BIOS 6312
Modern Regression Analysis
Spring 2023

Instructor:	Andrew Spieker (andrew.spieker@vumc.org) Office: 2525 WEA, 11125
Instructor office hour:	Tuesday 12:00 PM - 1:00 PM (or by appointment)
Class sessions:	Tuesday 10:30 - 12:00 PM [2525 WEA, Rm. 802] Thursday 10:30 - 12:00 PM [2525 WEA, Rm. 802]
Lab section:	Monday, 1:00 - 2:00 PM [2525 WEA, Rm. 802]
TAs:	Marisa Blackman (marisa.h.blackman@vanderbilt.edu) Siwei Zhang (siwei.zhang.1@vanderbilt.edu)
TA office hour:	Wednesday, 3:00 - 4:00 PM [2525 WEA, Rm. 11139]
Course website:	`https://andrewspieker.com/b6312`

Course description: This course covers modern regression analysis from an applied and methodological perspective. Specific topics to be covered include regression modeling for continuous, binary, nominal, ordinal, count, and time-to-event outcomes. This course will also cover more advanced topics such as basis splines, weighted least squares, longitudinal data analysis, prediction, and nonlinear regression. Emphasis will be placed on approach, strategy, and interpretation of results. Additionally, this course will introduce approaches to handle challenges encountered in the real world (e.g., missing data). Theoretical principles will be demonstrated with real-world examples from biomedical studies.

Pre/co-requisites: Prior mastery of material covered in BIOS 6311 (or equivalent) is presumed. I expect students to have a strong understanding of descriptive statistics, estimation, and inference (both in terms of conceptual understanding and implementation). The lab section, BIOST 6312L, features essential course content and must be taken concurrently. This course relies on the statistical software program Stata, prior experience with which is not required (however, some familiarity with statistical programming is expected).

Required notes: The lecture slides are posted on the course website and are the major source of required material for this course; there is no required textbook.

References: The (non-required) texts below can serve as solid reference material, but where they disagree with lecture material, lecture material takes precedence.

- Vittinghoff, Glidden, Shiboski, and McCulloch. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models, 2nd edition. Springer.
- Kleinbaum, Kupper, Muller, and Nizam. Applied Regression Analysis and Other Multivariable Methods, 3rd ed. Duxbury Press, 1998
- Kleinbaum and Klein. Logistic Regression: A Self-Learning Text (Third Edition) Springer, New York 2010.
- Kleinbaum and Klein. Survival Analysis: A Self-Learning Text (Third Edition) Springer, New York, 2012.

Learning objectives: By the end of the course, students should ordinarily expect to meet the following objectives:

- Understand how statistical methodology is motivated by medical problems.
- Become familiar with statistical concepts including exploratory data analysis, estimation, testing, and prediction.
- Select/implement regression methods to answer scientific questions, including those appropriate for continuous, binary, nominal, ordinal, count, and time-to-event outcomes.
- Properly specify transformations, categorization, and interaction terms in regression models and articulate proper parameter interpretation.
- State the statistical assumptions that are the basis for the conclusions of your analysis, and use diagnostic procedures to characterize evidence of violations.
- Know what it means for a model to be saturated (and know the implications).
- Develop data analytic skills including familiarity with statistical software.
- Develop writing skills needed to communicate the results of a data analysis to a statistically naive reader.

While some of the mathematical underpinnings are provided in the notes, assignments will be heavily skewed toward methodological principles. This course is targeted to people who will be regularly analyzing data in their research.

Expectations and policies:

Expectations you can have of me: You can expect assignments to be graded in a timely fashion. You can expect me to be responsive to your questions and concerns. If you don't receive a response to an email within 24 hours, please do feel free to e-mail again.

Attendance: The course will be offered in-person. If you are planning to be absent, please notify me (in advance, where possible) and make plans to catch up.

Recorded lectures: I plan to record audio and slides. I do occasionally use the whiteboard in in-class discussions; this will not be captured by the recordings. The availability of recordings is subject to technology working as expected. Further, these recordings are in no way an appropriate substitute for synchronous, in-person attendance; rather, the recordings should be considered as a way to augment your experience in the course. I reserve the right to halt recordings if attendance is dropping too low—a stipulation I feel compelled to make clear but hope never to invoke.

Collaboration: You are highly *encouraged* to work together on problem sets, with the caveat that your write-up must be your own words. Exams are an individual effort.

Academic honesty: Students are encouraged to familiarize themselves with academic honesty policies. If you hand in an assignment not written in your own words, I cannot give you credit.

Extra help: Some concepts we cover in this course may be challenging. If you cannot attend office hours, I strongly encourage you to make an arrangement with me to get help.

Late work: While I expect you to make every reasonable effort to turn work in on time, I understand that life does happen. Late work must be approved by me, in advance when possible.

Course evaluations: Please complete end-of-semester course evaluations. I read all comments closely and take them seriously. Changes from year to year are driven in part by prior students' thoughtful feedback. Comments about what works well and specific, constructive suggestions provide a mechanism for me to improve the course in future years.

Voicing concerns: Keeping the above in mind, please do not feel obligated to wait until the end of the semester evaluations for your voice to be heard. If you have concerns about the material, its presentation, or how you're being evaluated, please schedule a time to meet with me and discuss; I want you to know that your voice will be heard.

Assignments and grades: The grading scheme for the course is shown below.

Problems (**A**ll students): 48% (non-biostatistics)/32% (biostatistics)
There will be eighteen problems required for all students. Each problem will be graded holistically on a scale from 0 to 10. You are expected to attempt all required problems with good faith effort; among problems turned in with good faith effort, your lowest-scoring two problems will be dropped and your total score will be based on the other sixteen. Collaboration is highly encouraged, but the assignment you turn in *must* be in your own words (see policy on collaboration).

