PH0206; NUTR323
Intermediate Biostatistics:
Regression Methods

Fall 2017

Kenneth K. H. Chui

Contents

1 Version history

This is version 1.00

Future changes on this syllabus will be posted on Trunk (www.trunk.tufts.edu) and the class Wiki (wikis.uit.tufts.edu/confluence/display/MPH206/Home), from where the most updated version will also be made available.

2 About this course

2.1 Class meetings

We will be meeting on Monday from 5:30pm to 8:30pm at Sackler 851.

2.2 Teaching staff

Kenneth Kwan Ho Chui, Course director
Office: Room 117, M&V
E-mail: Kenneth.Chui@tufts.edu
Phone: 617-636-0853

Micaela Karlsen, Teaching Assistant
Office: By appointment
E-mail: Micaela.Karlsen@tufts.edu

2.3 Graduate credits

This is a 1-credit course.

2.4 Prerequisites

Grade B or above in PH0205, or grade B– or above in NUTR307 or NUTR309. Students who wish to use another statistics course as prerequisite please obtain a syllabus of the said course and contact the course director for consent before the end of the add/drop period.

2.5 Description and goals of the course

This course provides a survey of regression techniques for outcomes common in biomedical and public health data including continuous, count, and binary data. Emphasis is on developing a conceptual understanding of the application of regression techniques to solving problems, rather than on numerical details.

The four course goals and their corresponding learning objectives are as follows:

1. Monitor health status in order to identify and solve community health problems.
   (a) Conduct systematic literature review on selected health problems.
   (b) Evaluate quantitative evidence.
   (c) Formulate sound, relevant, and testable hypotheses.
   (d) Expand the hypotheses into research questions.
2. Inform, educate, and empower people about health issues using statistical methods and reasoning.
   (a) Create a data management and analysis plan.

(b) Evaluate the fulfillment of various assumptions required by the chosen statistical procedures.

(c) Suggest remedies or alternatives if assumptions are violated.

(d) Carry out the planned procedure with student's chosen software.

(e) Compose and interpret the results in the format of scientific journal article.

(f) Enumerate pros and cons of the study design and analytic methods.

(g) Communicate results of the analyses in both verbal and written formats that are suitable for target audiences.

3. Research for new insights and innovations in understanding and solving health problems.

   (a) Create the underlying causal pathway for the statistical models.

   (b) Collect and curate resources in order to stay informed in emerging biostatistical methods.

4. Practice and communicate statistics professionally and ethically.

   (a) Maintain a clear and professional documentation, including data management and analysis plans, software syntax files, and the personal wiki page.

   (b) Respect and acknowledge contribution of others by citing sources properly.

   (c) Recognize the susceptibility to false positives in statistical models and be able to detect and safeguard against "fishing expedition."

The learning objectives above serve to guide the composition of the course materials, exercises, and assignments.

3 Textbooks

3.1 Required text

E. Vittinghoff, D.V. Glidden, S.C. Shiboski, and C.E. McCulloch. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models, 2^{nd} edition*. Springer, 2012.

This main text is relatively inexpensive (about $80.00 for the hardcover book, or $0.00 for the e-book) and contains most of the fundamental statistical techniques applied in the field of biostatistics. Read the assigned sections before class. Two copies are put on reserve at the Hirsh Health Sciences Library (HHSL). The call number is WA 950 R343 2012.

Because Tufts has a site license with the book's publisher Springer, students may also read the text online for free.
The text also provides codes in Stata format and data sets in Stata, SAS, and Excel formats. The text's link is:

www.biostat.ucsf.edu/vgsm/

3.2 Supplementary text

J.E. Miller. *The Chicago Guide to Writing about Multivariate Analysis, 2nd edition*. The University of Chicago Press, Chicago, 2013.

This book is a simple and yet elegant work on how to express numerical information through writing. It is strongly recommended for students who would like to learn more about scientific writing. Two copies of the 2nd edition are on reserve at the HHSL (T11 M484 2013).

JAMA and Archives Journals. *AMA Manual of Style: A Guide for Authors and Editors, 10th edition*. Oxford, New York: Oxford University Press, 2007.

Highly recommended for students who would like to learn more about medical or scientific publishing. This book does not just cover problems related to writing or formatting, it also touches on issues such as ethical and legal issues, authorship, conflicts of interest, scientific misconduct, and intellectual property, etc. A copy is on reserve on HHSL (WZ 345 A511 2007).

