Advanced Topics in Multivariate Analysis
Generalized Linear Models
PLAD 8320, University of Virginia
Fall 2016

Instructor: Constanza F. Schibber

Time and Location: Tuesday and Thursday 11:00 AM - 12:15 PM, Nau Hall 142

Contact: constanza@virginia.edu

Office Hours: Thursday 2:00 PM - 4:00 PM, and by appointment (266 Gibson Hall)

Overview

This is a graduate-level course on the theory and application of generalized linear models (GLMs). In a generalized linear model (GLM), the response variable has a distribution in an exponential family and the mean response is related to covariates through a link function and a linear predictor. GLMs allow a unified theory for many of the models used in statistical practice, including normal theory regression and ANOVA models, loglinear models, logit and probit models for binary data, and models for poisson or gamma responses, *etc.*. Emphasis will be placed on statistical modeling, building from standard normal linear models, extending to GLMs, and going beyond GLMs.

Objectives

Upon successful completion of this course, students should be able to:

1. Translate political phenomena into mathematical notation.
2. Understand the value and limitations of generalized linear modeling.
3. Given a data generating process, select an appropriate statistical model and method.
4. Test substantive hypotheses using generalized linear models.
5. Interpret a variety of types of model estimates.
6. Describe the assumptions of generalized linear models and address violations of them where possible.
7. Use `R` to import, manipulate and describe data, implement models, conduct diagnostics and sensitivity analysis, and produce publication-quality figures.

Evaluation

Participation & Attendance: I expect students to attend all lectures and to arrive to class on time. Students who use laptops in class must do so exclusively for the purpose of note taking. Forms of participation may include asking questions, answering questions from the instructor or from other classmates, participating in in-class group activities and class discussion, among others. Using the course email list to ask and answer questions is strongly encouraged and it will contribute to your participation evaluation

Assignments: Most weeks you will have readings and problem sets. Assignments will consist of a combination of analytical problems, programming, and data analysis. I encourage you to work on the problem sets in groups, but you must write up your work on your own. Assignments should be written in a professional fashion and also include the `R` code you used to address specific problems. I recommend that you prepare your assignments using Latex plus the `R` library `knitr` (instructions will be provided separately), because it will be less time consuming.

Unless otherwise noted, assignments 1 through 3 should be completed by the time of class, 11 AM. Starting from HW#4, the *R component* of the assignments should be completed by 11:59 PM on the designated dates on the schedule and submitted via the Uva Collab site (`R` code and a solution in PDF). If the `R` code submitted does not run, the exercise in question will not be graded. Any *written component* can be handed in to the instructor personally or left in the instructor's mailbox in the designated due date before the office is closed, or scanned and submitted via the to Uva Collab site by the end of the day.

You are encouraged to work together with your fellow students to help each other complete the problem sets. *But do not copy answers from another student, or allow your answers to be copied, or look for and copy solutions to the assignments on the internet.* There is a clear difference between collaboration on assignments and copying. Examples of cheating include copying answers from another student, directly copying computer code from another student, or allowing answers or code to be copied, or directly sharing partial or complete code with a student to be submitted as their own assignment solution. If you are unsure about the difference, please come speak to me before there are any potential problems. Copying is cheating and will be referred to the Honor Committee.

Midterm Exam: There will be one in-class written midterm exam on November 10^{th}, 2016. The exam will be closed book.

Research Paper: The main assignment is to write a research paper that replicates an existing piece of scholarship. The following are some *minimum* requirements the article you select must meet:

- Be from your field of interest and, preferably, answer a question you are interested in and/or uses data you will eventually use in your own research.
- Be published in a leading journal in Political Science.
- Use methods at least as advanced as those introduced in this class.

- Is not being replicated by another student enrolled in the class.

You are encouraged to ask for my advice on the article you are considering replicating before submitting the "Replication Proposal" due on October 13^{th}, 2016. You can stop by my office with a hard copy of the article and some descriptive statistics of the variables included in the statistical model you seek to replicate. Although the "Replication Proposal" is not graded, I will decide whether to approve, approve with revisions, or reject a proposal. If a proposal is approved with revisions, you will have to submit a short written response to my comments on October 25^{th}, 2016. If a proposal is rejected, you will have to find another published article to replicate and submit a new proposal on October 25^{th}, 2016.

Your research paper should *not* simply reproduce the table of results and figures included in the article you select. You will conduct a thorough reanalysis of the paper by embracing the authors' theory and hypotheses, but writing your own `R` code to analyze the data and fit the model(s) presented in the article. The following is a brief list of what the replication entails (detailed instructions will be provided separately):

- Assess the validity and reliability of the measures.
- Assess the authors' modeling choices and present an alternative modeling strategy *if necessary*.
- Use mathematical notation to write down the model. Describe the model and equation.
- Conduct sensitivity analyses and provide an explanation.
- Put the authors' hypotheses to a test and provide an explanation.
- Create high-quality graphics presenting the results and provide an explanation.
- Provide reproducible `R` code.

