Tufts University | Friedman School of Nutrition Science and Policy

Updated: Jan 12, 2024

NUTR 0309: Biostatistics II
Spring 2024

Class and Lab Meetings: **Monday &** Wednesdays 1:30pm to 3:00pm

Instructors: Mei Chung, PhD, MPH
Email: mei_chun.chung@tufts.edu
Office hours: after class and by appointment

Teaching Asst.: Yongyi Pan, MS, MPH (Optional lab instructor)
Email: Yongyi.Pan@tufts.edu
Office hours: see optional lab schedules (TBD)

Muya Li
Email: Muya.Li@tufts.edu
Office hours: by appointment

Semester Hour Units: 3.0

Prerequisites: Biostatistics I (NUTR 0206) or Statistical Methods for Nutrition Science and Policy (NUTR 0207) or equivalent, and graduate standing or instructor consent. **NOTE: Students cannot receive semester hour units for both NUTR 309 and NUTR 307: Regression Analysis for Nutrition Policy. Students who have not taken Principles of Epidemiology (NUTR 0204) or equivalent course are strongly encouraged to take Principles of Epidemiology (NUTR 0204) concurrently with NUTR 0309.**

Course Description: This course is part two of a one-year, two-semester course on statistical methods for nutrition research (any study designs). The focus of this course is on multiple regression models for continuous, binary, time-to-event, and count data. Emphasis is on developing a conceptual understanding of the application of these techniques to solving problems, and on exposure to statistical modeling concepts and theory (non-mathematical) within the context of nutrition research. In the computer lab sessions, students will use concepts learned during lecture to analyze data using statistical software STATA, R and RStudio, or SAS.
Note: While students are not required to use a particular software program for this course, examples are provided primarily using Stata and R. Students who are using SAS may not receive sufficient support from the instructor. In addition, the statistical examples in the textbook and class slides were written in Stata. Equivalent R codes for the examples in the class slides are available.

Course Goals: Students will learn to analyze quantitative data using multiple regression models. Upon successful completion of the course, students should be able to:

- Gain working knowledge of statistical modeling procedures common in nutrition research.
- Analyze continuous outcomes using multiple linear regression models
- Analyze binary outcomes using logistic regression models
- Analyze count outcomes using poison regression models
- Analyze time-to-event outcomes using survival analysis
- Build and run proper multiple regression models in R and RStudio, STATA, or SAS.
- Document and report data analysis using data management plans and either STATA do-files, R scripts, or SAS code files that are liberally commented.
- Practice communicate statistics professionally and ethically (written and oral).

Textbooks: Required: E. Vittinghoff, D.V. Glidden, S.C. Shiboski, and C.E. McCulloch. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models, 2nd edition. Springer, 2012. This book is available electronically, for free, through the Tufts Library.
Note: Statistical examples in Vittinghoff textbook were written in Stata. Other required readings (journal articles) are specified in the syllabus.

Class Materials: All class materials, including lecture notes and assignments, will be posted on Canvas (https://canvas.tufts.edu).

Statistical Software: Students can choose which statistical software they would like to work with: STATA, R, or SAS. R statistical software and RStudio, an integrated development environment for R, will be used for this course. The Tufts **Data Lab** has instructions for downloading R and RStudio. R is available on the computers in the Jaharis student room, in the Boston Data Lab (Sackler 510 and 514), and on laptops at the HHSL library. R is also available on the computers at the Eaton computing lab and the Data Lab on the Medford campus.

STATA: Tufts provides an enterprise site license for Stata SE enabling students, faculty, and staff to install Stata SE on their Tufts-owned or personal computer for educational use. Follow the self-service installation instructions at the site provided below or contact Tufts Technology Services (617-627-3376). You can install a new version or upgrade your existing version to Stata SE 16.1 on your computer by following the steps outlined at Tufts STATA access.

SAS: SAS is a PC-based statistical software suite provided by Tufts for free for students, faculty, and staff. Tufts Technology Services can install SAS on your personal device (in person at either the Boston or Medford campuses, or via remote install) through a request to Tufts SAS access. If you have a Mac, you can run SAS through the TTS Virtual Lab. Contact Tufts Technology Services (617-627-3376) for help with either of these.

Communication Policy:

Students should try to seek out information for themselves before contacting the instructor or the TA. If you cannot find your answer, we have included a 'Clarity Thread' Discussion on Canvas. Please go there first to post your question and check if any student has already asked the same question. We will aim to check the Clarity Thread daily for any inquiries, but if other students know the answer they should go ahead and post it. If you need to ask a more personal question, please get in touch with either the TA or instructors by email.