4 Statistical software

A working copy of statistical software is required for completing the assignment and final project. However, *this course does not specify which software students should use*. We provide practice materials and examples for Stata, SAS, and R. Students may use any of these three. Inform the course director if you plan to use another software that is not listed here.

4.1 Stata (Stata.com)

Stata, currently at version 15, is a script-based software. It also supports point-and-click input. Users can often find useful stand-alone programs, codes, and help from the active online community. Because the development of this software has been more recent, the organization of command codes is relatively more coherent and intuitive compared to other packages.

Tufts has a license agreement with Stata at:

www.stata.com/order/new/edu/gradplans/campus-gradplan

Students can purchase a one-year license for about $125. However, a perpetual license for Stata/IC (maximal number of variables: 2,048) is only $225 and should be considered if the student would like to continue using the software for their upcoming school projects, internship, or work. If you envision yourself working on very large data, then consider the next tier Stata/SE, which can handle up to 32,767 variables.

4.2 SAS (sas.com)

SAS, currently at version 9.4, primarily uses script-based input. The syntax of SAS is not very intuitive and requires some time to adapt. However, once you become familiar with the system, it

will serve you very well.

Tufts has a site license with SAS and students can install a copy on their laptop for free. The license is one-year long and can be renewed. Mac users would need to install virtual Windows through software such as VMware and Parallels Desktop in order to run SAS. Consult the front desk at HHSL Compute Lab on 5^{th} floor for details. A typical installation can be half an hour long or more, so please bring the AC cord. Alternatively, students may also find SAS installed in the computers in the HHSL library and PHPD student lounge.
Notice that SAS may not work equally well on all operation systems, check for the compatibility at:

support.sas.com/supportos/list

4.3 R (r-project.org)

R, currently at version 3.4.2, is a script-based software developed from its predecessor S language. R is an open-source freeware but in no way any inferior to the other three. It can download and run different "packages" contributed by many users to perform functions that are not available in the basic installation. The online community of R is also vibrant and extremely helpful.

For new R users, consider a very nicely made graphical user interface called RStudio. It is freely available at rstudio.com. Remember to install R first, then RStudio.

5 Academic and professional conduct

Students are expected to uphold the highest standards of academic integrity. For details, refer to the following resources:

For students from the Public Health Professional Degree and MS-Biomedical Science Program, refer to the Tufts School of Medicine Student Handbook 2017–2018 (PDF at bit.ly/2vNf3WD).

Professionally, as aspiring biomedical and statistical practitioners, students are also required to read the Ethical Guideline for Statistical Practice of the American Statistical Association and learn to conduct statistical analysis responsibly (Website at bit.ly/1qCBP2).

In June 2016, Wasserstein and Lazar published *The ASA's Statement on p-Values: Context, Process, and Purpose*. The statement contains six principles that "in nontechnical terms [...] could improve the conduct or interpretation of quantitative science." Students are required to read this two-page statement (Website at bit.ly/2bIV3FG).

5.1 Absence and early leave

Notify the course director in advance if you will be absent for the class. Also, inform the lecturer in charge if you decide to leave the class early. Students who miss the class are responsible for downloading and reviewing the lecture materials from the class Wiki page and gathering class

notes from their classmates. After an initial attempt to review the materials, students are welcomed to meet with the teaching staff if they have lingering questions.

5.2 Using computer in class

Students who wish to use computer in class should consider these recommendations:

- If your computer screen is bigger than 12" (about 30 cm), take the seats at far right or left to avoid blocking the view of the students behind you. Another alternative would be sitting at the back of the classroom.
- As an extension of the point above, should you wish to use an AC cord, make sure the wire will not cause potential accident to other people walking by.
- Do not leave your computer unattended.
- Turn off the speakers.
- Minimize the sound of typing.
- Limit your use of computer solely for class-related purposes.

5.3 Cell phone etiquette

Silence all cell phones and beeping devices during the class. If you need to exchange texts with someone, treat it as a phone call: excuse yourself, leave the classroom to finish the communication, and then return to the class.

5.4 Scheduling and canceling an appointment

If you wish to meet with any of the teaching staff, make an appointment through e-mail *one school day prior* to the suggested meeting time. Full attention is not guaranteed for unannounced visits. If for any reason you cannot make it to the meeting, contact the teaching staff at least *one hour prior* to the suggested meeting time. No-shows will be considered when grading student's Participation.

5.5 Have questions?

Our primary Q&A platform is Piazza. A link is available on our Trunk site and Wiki site. Students are encouraged ask all questions related to the course on Piazza. They are *also strongly encouraged to answer others' questions or take part in the discussion within a thread of interest*. To ask a question, simple click "New Post" and then choose "Question" next to post type.