Be careful! Anyone can find "reasonable" ways of changing someone else's regressions so that coefficient estimates change. That is not the objective of this assignment. The goal of this research paper is for you to grasp the complete research process by focusing on characteristics of the data, the most appropriate quantitative method for establishing a clear connection between theory and empirics, hypothesis testing, and the substantive interpretation and visualization of the results. The end product should look like the statistical and empirical section of a paper published in a lead journal, along with a general assessment of article you replicated and your `R` code.
The research paper is due on December 15^{th}, 2016, by 11 AM, via the Uva Collab site. You are welcomed to (and should) stop by my office or send me a draft of your research paper for feedback. I recommend that you finish all your `R` portion of the paper (including graphics) before December 6^{th} and you use the rest of the time to polish the writing/presentation/explanation part of the paper (which is important and not easy to accomplish).

Required readings:

Clemens, Michael A. "The Meaning of Failed Replication: A Review and Proposal." *Journal of Economic Surveys*, *Forthcoming*. Copy at ftp.iza.org/dp9000.pdf.

King, Gary. 2006. "Publication, Publication." *PS: Political Science and Politics*, 35(1): 119-125. Copy at http://gking.harvard.edu/papers

Wainer, Howard "How to display data badly." *American Statistician* 38(2): 137-147. Copy at http://www.rci.rutgers.edu/~roos/Courses/grstat502/wainer.pdf

Presentation of the Research Paper: Each student will conduct a 15-minute in-class presentation.

Grading

Your grade will be structured as follows:

- Participation and attendance: 5%
- Assignment: 5% each; total of 45% (Out of 10 homeworks, the lowest grade will be dropped when computing each student's score)
- Midterm: 20%
- Research Paper (replication): 20%
- Presentation of the Research Paper: 10%

Late assignments will not be accepted and no incompletes will be given for assignments, exams, or the course. Exceptions will be granted only under truly extraordinary circumstances.

The procedure to have any grade revised is as follows. Please write up a short description of your argument as to why your grade should be changed and hand it in, along with your initial assignment, within one week of receiving your grade. The instructor will respond in writing. The instructor's decisions regarding grades are final.

Required Text

Faraway, Julian J. 2005 *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Boca Raton, FL: Chapman & Hall/CRC (Referred to as Faraway in the Reading List)

A variety of papers and chapters from other books will be assigned as well.

Installing R

All students will need to download and install the latest R software. R is a free statistical programming language that we will use to fit models, simulation, computing probabilities, creating graphics, *etc.*. It may be obtained at the CRAN website. Go to http://lib.stat.cmu.edu/R/CRAN and click your choice of platform (Linux, MacOS X or Windows) for the precompiled binary distribution. Note the FAQs link to the left for additional information.

UVa Collab

All assignments, details on the replication paper, exam, grades, and readings other than the ones in the textbooks will be posted on the UVa Collab site for the course, accessible at https://collab.itc.virginia.edu/portal. If you are officially enrolled in the course, the course website should already be accessible to you.

Schedule

Tueday	Thursday
Aug 23rd	25th **1** Introduction to the Course
30th **2** 1. PMFs and PDFs	Sep 1st **3** 2. Parameters and Moments
6th **4** 3. Logarithms, Summations, Long-Products, Derivatives	8th **5** 4. Chain rule, Optimization
13th **6** **HW #1 Due** 5. The Likelihood Model of Inference	15th **7** 6. The Likelihood Model of Inference

Tueday	Thursday
20th **8** **HW #2 Due** 7. Models for Dichotomous Outcomes: Logit, Probit & Cloglog	22nd **9** 8. Models for Dichotomous Outcomes: Log-Odds, Predicted Probabilities & Marginal Effects
27th **10** **HW # 3 Due** 9. Models for Dichotomous Outcomes: Interaction Terms	29th **11** 10. Models for Dichotomous Outcomes: `R` Lab
Oct 4th **Reading Day**	6th **12** **HW # 4 Due** 11. Models for Count Outcomes: Poisson, Negative binomial, & Zero-inflated count models 12. Models for Contingency Tables
11th **13** 13. Models for Unordered Categorical Dependent Variables: Multinomial Logit & Multinomial Probit, Conditional Logit	13th **14** **Replication Proposal Due** **HW # 5 Due** 14. Models for Unordered Categorical Dependent Variables: `R` Lab
18th **15** 15. Models for Ordered Categorical Dependent Variables: Ordered Logit & Ordered Probit	20th **16** **HW#6 Due** 16. Models for Ordered Categorical Dependent Variables: `R` Lab
25th **17** 17. The GLM Theory and the Exponential Family Form	27th **18** **HW#7 Due** 18. The GLM Theory and the Exponential Family Form

