BIOSTAT 705: Applied Biostatistical Methods II
Spring 2024

Instructor:

Hussein R. Al-Khalidi, PhD
Professor
Department of Biostatistics & Bioinformatics
Duke University
Email: hussein.al-khalidi@duke.edu
Phone: (919) 668-2923
Office Hours: Thursdays 12:15-1:15 pm
Location: Room 10090

Note: Microsoft Teams (MT) will be the platform to ask questions about the class materials, quizzes/exams and HWs. Make sure your MT notification is enabled.

Class TAs:

Zigui Wang
PhD Student
Email: zigui.wang@duke.edu
Office Hours: Wednesdays 11:30 am - 12:30 pm
Location: Room 10090

Jiajun Liu
PhD Student
Email: jiajun.liu@duke.edu
Office Hours: Mondays 11:30 - 12:30 pm
Location: Room 10090

Course website:

sakai.duke.edu

Course Schedule:

First class: Thursday Jan 11, 2024; Last class: Tuesday Apr 16 2024
Class time: Tuesdays & Thursdays 1:25pm-2:40pm
Class location: Hock 10089 (Hock 10th Floor)
In-class Mid-term Exam: Thursday Mar 7, 2024, 1:25pm-2:40pm EST
Spring Break (no classes): Mar 11-15, 2024
Reading Period: Apr 18 - 28, 2024
In-class Final Exam: Friday, May 3, 2024, 2:00 - 4:00pm EDT
Both mid-term and final exams will be given online with a time limit. The exams will be released and collected via Sakai. All students MUST abide by the Duke Honor Code.

Course rational:

This course focuses on the *practical* application of statistical methods, linear models and procedures to answer biologic/medical research questions utilizing *real* data. As such, the material covered in this course forms a bridge between statistical theory (i.e., BIOSTAT 701 & 704) and the biological/medical problem being addressed (i.e., BIOSTAT 703 & 706). This course will expand on the statistical methods, linear models and analytic technics that the students have learned in the first sequence (BIOSTAT 702) which provides the tools needed to become a practicing professional biostatistician.

Expected background:

Prerequisites: multivariable calculus, linear/matrix algebra
Corequisites: BIOSTAT 701; BIOSTAT 702; BIOSTAT 703

Course objectives:

The primary objective of biostatistical methods sequence (BIOSTAT 702 & 705) is for students, by the sequence's end, to be able to competently navigate the analytic process for a subset of problems commonly encountered by professional biostatisticians. This process involves: (1) identifying the appropriate statistical method required to address a specific inferential problem, (2) implementing the analytic procedure using standard statistical software, and (3) correctly interpreting the results[1].

Course format:

Lecture notes will be posted on Sakai ahead of the class. The course will be a mix of lecture, in-class quizzes (given weekly or every other week in the last 10 minutes of Thursday's class).

Regular attendance and in-class participation is required for this class. Zoom video recordings of classes will be made available via Sakai. These recordings are intended for students who want to review part of the lecture, but it's NOT a substitute for class attendance.

You are asked to bring your laptop to each class so that you can follow along with presentation as well as completing the quizzes and exams via Sakai.

Textbook:

Although, there is no official textbook for this course. However, students are encouraged to download the eBooks below (available for free through Duke Library), which are excellent resources for this class, in addition to the course slides, hand-written notes and handouts.

- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2021). "*An introduction to statistical learning: with applications in R*", 2nd Edition, Springer Series in Statistics.
- Harrell (2015) "*Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*", 2nd Edition, Springer Series in Statistics.

[1] There is often another step in this process: collaborating with a subject area scientist to figure out what question they are trying to answer and then to translate that question into statistical hypothesis and be addressed statistically. This step is systematically addressed in BIOSTAT 703 & BIOSTAT 706

Grading: (Notice there will be no make-up quizzes or exams)

In-class Bi-weekly (or weekly) quizzes: 20%

Topics covered in previous lectures or in the readings is fair game for a quiz question. These quizzes are given during the last 10 minutes of the class on Thursdays. All quizzes are closed notes and they are comprehensive (ie.,questions could be taken from any of the topics covered in previous lectures). A missed quiz will result in a score of 0. Over the course of the semester, the lowest quiz score will be dropped from final grade calculation.