To make it a non-threatening environment, our Piazza site is customized so that students can either fully disclose their name, remain anonymous to the students but disclosed to the staff, or keep a mysterious presence by being fully anonymous. I am fine with either one. However, for questions about your individual analysis project, please at least let the staff know your identity so that we can refer to your analysis plan for details.

Prior to posting the question, perform a general search (literature, online help menu, notes, etc.) on your problem first. Ask for help if you are still struggling on the same matter after *15 minutes* of investigation. Keeping a time limit is particularly important for tackling problems in statistical software because sometimes the answer for a seemingly hard software error could be very simple but unapparent to new users, who may then have to spend hours on this issue.

Do use the tagging function to efficiently categorize the questions for everyone's easy searching. For instance, if the question is about a SAS function applied to weekly assignment 3, tag both "SAS" and "wa3".

If you question is about how the class is conducted (for instance, you may feel unclear about a certain guideline on formatting or how to use Trunk to submit your assignment,) please also post on Piazza right away for clarification. *Do not rely on words of mouth or rumors*.

5.6 How to ask a statistical question

To best help us answer the question, include the following information if appropriate:

1. One to two sentences to briefly summarize what you wish to achieve. You may use your research question as a start.
2. List the variable involved. Particularly what is the dependent variable and what are the independent variables.
3. Mention the type of analysis you plan to do, if applicable.
4. If the question is about software, state what software it is.
5. If the question is about an error returned by a software, paste the error message into the Piazza post or attach a screen shot as an image. (Also, try perform an Internet search with the error message, most of the times you'll find the error has been explained elsewhere.)
6. If a particular chunk of codes is causing the problem, paste that into the Piazza post as a code block.

5.7 Academic integrity statement

Students are expected to abide by the School of Medicines Standards of Academic and Professional Conduct, which include a commitment to academic integrity. Examples of violations of academic integrity are as follows: plagiarism, submitting work used in another course without the permission of the instructor, violating the code of conduct for exam-taking, submitting another persons work as your own and altering or misrepresenting data. As faculty, I am required to notify the Program Director if I have concerns about plagiarism by any student in my course.

Plagiarism is the unacknowledged use or inadequate citing of someone elses work. It is important to note that plagiarism does not need to be intentional. It is your responsibility to learn the rules of citing and documenting sources and to conduct your research carefully. As a class we will be

using Turnitin (plagiarism prevention software) for the two major writing assignments to support student efforts to avoid plagiarism.

If you have any doubt at all as to what constitutes plagiarism I strongly encourage you to speak with Amy Lapidow, the research librarian assigned to students in PHPD. You also might want to familiarize yourself in advance with standard citation formats and become familiar with one or more of the citation management tools available to the Tufts community. Amy can help you with this, or you can find information on the Library website:

hirshlibrary.tufts.edu/research/citation-tools.

Good time management and careful note-taking are critical in avoiding plagiarism. The Library also has workshops and one-one consultations for students:

hirshlibrary.tufts.edu/research/schedule-consultation.

5.8 Writing assistance

Free writing assistance is available to all health sciences students. Writing coaches will help you plan, organize, draft and fine tune your papers as well as help improve your writing skills in the process. Whether you need to clarify your ideas, interpret the assignment, structure your thoughts, connect your paragraphs, or test your success at communicating complex information, the coaches can help. For more information and to schedule an appointment visit:

researchguides.library.tufts.edu/writingconsultants

6 Special accommodation

Students who have documented physical or learning disabilities and need accommodations, must complete an accommodations request form (PDF at bit.ly/2bJXAWo) and submit with required supporting documentation for approval to Robin Glover (Robin.Glover@tufts.edu), Associate Dean for PHPD Programs.

7 Methods of evaluation

7.1 Proportion of the components

Components	Proportion of final grade
Weekly assignments ($7 \times 5\%$)	35%
Scientific writing assignment (SWA)	10%
Individual research project	
Data management & analysis plan	5%
Analysis report	10%
Final presentation	10%
Final report	15%
Personal wiki page	5%
Class participation	10%
Total	**100%**

7.2 Weekly assignments

7.2.1 Description

Seven short assignments will be given throughout the course. The assignments will be largely based on the materials covered in the class.

These assignments serve two purposes: to reinforce the contents through hands on experience and to encourage self-learning. The instructor will evaluate the results and address any weaknesses or points of caution in the following class.

