Generalized Linear Models
PLS 900, Michigan State University
FALL 2018

Instructor: Constanza F. Schibber

Time and Location: Tuesday and Thursday 2:40 AM - 4 PM, S104 South Kedzie.

Contact: schibber@msu.edu

Office Hours: Tuesday and Thursday 11:40 AM - 12:00 PM, 344 South Kedzie.

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 or solving in-class exercises in `R`; if a student uses the computer for other activities he or she will receive a participation grade of zero. 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, being active on Piazza or Perusall, among others. Each week your participation will be graded (−, ✓, or +) basis.

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. Moreover, in order to receive credit for a question on a problem set, you must show your own work. Correct answers with no work will receive no credit. Assignments should be written in a professional fashion and also include the commented `R` code you used. Assignments should be written using `knitr` with `R Markdown` and, if you need to write a mathematical equation, you must use LaTeXwithin the same document. Unless otherwise noted, assignments should be completed by 2:40 PM on the day of the deadline. You should upload the assignment to Piazza (both the Rmd and a compiled PDF) and bring a hard-copy to class. If the Rmd file you submitted cannot be compiled due to an error, it will cost you 50% of the grade regardless of the reason for the error. No late homework will be accepted, except in the case of a documented emergency. It is your responsibility to request specific feedback on your assignment through Piazza, in class, or during office hours.

Class Discussion: On several weeks we will have special readings that will be posted on Perusall. You will read the readings in detail on Perusall, add comments or questions, and also answer to other students' questions through the online platform. Your work prior to the class discussion will be graded on a 0 to 100 scale. Your participation in the class discussion will be graded (−, ✓, or +) basis.

Midterm Exam: There will be two closed-book written midterm exams. The first one will take place on October 16 during class. The second one will be on Thursday, December 13, 2018, 10:00am - 12:00pm in S104 South Kedzie – notice that this date is different to the one provided by MSU, if you have a conflict you must tell me as soon as possible.

Research Paper: The main assignment is to write a research paper that replicates and reanalysis an existing piece of scholarship. You are not allowed to work in groups for the research paper.

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 (e.g. 3rd year paper).
- 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.

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 the model. Describe the model and equation.
- Conduct sensitivity analyses and cross-validation; provide an interpretation of the results.
- Put the authors' hypotheses to a test.
- Create high-quality graphics to present the results and provide an interpretation.
- 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 Wednesday, December 13, 2018 at 10 AM. Bring a hard-copy to the midterm exam and upload an electronic copy along with replication material to D2L.

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

A **Replication Proposal** is due on October 23, 2018. You are strongly encouraged to start working on the proposal as early as possible. You should also seek my advice to make sure the paper is a good fit for the class. To do so, send me a copy of the article, descriptive statistics and figures of the key variables included in the statistical model you seek to replicate, and the R code and dataset you used to produce them. I will not provide any feedback or advice requested after October 8, 2018.

Instructions on what the proposal entails will be shared with the class and proposals must follow these guidelines. Proposal will be approved, approved with minor revisions, approved with major revisions, or rejected. If a proposal is approved with revisions, you will submit a written response to my comments along with the revised proposal by November 6, 2018. If a proposal is rejected, you will have to find another published article to replicate and submit a new proposal by November 6, 2018. A proposal can only receive a final grade after the approval of the revisions or the new proposal. They will be graded as follows,

- Approved: 93-100
- Approved with minor revisions: 87-92
- Approved with major revisions: 77-86
- Rejected: 0-76

If a student does not hand in an appropriate response to the revisions or does not comply with the deadline, the grade of the proposal might not reflect this scheme (as the grade could be lower than anticipated).

An update on the status of your proposal is due November 15, 2018. Failure to comply with this requirement or the deadline will negatively affect the grade of your research paper.

Presentation of the Research Paper: Each student will conduct a 15-minute in-class presentation on the last week of classes. Students will be graded on their own individual presentation (90%) and on their interaction/attention/questions during the rest of the presentations (10%). The presentation will follow the standards of a professional conference and the slides will be submitted by e-mail before the presentation.

The student or students with the best research paper and the best presentation will have the opportunity to present in the Quantitative Methods Workshop during Spring 2019.