Quiz will be released via Sakai "Assignments" @2:30pm and the answer should be uploaded in a single PDF file into Sakai no later than 2:50pm. The uploaded PDF file name MUST BE as:
quizx_Lastname_Firstname.pdf

In-class quizzes dates: 1/25, 2/8, 2/22, 2/29, 3/21, 4/4 and 4/11

Homework Assignments: 15% & class project: 10% (Zigui Wang & Jiajun Liu)
Homework (HW) will be comprised of written assignments and data analysis projects to be done outside of class time. The homework will be used to evaluate how well you understand analysis procedures: when they are valid, their limitations, and how they are implemented. The write-up of the data analysis projects will follow a standard statistical report format and will be an opportunity for you to get feedback on how well you communicate what you did and what the results mean. ALL HWs assignments will be released to students via Sakai "Assignments" or "Gradescope". Collaboration is encouraged on HWs, but the final write-up should be yours alone. Make sure that you understand every step in the process and are able to defend your approach. Copying elements of someone else's report is a violation of the honor code (see "Duke Community Standard" below) and will get you in trouble.

Notice: HW assignment needs to be posted on due date/time.
(Points will be taken-off for late HW, unless prior arrangement has been made with the TA)

The TA will be grading your HW assignments and provides Key Solution for each assignment. Your HW assignment needs to be written as a scientific report using R Markdown. The TA expects your HW assignment to be uploaded into Sakai as a single PDF file on the due date.

R Markdown Guidelines to be followed during your HW prep:

1. Use R markdown to write your answers (no screenshots)
2. Show all code chunks (not just the output)
3. Write answers outside of code blocks (not as comments)

Class Project:
This will involve data analysis and final statistical report writing using R markdown. The class TAs will randomly divides the class into group of 5 students assigned to each project. Each student in the group will be required to contribute effectively to their group assigned project and be able to present the project findings in oral presentation (if requested by the TA). Statistical methods required for the data analysis for each project are those cover in 705 class, such as multiple linear regression models, variable-selection, ANOVA/ANCOVA and logistic regression models. All class projects (with its presentation) need to be completed no later than last-day of the class (Apr 16^{th}).

In-class Midterm Exam: 25%
This will be in-class exam (closed book and notes) will cover materials through Mar 7, 2023. You are allowed to have one-page (one-side) formulae sheet. The exam will be released to the students at 1:20 pm and submission is expected no later than 3:00 pm via Sakai. The duration of the mid-term exam is 75 minutes, but you're given extra time for down-loading and up-loading the exam via Sakai.

In-class Comprehensive Final Exam: 30%
This will be similar to in-class exam (closed book and notes) with questions taken from any of the topics covered during the semester. You are allowed to have one-page (both sides) formulae sheet.
The exam will be released to the students at 2:00 pm and submission is expected at 5:00 pm via Sakai. The duration of the final exam is 2 hours, but you're given extra time for down-loading and up-loading the exam via Sakai.

Both midterm and final exams require students to bring their laptop, a calculator and common statistical tables such as normal, chi-square, F (or use R functions to calculate the p-value).

Note: Quizzes and exams will not be curved.
Course grades will be assigned as follows:

- A+ : 98% or higher
- A : 93%-97%
- 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 : below 60%

Duke Community Standard:
Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and non-academic endeavors, and to protect and promote a culture of integrity. To uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

Software:
The SAS and R programming language will be our major platform for statistical computing, with more emphasis on SAS. Though not required, we strongly encourage students to typeset their statistical reports using LaTeX.

- **R**: An environment for statistical computing and graphics. Freely available for Linux, Mac, and Windows from http://cran.r-project.org/
- **SAS®**: Statistical analysis software (Version 9.4 or higher). SAS software licenses are available at no cost to the Duke community through a grant from SAS to NCICU. For more information see http://www.oit.duke.edu/software
- **LaTeX**: A document preparation system. Freely available for a number of computer platforms. See "Installation" at http://en.wikibooks.org/wiki/LaTeX for details

Main Topics

Linear models

multiple regression models, regression model in a matrix form; confounding and interaction, dummy variables; variable selection; regression diagnostics; multi-collinearity; lack-of-fit; introduction to ridge regression and LASSO (least absolute shrinkage and selection operator)

Analysis of variance (ANOVA)

One-way and two-way ANOVA; sum of squares and degrees of freedom partition; multiple comparisons (Tukey, Scheffe, Bonferroni); linear contrasts; interpretation of main and first-order interaction effects

Analysis of covariance (ANCOVA)

Relationship between ANOVA, linear regression and ANCOVA; Random and mixed-effects models; variance components in random and fixed effects models; hypothesis testing under mixed-effects models

Simple and multiple logistic regression

Model assumption; coefficients interpretation and relationship to odds ratios; model building and regression diagnostics; hypothesis testing under logit model; Pseudo R^2 measures in logit models;

Introduction to probit and complementary log-log models (if time permits)