Classroom Conduct:

Students are expected to attend all classes and lab sessions. Absences should be explained in writing at least 24 hours before class. Missing more than one or two classes/labs per semester will usually result in substantial underperformance. Students are also expected to read all assigned materials before class and come prepared to participate in class discussions and group activities.

Academic Conduct:

Each student is responsible for upholding the highest standards of academic integrity, as specified in the Friedman School's Policies and Procedures Handbook and Tufts University policies (http://students.tufts.edu/student-affairs/student-life-policies/academic-integrity-policy). It is the responsibility of each student to understand and comply with these standards, as violations will be sanctioned by penalties ranging from failure on an assignment and the course to dismissal from the school.

Accommodation of Disabilities:

Tufts University is committed to providing equal access and support to all students through the provision of reasonable accommodations so that each student may access their curricula and achieve their personal and academic potential. If you have a disability that requires reasonable accommodations please contact the Friedman School Assistant Dean of Student Affairs at 617-636-6719 to make arrangements for determination of appropriate accommodations. Please be aware that accommodations cannot be enacted retroactively, making timeliness a critical aspect for their provision.

Tufts Zoom: The Friedman School's on-campus courses may be offered by Tufts Zoom (https://access.tufts.edu/zoom) or WebEx (https://it.tufts.edu/webex) on days when the Boston campus is closed due to pandemic, weather or a temporary cancellation issue. Students should expect to be notified by email in the event that class is cancelled and will be provided with the Zoom link for students to attend any remote class sessions during the normally scheduled class period. The Zoom or WebEx meeting video and audio will be recorded and posted on Canvas when completed. If an on-campus Examination/Presentation was scheduled on a day when the Boston campus is closed due to weather or a temporary cancellation issue and cannot be conducted by zoom, the exam/presentation will be rescheduled for an alternate on-campus class session date.

Diversity Statement:
We believe that the diversity of student experiences and perspectives is essential to the deepening of knowledge in this course. We consider it part of our responsibility as instructors to address the learning needs of all of the students in this course. We will present materials that are respectful of diversity: race, color, ethnicity, gender, age, disability, religious beliefs, political preference, sexual orientation, gender identity, socioeconomic status, citizenship, language, or national origin among other personal characteristics.

Assessment and Grading:
Grading for the course will be based on the following distribution:

Components	**Proportion of final score**
Class participation	10%
Lab participation	10%
Homework	30%
Midterm project	25%
Final exam	25%
Total	**100%**

A passing grade in the course is B- or better. Course grades will be based on the below (subject to revision during the course):

Final score	**Letter grade**
≥ 97 %	A+
94 to < 97 %	A
90 to < 94 %	A-
87 to < 90 %	B+
84 to < 87 %	B
80 to < 84 %	B-
77 to < 80 %	C+
74 to < 77 %	C
70 to < 74 %	C-
<70	F

Class participation (10%)

Students are expected to attend all classes and read all assigned materials before class. Students will be expected to be active class participants by participating in class discussions and activities.

Lab participation (10%)

Students are expected to attend all labs and complete all lab exercises. Missing a lab without prior agreement with the instructor will result in receiving a 0% for that lab exercise.

Homework (30%)

Four problem sets, each worth 7.5% of the final grade, will be assigned throughout the semester. Guidelines for homework will be discussed during the first class. Homework assignments must be submitted through Canvas prior to the lab session. Students who are unable to complete a problem set on time for any reason should contact the instructor by email at least 48 hours prior to the deadline. Without prior notification, assignments received after their deadline will be penalized by a 20% grade reduction for the first 24 hours and 50% for the next 24hours. Late submissions will not be accepted if late more than 48 hours.

Midterm project (exploratory model write-up) (25%)

Scenario: You have been asked, as a statistical expert, to co-author an article. While you may not be a subject matter expert, your co-authors are certainly not experts in statistical modeling. So, you split up the work on the manuscript. Your co-authors will write the article Introduction and Discussion sections, and you will not only carry out the statistical modeling, but you will write the majority of the Statistical Methods section and the entirety of the Results section. For this assignment, you will write the enhanced Statistical Methods and Results sections for a journal article. Thus, you should follow the author guidelines for a major journal. Additional information on the format, grading, and datasets of the project will be provided in session 5.