Grading

I do not give out grades, you *earn* your grade. Your grade will be structured as follows:

- Participation and attendance: 2%
- Assignment: 3% each; total of 21%
- Class Discussion: 8%
- Midterm: 17% each; total of 34%

- Paper Proposal: 5%
- Research Paper (replication): 20%
- Presentation of the Research Paper: 10%

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. Decisions regarding grades are final.

Late assignments will not be accepted and no incompletes will be given for assignments, exams, or the course. Late paper proposals will be graded as a 'rejected & resubmit' proposal (with a grade 0-76). Late research papers will receive a penalty of 10 points per hour late (e.g. if a paper that receives a grade of 100 out of a 100 was submitted three hours late, the final grade would be 70). Late presentation will not be accepted and will receive a grade of zero. Exceptions will be granted only under truly extraordinary circumstances.

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.

Schedule

Tueday	Thursday
Aug 28th **1**	30th **2** 1. Introduction to the Course
Sep 4th **3** 2. Review	6th **4** 3. Review 4. More Review
11th **5** 5. The Likelihood Model of Inference **HW #1 Due**	13th **6** 6. The Likelihood Model of Inference
18th **7** 7. Models for Dichotomous Outcomes: Logit, Probit & Cloglog	20th **8** 8. Models for Dichotomous Outcomes: Model Derivation
25th **9** 9. Models for Dichotomous Outcomes: Log-Odds & Predicted Probabilities **HW # 2 Due**	27th **10** 10. Models for Dichotomous Outcomes: Interaction Terms & First Differences
Oct 2nd **11** 11. Models for Dichotomous Outcomes: `R` Calculation of Predicted Probabilities Using the Observed Value Approach	4th **12** 12. Models for Dichotomous Outcomes: Rare Events, Separation
9th **13** 13. Models for Censored and Truncated Data; Sample Selection Models **HW # 3 Due**	11th **14** 14. Group Presentations of “Mini Replications” (part of HW # 3 Due)

Tueday	Thursday
16th **15** **Midterm Exam 1**	18th **16** 15. The GLM Theory and the Exponential Family Form
23rd **17** 16. The GLM Theory and the Exponential Family Form **Replication Proposal Due**	25th **18** 17. Models for Count Outcomes: Poisson
30th **19** 18. Models for Count Outcomes: Poisson, Negative-Binomial, and Zero-inflated Models **HW # 4 Due**	Nov 1st **20** 19. Models for Count Outcomes: `R` Lab
6th **21** 20. Models for Unordered Categorical Dependent Variables: Multinomial Logit & Multinomial Probit, Conditional Logit **HW # 5 Due**	8th **22** 21. Models for Unordered Categorical Dependent Variables: `R` Lab
13th **23** 22. Models for Ordered Categorical Dependent Variables: Ordered Logit & Ordered Probit **HW # 6 Due**	15th **24** 23. Models for Ordered Categorical Dependent Variables: `R` Lab **Project Update Due**
20th **25** 24. Missing Data **HW # 7 Due**	22nd **26** **Thanksgiving Day**

TUEDAY	THURSDAY
27th **27** 25. Advanced R Programming: Bootstrapping, Model Checking, Sensitivity Analysis, Cross-validation	29th **28** 26. Workshop Day
Dec 4th **29** 26. Class Presentations, Part I	6th **30** 27. Class Presentations, Part II

1 Reading List Organized by Topic Number

1. Class 1: Read the Syllabus!

2. 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), Chapter 3 (linear regression)

3. Gelman & Hill, Chapter 4 (linear regression)

4. Thomas Brambor, William Roberts Clark, & Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." Political Analysis 14: 63-82.

5. Gary King. 1998. *Unifying Political Methodology. The Likelihood Theory of Statistical Inference*, Michigan University Press. Read chapters 1 and 2.

6. Faraway, Chapter 1

7. Readings:

   - Faraway, Chapter 2, Binomial Data

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. 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. 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

12. TBD

13. TBD

14. Group Presentations

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

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

17. Faraway, Chapter 3, Count Regression & Faraway, Chapter 4, Contingency Tables

18. Gelman & Hill, Chapter 6

19. Martin, Lanny W. and Georg Vanberg. 2005. "Coalition Policymaking and Legislative Review." *American Political Science Review* 99(1): 93-106.

20. Faraway, Chapter 5, Multinomial Data

21. 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

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

23. TBD

24. 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

25. Bring your projects!

26. Papers selected by students.

27. Papers selected by students.