7.2.2 Time line

The seven assignments will be sprinkled across the semester. For actual dates, please see the schedule on page 21. For each assignment, the default open date is noon of the following Tuesday and the due date is 9:00 pm of the following Sunday.

7.2.3 Distribution and submission

Questions can be downloaded from Trunk. Submit your work in either PDF or DOCX format through Trunk. Refer to the format guideline on page 28 for details on formats.

7.2.4 Grading and late policy

Grade will be in number of points out of 5. Unannounced late submissions will receive 0%.

7.3 Scientific writing assignment (SWA)

7.3.1 Description

The Scientific writing assignment resembles a mini-paper. Students will be asked to download a publicly available data set and perform a semi-guided analysis. The purposes of this assignment are to let students better familiarize themselves with large data management and scientific writing.

These assignments serve two purposes: to reinforce the contents through hands on experience and to encourage self-learning. The instructor will evaluate the results and address any weaknesses or points of caution in the following class.

7.3.2 Time line

Details on data set and analysis will be announced in session 2. Students will have more than a month to work on this paper. For actual due date, see the schedule on page 21.

7.3.3 Distribution and submission

Guideline will be posted on Trunk. Submit your work in either PDF or DOCX format through Trunk. Page limit (excluding bibliography) is 12 pages. Refer to the format guideline on page 28 for details on formats. For this assignment students are **not** required to submit their syntax file.

7.3.4 Grading and late policy

Grading will be done using a rubric that will be shared among the students prior to the due date of this assignment. The final rubric scores range from 1 (beginning) through 4 (outstanding.) Grade and point conversion are as follows:

Point range	Grade	Percent obtained
> 3.4	A	100.0
> 2.8 to 3.4	A–	87.5
> 2.2 to 2.8	B+	75.0
> 1.6 to 2.2	B	62.5
≤ 1.6	B–	50.0
	C+	37.5
	C	25.0
	C-	12.5
	F	0.0

Students will receive reviewed copies from the teaching team and learn more about the evaluation from the comments in the margin.

Unannounced late submissions will receive a 12.5 percent point reduction per day, including holiday, until all points are depleted. A fraction of a day will be counted as one full day: i.e. 2.2 days delay will be counted as 3.

7.4 Individual research project—Data management and analysis plan (DMAP)

7.4.1 Description

This is the first component of the individual research project. Students will given a sample plan and they will need to compose a data management and analysis plan to detail the data source, research questions, key variables, and main analysis for their project.

The analysis should be driven by hypotheses. Students are required to perform a literature review and generate three testable hypotheses which are to be answered by any combination of the following statistical techniques:

- Linear regression
- Logistic regression
- Log-linear (Poisson) regression
- Other analytical methods of your choice[1]

The three hypotheses can be sequential (An unadjusted model, an adjusted model, and another model that treat the same outcome differently) or separated (A model to generate an index, another model to predict the outcome with the index as a validation, a third model to apply the index on another totally different data set). However, a general thematic coherence is expected.

7.4.2 On choosing data set

The data sets should be considerably big[2] and should be either obtained from the public domain or permitted to be used by the owner of the data or the associated principal investigator.

If students wish to use data from their work or research, they would need to acquire written permission from their supervisor and ensure that they are included in the IRB document and any pertinent data use agreement as a legitimate data user. If your work or study involves human subjects, the subjects' identification[3] shall not be released in any form throughout this course.

7.4.3 Time line

Format requirement of the sample plan will be distributed in session 2. The due date is about one month into the course, see page 21 for the actual due date.

[1] Past examples include using spatial regression to describe disease clusters, harmonic regression to quantify seasonality of infectious disease, mixed-effects model to identify factors affecting children's growth pattern.

[2] The definition of "considerably big" differs according to the research design and complexity of the models. As a rule of thumb, aim for 15 cases for every explanatory variable. Consult the teaching staff if you have any difficulty in determining if a data set is adequate.

[3] Include but not limited to: 1) name and social security number; 2) street address, e-mail address, telephone and fax numbers; 3) certificate/license numbers; 4) vehicle identifiers and serial numbers; 5) URLs and IP addresses; 6) full face photos and any other comparable images; 7) medical record numbers, health plan beneficiary numbers, and other account numbers; 8) device identifiers and serial numbers; and 9) biometric identifiers, including finger and voice prints.

7.4.4 Distribution and submission

Samples can be downloaded from wiki. Submit your work in either PDF or DOCX format through Trunk. Page limit (excluding bibliography) is 6 pages. Refer to the sample plan for format guidance.