Final take home exam (25%)

Take home exam is cumulative. The format will be similar to homework assignments with some multiple-choice questions, short answers, and computational questions (computations using a statistical software).

Course Topics and Assignment Schedule at a Glance*:

Session	Date	Topic	Assignments	Readings
1	1/17 (W); 1/22 (M)	Course overview and expectations Review bivariate analyses (correlations, covariance, and simple linear regression) In-class lab: Software introduction (Stata & R) and basic statistic review	Reply to Welcome Message and Pre-class Questions on the course Canvas site before class	Vittinghoff 3.2-3.3
2	1/24 (W); 1/29 (M)	Multiple linear regression models and model diagnostics Computer lab: correlations and linear regression models	Homework 1 assigned	Vittinghoff 4.0-4.3.4; 4.7
3	1/31 (W); 2/5 (M)	Principles of model building – Part 1 (Types of models; research questions and study designs; causal inference; confounding) Computer lab: Dealing with variables with non-linear relationships with the outcome variables (very common for nutrient-disease relationships) in regression	Homework 1 due	Vittinghoff 4.4; 4.6
4	2/7 (W); 2/12 (M)	Principles of model building – Part 2: Regression model building techniques Computer lab: regression model building techniques		Vittinghoff 10.0-10.1.3.2, 10.1.4, 10.2-10.6
5	2/14 (W); 2/22 (Thur)	Logistic regression model building Computer lab: Introduce the mid-term project dataset and provide detailed instructions on the mid-term project	Homework 2 assigned	Vittinghoff 5.1-5.2
6	2/26 (M); 2/28 (W)	Logistic regression model: diagnostics Computer lab: logistic regression building and diagnostics		Vittinghoff 5.3-5.4
7	3/4 (M); 3/6 (W)	Reporting Guidelines for Statistical Analysis Computer lab: Data and syntax for the mid-term project	Homework 2 due	
8	3/11 (M); 3/13 (W)	Students' presentation of the draft methods of the mid-term project for peer and faculty feedback (ungraded) Computer lab: Data and syntax for the mid-term project	Draft methods of the mid-term project due	
Spring break (3/16 to 3/24)				
9	3/25 (M); 3/27 (W)	Introduction to Poisson regression model: building Computer lab: Poisson regression model building	**Mid-term project due**	
10	4/1 (M); 4/3 (W)	Survival analysis – part 1 (Time-to-event outcomes and model) Computer lab: run survival analysis and check proportional hazard assumptions	Homework 3 assigned	Vittinghoff, 6.0-6.2.6.2, 6.2.7-6.2.13

11	4/8 (M); 4/10 (W)	Survival analysis – part 2 (Carrying out and interpret Cox regression model and interactions) Computer lab: Check proportional hazard assumptions and interactions	Homework 3 due Homework 4 assigned	Same as previous week
12	4/15 (No class) 4/17 (W)	Survival Analysis – Part 3 (Cox regression model diagnostics) &		Vittinghoff, chapters 6.4.2
13	4/22 (M); 4/24 (W)	Analyses for Randomized Controlled trials; Interactions/Effect modification ; Sample size and power; Introduction to multinominal and ordinal logistic regression	Homework 4 due	Vittinghoff Chapter 8.3 (Generalized linear models)
14	4/29 (M); 5/1 (W)	Putting everything together (final review for final exam)	**Take-home exam assigned and due in a week**	

*This schedule is subject to modification at the instructor's discretion

Detailed Description of Course Topics, Assignment Schedule, and the Learning Objectives:

Session 1: Course overview and expectations. Review bivariate analyses (correlations, covariance, and simple regression)

Learning Objectives:

During the first part of the class, we will meet and greet and go over the syllabus in detail. Emphasis will be put on the assignments, lab exercises, midterm, and final project, as well as on the expectations for using R/RStudio, STATA, or SAS, and Canvas. During the second part of the class, we will review correlations, covariance, and simple linear regression.

Upon completion of this week, students should be able to:

1. Reference the syllabus for information on expectations and requirements of the course.
2. Use the syllabus to find information on the structure of the course, including important dates and leaning objectives for each class.
3. Get to know their fellow classmates
4. Review correlations and simple regression analysis
    a. Differentiate between different types of correlations
    b. Understand the relationship between correlation and simple regression analysis

Required Reading **before class**:

Vittinghoff 3.2-3.3

Assignments Due **before class**:

1. Reply to Welcome Message and Pre-class Questions on the course Canvas site

Optional:

If you need a quick refresher of Biostats I, read Vittinghoff 3.1

If you need a quick refresher on the use of R and RStudio, the DataLab has tutorials including: R and RStudio basics.