Tueday	Thursday
Nov 1st **19** 19. Advanced R Programming: Hypothesis Testing and Presentation of Results	3rd **20** **HW#8 Due** 20. Advanced R Programming: Hypothesis Testing and Presentation of Results
8th **21** *Election Day* 21. Model Checking, Sensitivity Analysis, Crossvalidation	10th **22** **Midterm Exam**
15th **23** 22. Missing Data	17th **24** 23. Missing Data: R Lab
22nd **25** 24. Other GLMs, Quasi-Likelihood Estimation	24th **Thanksgiving Day**
29th **26** **HW#9 Due** 25. Hierarchical Modeling: Varying Intercepts (Random Effects)	Dec 1st **27** 26. Hierarchical Modeling: More on Varying Intercepts
6th **28** **HW #10 Due** 27. Hierarchical Modeling: Varying Intercepts & Varying Slopes	8th
13th	15th **Research Paper Due**

1 Reading List Organized by Class Number

1. Class 1: Read the Syllabus!
2. The Real Class 1: Read the Syllabus!
3. Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press. Chapter 2 (Concepts and methods from basic probability and statistics)
4. Jonathan Kropko. 2016. *Mathematics for Social Scientists*, Sage. Read chapters 4 (sections 4.5-4.9) and chapter 5 (sections 5.1-5.3).
5. Jonathan Kropko. 2016. *Mathematics for Social Scientists*, Sage. Read chapters 6 and 7 (sections 7.3).
6. Faraway, Chapter 1
7. Gary King. 1998. *Unifying Political Methodology. The Likelihood Theory of Statistical Inference*, Michigan University Press. Read chapters 1 and 2.
8. Altman, Douglas G and Patrick Royston. 2006. "The cost of dichotomising continuous variables." *BMJ* 332:1080 http://www.bmj.com/content/332/7549/1080.1
9. Readings:
   - Faraway, Chapter 2, Binomial Data
   - Gelman and Hill, Chapter 5, Logistic Regression
10. Interaction Terms in GLMs
    - Ai, Chunrong and Edward C. Norton. 2002. " Interaction terms in logit and probit models." *Economic Letters* 80: 123-129.
    - Berry, William, Jacqueline H. R. DeMeritt, and Justin Esarey. 2009."Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?." *American Journal of Political Science*, 54(1):248-266.
    - Tsai, Tsung-han and Jeff Gill. 2013. "Interactions in Generalized Linear Models: Theoretical Issues and an Application to Personal Vote-Earning Attributes." *Social Sciences*, 2(1):91-113. Copy at http://www.mdpi.com/2076-0760/2/2/91
11. Revise all of the readings on Models for Dichotomous Outcomes
12. Reading Day
13. Faraway, Chapter 3, Count Regression & Faraway, Chapter 4, Contingency Tables
14. Faraway, Chapter 5, Multinomial Data
15. Dow, Jay K. and James W. Enders. 2004. "Multinomial probit and multinomial logit: a comparison of choice models for voting research." *Electoral Studies*, 23(1):107-122

16. Faraway, Chapter 5, Multinomial Data (last section of the chapter on ordered logit)

17. Gill, Jeff. 2001. *Generalized Linear Models: A Unified Approach.* Sage (Electronic Version available through the Library)

18. Gill, Jeff. 2001. *Generalized Linear Models: A Unified Approach.* Sage

19. Faraway, Chapter 6, Generalized Linear Models

20. Berry, William D., Matt Golder, and Daniel Milton. 2012. "Improving Tests of Theories Positing Interaction" *Journal of Politics*, 74(3):653-671

21. Hanmer, Michael J. and Kerem Ozan Kalkan. 2013. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models." *American Journal of Political Science* 57(1):263-277

22. No reading

23. Exam

24. Little, Roderick J. A and Donald B. Rubin. 2012. *Statistical Analysis with Missing Data.* John Wiley & Sons (selected chapters)

25. van Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. "`mice`: Multivariate Imputation by Chained Equations in `R`." *Journal of Statistical Software* 45(3):1-67. Copy at https://www.jstatsoft.org/article/view/v045i04/v45i04.pdf

26. Faraway, Chapter 7, Other GLMs

27. Faraway, Chapter 8, Random Effects

28. Optional: Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press. Chapter 11 (Multilevel structures) and 12 (Multilevel linear models: the basics).

29. Last day!