothesis of your choice and culminate in the writing of manuscript-type methods and results sections.

Student Learning Outcomes: At the end of this course, students will be able to:

- List motivation for learning a programming language
- Access online resources for R and import new function packages into the R workspace
- Import, review, manipulate and summarize data-sets in R
- Explore data-sets relevant to research questions to create testable hypotheses and identify appropriate statistical tests
- Perform appropriate statistical tests using R
- Communicate model choices and results for publication
- Create and edit visualizations with R for publication
- Document analysis and results in a manner that supports transparency and reproducible research

Requirements/Prerequisites: This course is aimed at graduate students. Students are expected to

- have no previous programming experience (but GREAT if you do!)
- have some basic statistical knowledge and a desire to learn more
- be motivated enough to work through the learning curve associated with learning any programming language.

Course Structure: The course will include 3 hours (3 credits) of combined lecture-computer tutorial per week. Students are expected to bring their own laptop (Mac, PC or Unix-based) to each session (all software is free; links will be provided before the class). Please contact the instructor if this limits your ability to take the course in any way (alternate resources are possible).

Class Times: One three-hour session; Mondays 1:30 – 4:20pm, Marine Sciences Building 315

Reading/Texts: There is no required reading for this course. Programming and statistical resources will be distributed in class.

Office hours: By appointment

Grading Scheme: Course grade will be based on i) participation during lecture/lab exercises, ii) weekly assignments that will practice programming and statistics topics from class, iii) a term project where students will explore a research hypothesis of their choice. Course can be taken for grade or CR/NC.

Please note:

- Any student who feels s/he may need an accommodation based on the impact of a disability is invited to contact the course instructors privately. We would be happy to work with you and the KOKUA Program (Office for Students with Disabilities) to ensure reasonable accommodations in the course. KOKUA can be reached at (808) 956-7511 or (808) 956-7612 (voice/text) in room 013 of the Queen Lili'uokalani Center for Student Services.

- UH's Counseling and Student Development Center is available for any personal, academic and career concerns. Their approach is "encouraging, collaborative, goal focused and culturally sensitive." The can be reached at 808-956-7927 and manoa.hawaii.edu/counseling/

- The University of Hawai'i is committed to providing a learning, working and living environment that promotes personal integrity, civility, and mutual respect and is free of all forms of sex discrimination and gender-based violence, including sexual assault, sexual harassment, gender-based harassment, domestic violence, dating violence, and stalking. If you or someone you know is experiencing any of these, the University has staff and resources on your campus to support and assist you. Staff can also direct you to resources that are in the community. Here are some of your options:

    - As members of the University faculty, your instructors are required to immediately report any incident of potential sex discrimination or gender-based violence to the campus Title IX Coordinator. Although the Title IX Coordinator and your instructors cannot guarantee confidentiality, you will still have options about how your case will be handled. Our goal is to make sure you are aware of the range of options available to you and have access to the resources and support you need.

    - If you wish to remain ANONYMOUS, speak with someone CONFIDENTIALLY, or would like to receive information and support in a CONFIDENTIAL setting, contact the confidential resources available here:

      http://www.manoa.hawaii.edu/titleix/resources.html#confidential

    - If you wish to directly REPORT an incident of sex discrimination or gender-based violence including sexual assault, sexual harassment, gender-based harassment, domestic violence, dating violence or stalking as well as receive information and support, contact: Dee Uwono Title IX Coordinator (808) 956-2299 t9uhm@hawaii.edu

Tentative Schedule

09/1	Week 1: Introduction to programming; Scripts – Documenting, commenting & sharing your code; Data types and structures; Creating, importing and manipulating objects – e.g. summarizing, sorting, sub-setting, merging
16/1	NO CLASS
23/1	Week 2: Creating, importing and manipulating objects (cont.); Formatting dates and times; Manipulating multiple components
30/1	Week 3: Manipulating multiple components – for, if and while loops; vectorization; Introduction to visualization; Saving; Installing packages; Reproducible research with R Markdown, etc.
06/2	Week 4: Introduction to statistical modeling; Research questions and model response variable(s); Predictor variable(s); From research hypothesis to statistical model
13/2	Week 5: Choosing a starting model; Assessing model fit; Linear models with continuous and categorical predictors (including ANOVA-type models); Model selection; Reporting your model – communicating statistical results for publication
20/2	NO CLASS
27/2	Week 6: When errors aren't normal – generalized linear models (GLMs)
06/3	Week 7: When relationships aren't linear with shapes unknown - generalized additive models (GAMs)
13/3	Week 8: When relationships aren't linear with shapes known - Testing relationships of known shape
20/3	Week 9: When observations aren't independent – Time- and space-series
27/3	NO CLASS
03/4	Week 10: When observations aren't independent – Mixed modeling
10/4	Week 11: When responses are multiple – Multivariate modeling (unconstrained & constrained ordination)
17/4	Week 12: From research hypothesis to statistical modeling – further examples; Visualization for publication – e.g. multi-panel plotting, ggplot2, maps & charts
24/4	Week 13: Visualization for publication (cont.)
01/5	Week 14: Topics may include: Working with large data sets – e.g. data.table, dplyr, reshape2 packages; Coding for parallel processing; Programming for theoretical modeling – e.g. designing and coding biophysical models; Making your own functions; Creating interactive figures with dygraphs; Publishing results to the web (with Shiny), etc.