If you need a quick refresher on the use of Stata, watch STATA Video Tutorials below:

L01: Intro to Stata
L02: Results and Do-files
L03: Creating Variables and Labeling

Session 2: Review multiple linear regression and model diagnostics

Learning Objectives:

During class, we will revisit the uses of multiple regression, basic principles for reporting the results of multiple regression models, including regression diagnostics and discuss the remedies to avoid violations of linear regression assumptions. During the computer lab, we will run bivariate analyses, a multiple linear regression, and model diagnostics.

Upon completion of this week, students should be able to:

1. Describe the assumptions and steps required to carry out linear regressions.
2. Run multiple linear regression models and interpret regression coefficients.
3. Understand how to evaluate regression models for outliers, influence and collinearity, and for violations of normality and homescedasticity.
4. Understand how to interpret log-linear transformed models

Required Reading/Assignments:

1. Vittinghoff 4.0-4.3.4; 4.7
2. Homework 1 assigned

Optional readings:

- Stata users: Regression With Stata, chapter 2, Regression Diagnostics: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter2/stata- webbooksregressionwith-statachapter-2-regression-diagnostics/
- R users: Introduction to Regression in R (Part 2, Regression Diagnostics): https://stats.idre.ucla.edu/wp-content/uploads/2019/02/R_reg_part2.html#(1)

Assignments Due:

1. None

Session 3: Principles of model building – Part 1: Types of models; research questions and study designs; causal inference; confounding and effect modification

Learning Objectives:

During the class, we will think about WHY we are creating models, types of models (explanatory, descriptive, and predictive models) and their associated research questions, review study designs (intervention and observational study designs) and causal inference, and finally discuss techniques to deal with confounding and evaluate effect modifications in the regression models. During the computer lab, we will practice variable transformation and other modeling techniques to investigate non-linear relationships with the outcome variable (very common for nutrient-disease relationships) in multiple regression.

Upon completion of this week, students should be able to:

1. Gain a basic understanding of model types and their associated research questions
2. Understand the basic principles of confounding and effect modification
3. Understand how to investigate and interpret non-linear relationships with the outcome variable in multiple regression.

Required Reading/Assignments:

1. Vittinghoff 4.4; 4.6

Assignments Due:

2. Homework 1 is due before the lab session

Session 4: Principles of model building – Part 2: Regression model building techniques

Learning Objectives:

During class, we will describe and contrast different multiple regression model building techniques, including methods for selecting predictors (including using causal reasoning [Directed Acyclic Graph]) and running and refining models to avoid overfitting. During the computer lab, we will build multiple regression models using the variable selection techniques covered in class.

Upon completion of this week, students should be able to:

1. Recognize the differences in model building strategies for explanatory, descriptive, and predictive models.
2. Understand the methods for selecting predictors
3. Gain basic understanding of Directed Acyclic Graph and its use for model building
4. Generate and interpret statistics for comparing different regression models, including R^2, adjusted R^2, extra sum of square F-test, AIC, and BIC.
5. Discuss the drawbacks of predictor selection procedures.
6. Build models with multiple predictors

Required Reading/Assignments:

1. Vittinghoff 10.0-10.1.3.2, 10.1.4, 10.2-10.6

Optional Reading:

- Williams TC, Bach CC, Matthiesen NB, Henriksen TB, Gagliardi L. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatr Res*. 2018;84(4):487-493. doi:10.1038/s41390-018-0071-3

Assignments Due:

1. None

Session 5: Introduction to logistic regression

Learning Objectives:

During class, we will discuss logistic regression to model dichotomous outcomes and how to interpret the results of those models. We will also discuss the relationship of logistic regression models with odds ratios. During lab, we will fit simple and multiple logistic regression models.

Upon completion of this week, students should be able to:

1. Understand the basic principles of creating a multivariable model of dichotomous outcomes.
2. Understand how to interpret the results of logistic regression models in the context of odds ratio.
3. Use logistic regression to analyze binary outcomes using R, STATA, or SAS.

Required Reading/Assignments:

1. Vittinghoff 5.1-5.2

Assignments Due:

1. Homework 2 is due

Special Note:

- This week we will also introduce the mid-term project dataset and provide detailed instructions on the mid-term project.