Problems (**B**iostatistics students): 0% (non-biostatistics)/16% (biostatistics)
Biostatistics students require a strong mathematical and computational foundation in this course to prepare for comprehensive examinations and advanced biostatistics courses. Therefore, there will be nine additional problems required for biostatistics students (at *most* one per assignment), among which the lowest-scoring problem attempted with good faith effort will be dropped. For non-biostatistics students, these problems are optional and good performance will be taken into account at the end of the semester.

Midterm exam: 24% (all students)
There will be an in-class midterm (emphasizing material up to and including weighted least squares), scheduled for Thursday, March 2 from 10:30a to 12:00p. It will be closed-everything, but you will be permitted the use of a scientific calculator.

Final exam: 24% (all students)
There will be an in-class final exam (emphasizing material from binary outcome regression onward), scheduled for Tuesday, April 25 from 10:30a to 12:00p. It will be closed-everything, but you will be permitted the use of a scientific calculator.

Participation: 4% (all students)
Your engagement is essential to the course. There are multiple ways to participate, but some ways include joining class on time, paying attention, asking questions, and engaging in discussions.

Notes on exams: Any optional problems given on exams will be optional for everyone. Further, I reserve the right to borrow exam questions from prior years (all of which are available on the course website).

Note on grading: If you believe there was a grading error, please direct your concern to Andrew. I do not grade "on a curve," but scores are usually slightly scaled upward at the end of the semester.

Accommodation: If you have established accommodations with Disability Services, please communicate them to me at your earliest convenience so we can discuss your needs in this course (`https://www.vanderbilt.edu/eeo/disability_services/contact_us.php`).

Topic outline by date: Below is a tentative and approximate outline of lectures and labs.

Unit	Date	Topic
1: Introduction and review	M – 1/09	Welcome!
	T – 1/10	Review: Principles of biostatistics (estimation and inference)
	R – 1/12	Lab 1: Introduction to Stata [remember your computer :)]
2: Continuous outcomes	M – 1/16	**NO LAB (MARTIN LUTHER KING, JR. DAY)**
	T – 1/17	Simple linear regression: Introduction and ordinary least squares
	R – 1/19	Simple linear regression: Assumptions and geometry
	M – 1/23	Simple linear regression: Prediction and diagnostics
	T – 1/24	Multiple linear regression: Ordinary least squares, geometry
	R – 1/26	Multiple linear regression: Confounding and precision
	M – 1/30	Lab 2: Prediction, and diagnostics
	T – 1/31	Multiple linear regression: Nominal predictors and effect modification
	R – 2/02	Multiple linear regression: Subgroup-specific effects
	M – 2/06	Lab 3: Subgroups and ombnibus tests
	T – 2/07	Multiple linear regression: Transformations
	R – 2/09	Multiple linear regression: Basis splines
	M – 2/13	Lab 4: Transformations and nonlinearity
	T – 2/14	Multiple linear regression: More on prediction and diagnostics
	R – 2/16	Multiple linear regression: Weighted least squares
	M – 2/20	Lab 5: Summarizing analyses of continuous outcomes
3: Discrete outcomes	T – 2/21	Binary outcome regression: Introduction
	R – 2/23	Binary outcome regression: Further topics with examples
	M – 2/27	Exam review session
	T – 2/28	Regression of categorical outcomes
	R – 3/02	Midterm exam
	M – 3/06	Lab 6: Binary outcome regression
	T – 3/07	Regression of ordinal outcomes
	R – 3/09	Poisson regression
	M – 3/13	**NO LAB (SPRING BREAK)**
	T – 3/14	**NO CLASS (SPRING BREAK)**
	R – 3/16	**NO CLASS (SPRING BREAK)**
	M – 3/20	Lab 7: Discrete outcomes
4: Time-to-event outcomes	T – 3/21	Time-to-event outcomes
	R – 3/23	Proportional hazards regression
	M – 3/27	Lab 8: Survival analysis
5: Correlated data	T – 3/28	Correlated data: Introduction
	R – 3/30	Methods for longitudinal data
	M – 4/03	Lab 9: Longitudinal regression models
6: Machine learning	T – 4/04	Prediction and cross-validation
	R – 4/06	Penalized regression
	M – 4/10	Lab 10: Penalized regression
7: Advanced topics	T – 4/11	Nonlinear regression
	R – 4/13	Strategies for missing data
	M – 4/17	Lab 11: Missing data
	T – 4/18	The nonparametric bootstrap
	R – 4/20	Bayesian methods
	M – 4/24	Exam review session
	T – 4/25	Final exam

Problem sets and their due dates: This is a schedule of due dates for collections of problems. Should any changes be necessary, I will alert you of them.

Set	Date	Type A problems	Type B problems
1	R - 1/19	A1	B1
2	R - 1/26	A2, A3	B2
3	R - 2/02	A4, A5	B3
4	R - 2/09	A6, A7	B4
5	R - 2/16	A8, A9	B5
6	R - 3/23	A10, A11, A12	–
7	R - 3/30	A13, A14	B6
8	R - 4/06	A15, A16	B7
9	R - 4/13	A17	B8
10	T - 4/18	A18	B9

Please note that problems are due by email prior to the start of class (i.e., by 10:30a) on their respective due dates. Please submit your solutions in a word-processed document by email. You code should be attached as an appendix to your solutions. If your answer is not correct, having access to your code as an appendix will help us identify the error and, in turn, award partial credit more appropriately and provide more helpful feedback to you. Importantly, unedited software code or output should not be included as part of your response under any circumstance; anything of that nature that you wish to turn in should be in a supplementary appendix only.