7.4.5 Grading and late policy

Grade will be in number of points out of 5. Satisfactory plans will receive 5. Plans that need more work will receive 2.5, but can be revised and resubmitted within seven days for the remaining 2.5 if the improvement is significant. Unannounced late submissions will receive 0%.

7.5 Individual research project—Analysis report

7.5.1 Description

Prior to the presentation, students are required to compile an analysis report (resembles the result section in a scientific papers and all the associated tables and figures, plus the analytical syntax) for a review. The teaching staff will provide feedback to the modeling approaches and format of the tables and figures. Students will then revise their model and data presentation, and incorporate them into the final report.

7.5.2 Time line

Before the final presentation, see page 21 for the actual due date.

7.5.3 Distribution and submission

Submit your work in either PDF or DOCX format through Trunk. Page limit (excluding bibliography) is flexible depending on your proposed analysis. Analytical syntax can be submitted as is, as a compressed zip file, or transferred onto a PDF or DOCX document. Proper format will for computer syntax will be discussed in class.

7.5.4 Grading and late policy

Grade will be in number of points out of 10. Unannounced late submissions will receive 0%.

7.6 Individual research project—Final presentation

7.6.1 Description

In the last two sessions students will have a chance to share with their peers the project they have been working on.

7.6.2 Time line

The last two sessions will be dedicated to the presentations. The sequence of presentation will be decided randomly and announced at the middle of the semester. The time limit of the presentation ranges from 15–20 minutes depending on enrollment.

7.6.3 Distribution and submission

Submit your presentation file to Trunk by 3:00 pm on the day of your presentation.

7.6.4 Grading and late policy

Grade will be in number of points out of 10. The evaluation will be based on the followings:

1. Clear lay out of data source, background, and hypotheses being tested.
2. Appropriate application of statistical methods.
3. Clarity in interpreting the results and sound logical links between results, discussion, and conclusion.
4. Adequate uses of tables and graphs; attention to visual components such as good color and typographical selections.
5. Good command of oral presentation skills.

7.7 Individual research project—Final report

7.7.1 Description

The final report is the capstone of this course. Students will have a chance to combine all they learned and create a scientific journal article that stems from their chosen data set and planned analysis.

7.7.2 Time line

The final report is due three days after the presentation. Students are allowed to use the time to revise their report based on the feedback they got in the presentation. Notice that although the report is due at the very end, the composition of this report should start after the DMAP is approved.

7.7.3 Distribution and submission

Guideline will be posted on Trunk. Submit your work in either PDF or DOCX format through Trunk. Page limit (excluding bibliography) is 24 pages. Refer to the format guideline on page 28 for details on formats.

7.7.4 Grading and late policy

Grading will be done using a rubric that will be shared among the students prior to the due date of this assignment. The final rubric scores range from 1 (beginning) through 4 (outstanding.) Grade and point conversion are identical to that of the SWA listed on page 14. Students will also receive reviewed copies from the teaching team and learn more about the evaluation from the comments in the margin.

Unannounced late submissions will receive a 12.5 percent point reduction per day, including holiday, until all points are depleted. A fraction of a day will be counted as one full day: i.e. 2.2 days delay will be counted as 3.

7.8 Individual research project—Personal wiki page

7.8.1 Description

7.8.2 Time line

The wiki is usually due 7–10 days after the final presentation. It is the last activity for this course.

7.8.3 Distribution and submission

Just update your wiki by uploading your paper, data (if appropriate), code book, and analytical syntax file onto your personal wiki page by the due date. A sample wiki will be provided for your reference.

7.8.4 Grading and late policy

Grade will be in number of points out of 5. Satisfactory works will receive 5 points. Wiki pages with unsatisfactory format or missing components without justification will receive 2.5 points. Late submission will receive 0 point.

7.9 Class participation

This component rewards students who actively take part in asking and answering questions as well as participate in group discussions in class and online.