Session 6: Logistic regression model diagnostics

Learning Objectives:

During class, we will discuss different diagnostic methods to assess logistic regression model fit, and conditional logistic regression for case-control studies (including nested case-control studies). We will also discuss the relationship of logistic regression models with 2x2 tables. During the computer lab, we will build and assess model fit of multiple logistic regression models.

Upon completion of this week, students should be able to:

1. Assess model fit and carry out model diagnostics for logistic regression models.
2. State the relationship between 2x2 tables and logistic regression models.
3. Fit conditional regression models for case-control studies.

Required Reading/Assignments:

1. Vittinghoff 5.3-5.4
2. Draft methods of the mid-term paper

Assignments Due:

1. None

Session 7: Reporting Guidelines for Statistical Analysis

Learning objectives:

During class, we will discuss reporting guidelines for different study designs. Students will present the draft methods of their mid-term paper to get feedback from the instructor and their peers.

Upon completion of this week, students should be able to:

- understand the basic principles of good reporting of scientific articles

Required Reading/Assignments:

1. Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL guidelines. *Int J Nurs Stud*. 2015;52(1):5-9. doi:10.1016/j.ijnurstu.2014.09.006
2. STROBE-nut: Lachat C, Hawwash D, Ocke MC, Berg C, Forsum E, Hornell A, et al. Strengthening the Reporting of Observational Studies in Epidemiology-Nutritional Epidemiology (STROBE-nut): an extension of the STROBE statement. PLoS Med. 2016;13(6):e1002036.

Assignments Due:

1. Homework 2 is due

Session 8: Mid-term project discussion & feedback

Students' presentation of the draft methods of the mid-term project for peer and faculty feedback in class (ungraded)

Assignments Due:

1. Draft methods of the mid-term project due

Session 9: Introduction to Poisson regression model

Learning objectives:

1. Describe situations in which Poisson regression analysis is needed.
2. Interpret the results of Poisson regression models.
3. Assess model fit and carry out model diagnostics for Poisson regression models.
4. Use Poisson regression to analyze count data using R, STATA, or SAS.

Required Reading/Assignments:

Vittinghoff Chapter 8.3 (Generalized linear models)

Assignments Due:

1. Mid-term project due this week

Session 10: Survival Analysis – Part 1 (Time-to-event outcomes and model)

Learning objectives:

- understand the basic principles of modeling time-to-event outcomes
- be able to carry out and interpret survival models

Required Reading/Assignments:

1. Vittinghoff, chapters 6.0-6.2.6.2, 6.2.7-6.2.13

Assignments:

1. Homework 3 assigned

Session 11: Survival Analysis – Part 2 (Carrying out and interpret Cox regression model and interactions)

Learning objectives:

- be able to carry out and interpret survival models
- understand how to evaluate interactions in regression model

Required Reading/Assignments:

1. Vittinghoff, chapters 6.0-6.2.6.2, 6.2.7-6.2.13
2. Homework 4 assigned

Assignments Due:

1. Homework 3 is due

Session 13: Survival Analysis – Part 3 (Cox regression model diagnostics) & Analyses for Randomized Controlled trials; Interactions/Effect modification

Learning objectives:

- understand how to evaluate the validity of the assumptions of survival models.
- understand the remediations when proportional hazard assumption is violated.
- Understand analysis methods for randomized controlled trial design (parallel and cross-over) and interactions/effect modification

Required Reading/Assignments:

1. Vittinghoff, chapters 6.4.2

Session 13: Sample size and power; Introduction to multinominal and ordinal logistic regression

Sample Size and Power Learning Objectives:

1. Recognize that power calculations are an essential part of both study design and analysis.
2. Identify the factors which influence power and sample size.
3. Compute sample sizes for binary and measured outcomes.
4. Discuss strategies to calculate sample sizes when not all information is available, and to address sample sizes that are not feasible.

Multinominal and Ordinal Logistic Regression Learning Objectives:

- Extend the understanding of logistic regression to encompass multi-category outcomes and ordinal outcomes
- Extend the experience working with multivariable models of categorical outcomes

Required Reading/Assignments:

1. Stata users: https://stats.idre.ucla.edu/stata/dae/multinomiallogistic-regression/
2. R users: https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

Assignments Due:

1. Homework 4 is due

Session 14: Putting Everything Together

Learning Objectives:

This is a final review section before final exam.

Assignments:

1. Take home final exam assigned

*** This schedule is subject to modification at the instructor's discretion.**