GSERM 2020: "Regression for Publishing"

June 2020

Professor Christopher Zorn
Department of Political Science
Pennsylvania State University
E-mail: zorn@psu.edu
June 15-19, 2020

Course Description

This is a third course on regression analysis in the Global School for Empirical Research Methods (GSERM) at the University of St. Gallen. The course builds directly upon the foundations laid in *Regression I* and *II*, with a focus on successfully applying linear and generalized linear regression models. After a brief review of the linear regression model, the course addresses a series of practical issues in the application of such models: presentation and discussion of results (including tabular, graphical, and textual modes of presentation); fitting, presentation, and interpretation of two- and three-way multiplicative interaction terms; model specification for dealing with nonlinearities in covariate effects; and post-estimation diagnostics, including specification and sensitivity testing. The course then moves to a discussion of generalized linear models, including logistic, probit, and Poisson regression, as well as textual, tabular, and graphical methods for presentation and discussion of such models.

All course materials (including this syllabus, slides, notes, data, computer code, homework exercises, etc.) are available at the UNISG CANVAS system, as well on a dedicated Github repo, which can be found at https://github.com/PrisonRodeo/GSERM-2020-git. Throughout this syllabus, hot links are in GSERM green.

Prerequisites

Knowledge of Topic

Mathematics: Comfortable familiarity with univariate differential and integral calculus, basic probability theory, and linear algebra is required. Familiarity with discrete and continuous univariate probability distributions will be helpful.

Statistics: Students should have completed Ph.D.-level courses in introductory statistics and linear regression models, up to the level of GSERM's *Regression II*.

Hardware

Students will complete course work on their own laptop computers. Microsoft Windows, Apple OS-X, and Linux variants are all supported; please contact the instructor to ascertain the viability of other operating systems for course work.

Software

Basic proficiency with at least one statistical software package/language is highly recommended. Preferred software packages include the R statistical computing language and Stata. For purposes of consistency, course content will be presented using R; computer code for all course materials (analyses, graphics, course slides, examples, exercises) will be made available to students. Students choosing to use R are encouraged to arrive at class with current versions of both R (https://www.r-project.org) and RStudio (https://www.rstudio.com) on their laptops.

Work Load

At least 24 units 45 minutes each on 5 consecutive days. Main course times: 11:00-15:00 UTC (13:00-17:00 CEST; 07:00-12:00 EDT), 15-19 June 2020.

Course Content and Structure

This course builds directly upon the foundations laid in Regression II, with a focus on successfully applying linear and generalized linear regression models. After a brief review of the linear regression model, the course addresses a series of practical issues in the application of such models: presentation and discussion of results (including tabular, graphical, and textual modes of presentation); fitting, presentation, and interpretation of two- and three-way multiplicative interaction terms; model specification for dealing with nonlinearities in covariate effects; and post-estimation diagnostics, including specification and sensitivity testing. The course then moves to a discussion of generalized linear models, including logistic, probit, and Poisson regression, as well as textual, tabular, and graphical methods for presentation and discussion of such models. The course concludes with a "participants' choice" session, where we will discuss specific issues and concerns raised by students' own research projects and agendas.

Structure / schedule:

- **Day One**: Review of linear regression + presentation and interpretation of linear regression models.

  Readings:

- Berk, Richard. 2010. "What You Can and Can't Properly Do with Regression." *Journal of Quantitative Criminology* 26(4):481-487.
- Berk, Richard, Lawrence Brown, Andreas Buja, Edward George, Emil Pitkin, Kai Zhang, and Linda Zhao. 2014. "Misspecified Mean Function Regression: Making Good Use of Regression Models That Are Wrong." *Sociological Methods & Research* 43:422-451.
- Kastellec, Jonathan P., and Eduardo L Leoni. 2007. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics* 5:755-771.

- **Day Two**: Fitting and interpreting models with multiplicative interactions + nonlinearity: specification, presentation, and interpretation.

  Readings:

  - Brambor, Thomas, William R. Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14:63-82.
  - Esarey, Justin, and Jane Lawrence Sumner. 2018. "Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." *Comparative Political Studies* 51:1144-1176.
  - Hainmueller, Jens, Jonathan Mummolo, and Xiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27:163-192.

- **Day Three**: Anticipating criticisms: Model diagnostics and "robustness" + introduction to logit, probit, and other generalized linear models (GLMs).

  Readings:

  - Freedman, D. A. 2006. "On the So-Called 'Huber Sandwich Estimator' and 'Robust' Standard Errors." *The American Statistician* 60:299-302.
  - King, Gary, and Margaret E. Roberts. 2014. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What To Do About It." *Political Analysis* 22:1-21.

- **Day Four**: GLMs: Presentation, interpretation, and discussion + practical considerations and extensions.

  Readings:

- Mize, Trenton D. 2019. "Best Practices for Estimating, Interpreting, and Presenting Nonlinear Interaction Effects." *Sociological Science* 6:81-117.

- **Day Five**: "Participants' choice" session: Models for panel / time-series cross-sectional data + examination period.

  Readings:

  - Clark, Tom S. and Drew A. Linzer. 2015. "Should I Use Fixed Or Random Effects?" *Political Science Research and Methods* 3(2):399-408.
  - Keele, Luke, and Nathan J. Kelly. 2006. "Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables." *Political Analysis* 14(2):186-205.
  - Zorn, Christopher. 2001. "Estimating Between- and Within-Cluster Covariate Effects, with an Application to Models of International Disputes." *International Interactions* 27(4):433-45.

Literature

Mandatory:

See above. Additional readings may also be assigned as necessary or requested; a list of those readings will be sent to course participants. All additional readings will be available on the course CANVAS site, github repository, and/or through online library services (e.g., JSTOR).

Supplementary / Voluntary:

- Weisberg, Sanford. 2014. *Computing Primer for Applied Linear Regression Using R*, 4th Ed. Available at http://users.stat.umn.edu/ sandy/alr4ed/links/alrprimer.pdf

Mandatory readings before course start:

None.

Examination Part

Grading:

- Two written homework assignments (20% each)
- A final examination (50%)
- Oral / class participation (10%)

Supplementary aids:

The exam will be a "practical examination" (see below for content). Students will be allowed access to (and encouraged to reference) all course materials, notes, help files, and other documentation in completing their exam. Additional useful materials include:

- Fox, John, and Sanford Weisberg. 2011. *An R and S-Plus Companion to Applied Regression*, Second Edition. Thousand Oaks, CA: Sage Publications.
- Nagler, Jonathan. 1996. "Coding Style and Good Computing Practices." *The Political Methodologist* 6(2):2-8.

Examination content:

The examination will involve the application of the techniques taught in the class to one or more "live" data example(s). These will typically take the form of either (a) a replication and extension of an existing published work, or (b) an original analysis of observational data using linear and/or generalized linear regression. Students will be required to specify, estimate, and interpret various forms of regression models, to present tabular and graphical interpretations of those model results, to conduct and present diagnostics and robustness checks, and to give detailed explanations and justifications for their responses.

Literature:

- Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*, Third Edition. Thousand Oaks, CA: Sage Publications.

  Faraway, Julian J. 2002. *Practical Regression and ANOVA Using R*. London: CRC Press. (Available in the "Readings" folder.)