7.10 Final grade

The final grade is based on an absolute scale. Grade assignment according to the final accumulated score is tabulated as follows:

Final score	Grade
> 95	A
> 90 to 95	A–
> 85 to 90	B+
> 80 to 85	B
> 75 to 80	B–
> 70 to 75	C+
> 65 to 70	C
> 60 to 65	C–
≤ 60	F

8 Schedule

Date	Topic	Notes
Sep 11	Session 1. Course overview	Start looking for data
Sep 18	Session 2. Linear regression, a review	WA 1 and SWA distributed; Last day to add/drop
Sep 25	Session 3. Violation of the linearity assumption	WA 2 distributed
Oct 2	Session 4. Violation of the normality assumption	WA 3 distributed
Oct 16	Session 5. Data management and scientific writing	Start working on DMAP
Oct 23	Session 6. Model building: Addressing confounding	WA 4 distributed, DMAP due
Oct 25	–	Last day to withdraw
Oct 30	Session 7. Model building: Addressing interaction	WA 5 distributed
Nov 6	Session 8. Mid-course review & Modeling workshop I	Bring a computer, SWA due
Nov 13	Session 9. Violation of the independence assumption & survey with complex sampling	WA 6 distributed
Nov 20	Session 10. Logistic regression	WA 7 distributed
Nov 27	Session 11. Advanced topic	Analysis report due
Dec 4	Session 12. Modeling workshop II	Bring a computer
Dec 11	Session 13. Student presentation (Group I)	
Dec 14	–	Final report due (Group I)
Dec 18	Session 14. Student presentation (Group II)	
Dec 21	–	Final report due (Group II)
Dec 24	–	Personal wiki due
Jan 2	–	Final grades due

WA: Weekly assignment.
SWA: Scientific writing assignment.
DMAP: Data management and analysis plan.

Snow emergency

If the campus closes down due to snow, announcements will be sent via e-mail. You can also check the Tufts Emergency Preparedness website for updates:

emergency.tufts.edu/weather/closing/

Session 1. Course overview

Learning objectives:
We will be spending the first session to meet and greet and go over the syllabus in details. Emphasis will be put on the assignments and individual projects, as well as demonstration of statistical software packages SPSS, SAS, Stata, and R.

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

1. Understand and appreciate the structure of the course.
2. Understand the requirements and expectations of the course.

3. Make an informed choice on their statistical software package.

Preparation for class:
Please start looking for possible candidates for your individual project.

Session 2. Linear regression, a review

Learning objectives:
In the first half of the class we will revisit the basics of linear regression using a one-predictor linear regression as example. In the second part of the course, we will learn more about data visualization.

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

1. Name important stages in an analysis protocol including examination of assumptions, fitting the model, interpreting the regression coefficient, and diagnosis of the model.
2. Locate and interpret important statistics such as Sum of Square, F-statistics, t-statistics, p-values, R and R^2, intercept (constant), slope, and standardized regression coefficients.
3. Match the type of data and graphs properly.
4. Design, critique, and improve graphical devices based on theories introduced by different experts in the field.

Preparation for class:
Read Vittinghoff chapters 4.1–4.3

Session 3. Violation of the linearity assumption

Learning objectives:
In the first half of the class we will look at different approaches to tackle violation of linear assumption. Methods to be discussed include transformations, incorporating non-linear term, spline modeling, and categorization. We will also cover some technical challenges such as interpreting log-transformed data and checking linearity between a predictor and the outcome in a multiple linear regression.

In the second half of the class we will showcase how to incorporate syntax files into our data management and analysis routine. Examples on how to compose simple macro–syntax language that generates syntax chunks that carry out lengthy, regular, and repetitive tasks–using SPSS, SAS, Stata, and R will also be shown.

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

1. Suggest alternative through data transformation and model refinement to tackle the violation of linearity assumption.
2. Interpret regression models that involve log-transformed outcome or predictors.
3. Match the type of data and graphs properly.
4. Design, critique, and improve graphical devices based on theories introduced by different experts in the field.
5. Appreciate the importance of reproducible research and develop a good habit of documenting all data cleaning, analysis, and reporting procedures.

Preparation for class:
Read Vittinghoff chapter 4.7

Session 4. Violation of the normality assumption

Learning objectives:
In the first half of the class we will discuss the consequence when the normality assumption being violated. Detective methods and common remedies of the violation will be covered. We will also discuss the strategies on how to detect and deal with outliers and influential data points, which are very relevant to some violations of normality assumption.

In the second half of the class we will discuss some key issues in scientific writing. Topics to be covered include reporting guidelines of different study design, bibliography software, anatomy of a scientific article, strategies of writing and dealing with writer's blocks, online and printed resources on general and scientific writing. Students who have written or submitted a scientific article are encouraged to share their tips and experience.

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

1. Suggest alternative through data transformation and model refinement to tackle the violation of normality assumption.
2. Generate, graph, and interpret important diagnostic data including residual, leverage, and Cook's distance.
3. Understand the general structure of a scientific article.

Preparation for class:
Read Vittinghoff chapter 4.7

Session 5. Data management and scientific writing

Learning objectives:
In the first half of the class we will discuss the best practices of using syntax files to manage data and analysis work flow. In the second part we will have a review of important components in a scientific paper.

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

1. Understand the usefulness and structure of an analysis plan.
2. Obtain a basic understanding of reproducible research and be able to start practicing it by keeping a good analysis record and set of syntax files.
3. Review the individual components of a scientific paper.
4. Familiarize with and learn to avoid common stylistic and grammatical errors frequently seen in the past submitted coursework.

Session 6. Model building: Addressing confounding

Learning objectives:
In the first half of the class we will discuss the idea of confounding and how to use different approaches to minimize its impact. We will also introduce the basics of directed acyclic graphs (DAG) as a means to better conceptualize the regression model and identify potential variables to be (and not to be) adjusted. We will then introduce automatic model selections (and why never to use any of them) as well as model-based statistics such as R^2 and Adjusted R^2, extra sum of squares F test, Mallow's C_p, AIC, and BIC.

In the second half of the class we will introduce some basics on estimating power, sample size, and effect size. Examples using SAS's `proc power` and a free software *GPower 3* will be shown.

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

1. Suggest preventive measures and remedies to address confounding in various stages of a study.
2. Generate and interpret statistics for comparing different regression models.
3. Appreciate the relationship between sample size, power, and effect size.
4. Calculate sample size, power, and effect size in simple scenarios including χ^2 (chi-square) test, correlation test, one-sample t-test, independent sample t-test, paired sample t-test, and linear regression.

Preparation for class:
Read Vittinghoff chapters 4.4–4.5, 10.1–10.4.

Session 7. Model building: Addressing interaction

Learning objectives:
In this class we will discuss the consequences of and remedies for interaction. We will demonstrate how to model different type of interaction such as continuous-continuous, binary-binary, categorical-continuous, and categorical-categorical. Software-related techniques to facilitate testing interactions will be introduced. We will also discuss to relevant topics: the concept of using "dummy" variables to represent categorical predictors and collinearity.

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

1. Understand the impact of unadjusted interaction in a regression model.
2. Strategize the testing of interaction terms consisting of variables with different levels of measurement.

Preparation for class:
Read Vittinghoff chapter 4.6.

Session 8. Mid-course review & Modeling workshop I

Learning objectives:
In this class we will start with an open forum for questions and answers about the materials cover thus far. In the later part, we will have a modeling workshop. Scenario, data, and code book will be provided on site. The scenarios will be largely about the materials covered up to session 7.

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

1. Foster a clearer understand of basic modeling concepts and techniques.
2. Work collaboratively to design, compile, and discuss regression analyses.

Preparation for class:
Bring a laptop computer with a working copy of your preferred software installed. Make sure in advance that the laptop should be able to access the Tufts WiFi network. Due to the long work time, you're advised to bring the AC adaptor with you as well.

If students need to borrow laptop from the HHSL, do realize that only a minority of loan laptops at the HHSL come with statistical software. Contact the staff member at the HHSL computer lab front desk for details.

Assignments for this week: Not applicable.

Session 9. Violation of the independence assumption & analysis of survey with complex sampling

Learning objectives:
In this class we will discuss the last assumption of linear regression: independence between observations. We will showcase some work-arounds and introduce some advanced technique such as multi-level models and incorporating complex sample weight in the analyses. Students are not expected to become well-versed with these advanced techniques within this course, but they are expected to know the limit of classical linear regression analysis and be able to determine by what technique data generated from a complex study design is better analyzed.

The second part of the class will introduce the fundamental features of survey with complex sampling design. We will use examples such as Stata's svy: command and SAS proc surveyreg to showcase some analyses. Analytical precautions specific to complex survey data will also be discussed.
Upon completion of this week, students will be able to:

1. Appreciate the common lack of independence between observations in most study designs, and suggest solutions.
2. Identify relevant techniques, such as mixed-effects model or regression for complex sampling, that can be used to address the lack of independence.
3. Understand the reasons of applying complex sampling methods, and specific procedures required to analyze data collected through complex sampling.
4. Realize specific features in Stata and SAS that can adjust for complex sampling weights.
5. Identify some publicly available complex survey data including NHANES, BRFSS, etc.

Preparation for class:
Read Vittinghoff chapters 7 & 12.

Session 10. Logistic regression

Learning objectives:
In this class we will formally introduce generalized linear model that can deal with different types of distributions of outcome. We will first look at logistic regression which can incorporate dichotomous outcomes. Two extensions of logistic regression—ordered logistic regression and multinomial logistic regression—will also be introduced.

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

1. Use linear model approach to analyze binary, ordinal, and categorical outcomes.

2. Assess model fit and carry out model diagnostics for their logistic regression models.

3. Interpret the results of logistic regression models in the context of odds ratio.

Preparation for class:
Read Vittinghoff chapter 5.

Session 11. Advanced topic

Learning objectives:
This class serves as a place holder for a lecture on an advanced topic. Past topics include Poisson regression, sample and power calculation, survival analysis, and introduction to R. Topic for this year will be announced the week before.

Session 12. Modeling workshop II

Learning objectives:
In this workshop students will be randomly assigned into groups to carry out a more elaborated modeling task. Scenario, data, and code book will be provided on site. The groups will use the three hours to understand the data, survey the available variables, come up with research questions, fit the models, compose a short online presentation (e.g. PowerPoint), and present to the rest of the class.

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

1. Appreciate the pros and cons of collaborative data analysis.

2. Identify one self's strengths and weaknesses in a collaborative setting.

3. Compose a meaningful story using statistical evidence, and be able to suggest audience-oriented recommendations.

Preparation for class:
Bring a laptop computer with a working copy of your preferred software installed. Make sure in advance that the laptop should be able to access the Tufts WiFi network. Due to the long work time, you're advised to bring the AC adaptor with you as well.

If students need to borrow laptop from the HHSL, do realize that only a minority of loan laptops at the HHSL come with statistical software. Contact the staff member at the HHSL computer lab front desk for details.

Assignments for this week: Not applicable.

Sessions 13 and 14. Individual project presentation

Learning objectives:
In these two sessions we will celebrate our hard work by sharing our results of the individual analysis.

Upon completion of these weeks, students will be able to:

1. Appreciate the many research interests and approaches to answer different research questions.
2. Have opportunities to ask questions and provide feedback in a professional and friendly environment.
3. Realize that presentations are so much more enjoyable with pizzas.

9 Appendix I: Requirements on typesetting for weekly assignments

1. Submit through Trunk in either PDF, DOC, or DOCX format.
2. Hand-written assignments are not accepted.
3. Use letter size dimension (8.5 inch x 11 inch).
4. 1 inch margins.
5. Limit the font size of the main text to 11 through 13; be consistent within one document.
6. Lines in the main text should be of 1.5x to 2.0x spacing.
7. Number all pages.
8. Include your name on the first page.
9. Start a new page for each main question and number each answer clearly. For instance, questions 1.1 and 1.2 can be written on the same page, but once you start to answer question 2, start a new page.
10. Provide mathematical work for numeric answer unless being instructed otherwise.
11. Tables and graphs should be incorporated into the document. Do not submit graphs as a separated file unless being instructed otherwise.

10 Appendix II: Requirements on typesetting for SWA and Final Report

1. Submit through Trunk in either PDF, DOC, or DOCX format.
2. Hand-written assignments are not accepted.

3. Use letter size dimension (8.5 inch x 11 inch).
4. 1 inch margins.
5. Limit the font size of the main text to 11 through 13; be consistent within one document.
6. Lines in the main text should be of 1.5x to 2.0x spacing.
7. Texts in Abstract, tables, and figures can be of single line spacing.
8. Number all pages.
9. Page limit is 12 (excluding references) for SWA and 24 for the final report. If you need to go beyond this limit due to the amount of analysis output, contact the instructor for advice and approval.
10. Include your name on the first page.
11. No title page is required. Use page 1 for the Abstract.
12. Assign the main section as:
    (a) Abstract (Structured, with subtitles Background, Methods, Results, and Conclusion. Word limit is 350 words)
    (b) Background
    (c) Methods
    (d) Results
    (e) Discussion
    (f) Conclusion
    (g) Acknowledgment, if applicable
    (h) References

    Proper use of sub-titles within each main section is allowed.
13. Tables and graphs should be incorporated into the document. You may choose between incorporating the tables and graphs within the text or at the end of the text, as long as they are clearly captioned and indexed.
14. Table's caption should be put *above* the table; figure's caption should be placed *below* the figure.
15. Do ***not*** copy statistical outputs from the software and directly paste onto your assignment. Thoroughly edit the results and arrange them in a well formatted table shell.
16. In-text reference should be in numeric style (Vancouver style.) Documents using other style such as APA style (Author last name, year) will be returned for correction.
17. Do not use footnotes.

END OF THE SYLLABUS