POL036 (first part) Research Methods in Political Science I: Advanced Topics in Applied Regression

Class day/time	Fall semester 2017, 29 September – 1 October, room 41
Number of credits	2[1]
Class type	lectures and seminars
Type of completion	credit
Instructor	Constantin Manuel Bosancianu[2]
Contact person	Miroslav Nemčok (miroslav.nemcok@gmail.com)
Assistance	Michal Pink

1 Workshop plan

This intensive 3-day workshop will expand the standard OLS toolbox that participants are accustomed to using. The expansion ventures into territory where OLS estimation produces biased or inefficient results due to the violation of one or more of the standard regression assumptions. Given the relative simplicity, speed, elegance and robustness of OLS over more complex estimation procedures, this course tries to therefore "save" the *least squares* framework (as much as possible) by introducing adaptations of it.

We start with a brief coverage of the most important OLS assumptions, emphasizing in particular those that refer to the regression residuals: their Gaussian distribution, constant variance (*homoskedasticity*), and linear relationship to the predictors. We discuss, in turn, how OLS estimates of effect and uncertainty are impacted by violations of these assumptions, and what tools we have available in R to diagnose these problems. I make the point that these assumptions are frequently not met in the course of many analyses, leading to biased estimates and, therefore, shaky conclusions. The rest of the first session is dedicated to the issue of *heteroskedasticity*[3]: what its implications are for estimates, how it can be detected in the course of a standard analysis, and how commonly it appears as a problem. To address this issue, I present two potential solutions. The first is heteroskedasticity-consistent standard errors, which address the problem in cases in which we have no clear idea of the shape of the non-constant variance. Heteroskedasticity-consistent SEs continue to be a very popular approach in a variety of disciplines, which is why they are covered in depth here. The second, more general, solution is the use of Weighted Least Squares (WLS). Both subtopics are discussed from a theoretical perspective, as well as in a practical setting, in the laboratory.

---

[1] If a student attends only this class, s/he will be awarded 2 ECTS. In case of attendance of both classes, the resulting number of ECTS for a student is 4.

[2] Research Fellow in the *Institutions and Political Inequality* research unit, Wissenschaftszentrum Berlin für Sozialforschung (WZB), Berlin, Germany. Email: manuel.bosancianu@outlook.com

[3] Essentially, this is the violation of the assumption of constant variance (*homoskedasticity*).

In the second day of the workshop I take up the issue of effect heterogeneity across different subpopulations in the sample. In practice, this will involve an in-depth discussion of interactions in linear models. We will cover two-way and three-way inter-actions, both for continuous and dichotomous predictors, as well as how to present marginal effects in a graphical way. As we will see, interactions are frequently a source of confusion in published work, and continue to be misinterpreted. In the final part of the day, I bring up the issue of *fixed effects*, as a solution to omitted variable bias in regression models. Such a strategy is frequently invoked in the search for accurate causal estimates of effects. As in the previous days, the theoretical coverage is followed by applied lab work, using R and empirical data.

I conclude the workshop with a presentation of *semi-parametric models*, which can be used to model simultaneously both linear and non-linear relationships between variables. Such models allow us to test models where not all relationships between predictors and outcome are linear, in the search for a more faithful regression-based description of the data. We start from very simple bivariate specifications, based on smoothing splines, and end our discussion with full semi-parametric models.

2 Grading

The students are required to attend all classes to successfully complete the course.

3 Readings

Day 1 (room 41, 15:15-18:30)

- [mandatory] Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach*, 5th edition. Mason, OH: Cengage Learning. Chapter 8: "Heteroskedasticity" (pp. 268–302).
- [optional] Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. New York: Sage. Chapter 12: "Diagnosing non-normality, nonconstant error variance, and nonlinearity" (pp. 267–306).

Day 2 (room 41, 09:45-13:00)

- [mandatory] Kam, C. D., & Franzese Jr., R. J. (2007). *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor, MI: The University of Michigan Press. Chapter 3 ("Theory to practice") and Chapter 4 ("The meaning, use, and abuse of some common general-practice rules"), pp. 13–102.
- [optional] Brüderl, J., & Ludwig, V. (2015). "Fixed-effects panel regression". In Best, H., & Wolf, C. *The SAGE Handbook of Regression Analysis and Causal Inference*. London: SAGE Reference. Chapter 15, pp. 327–357. (please read only until roughly page 338).

Day 3 (room 41, 09:45-13:00)

- [mandatory] Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. New York: Sage. Chapter 17: "Nonlinear regression" (pp. 451–475).
- [optional] Keele, L. (2008). *Semiparametric Regression for the Social Sciences*. New York: Wiley. Chapters 1–3 and 5 (pp. 1–84 and 109–136).

4 Software requirements

Participants should bring their own laptops to the sessions. Please make sure that R version 3.4.1 or newer is installed on your computer. Additionally, you will want to have installed a GUI for R - my recommendation is RStudio, which is freely available from https://www.rstudio.com/products/rstudio/download/

5 Ancillary materials

I will circulate all slides, data sets and R script files on which the lectures and labs are based one day before the sessions are scheduled to commence.

(*Second part continues below*)

POL036 (second part) Research Methods in Political Science II: Introduction to Computer Assisted Text Analysis

Course Information

Class day/time	Fall semester 2017, 18-20 December, room 41
Number of credits	2[4]
Class type	lectures and seminars
Type of completion	credit
Instructor	Juraj Medzihorsky[5]
Contact person	Miroslav Nemčok (miroslav.nemcok@gmail.com)
Assistance	Michal Pink

Course outline

This course provides a concise hands-on introduction to computer assisted text analysis for social scientists. The participants will learn how to automate document collection and processing, scale text using dictionaries and dimensionality reduction techniques, and use machine learning techniques to automate text annotation. The course relies on the R language.

The course will meet in three days. Each day will consist of two 90 minute sessions, and contain both a theoretical exposition of the material as well as computer exercises.

Course requirements

Basic familiarity with content analysis and with the R language. While it is not necessary, the participants are strongly encouraged to install the R language on their computers.

Grading

At the end of each day of the course the participants will be given a take-home exercise. Each of the three exercises will contribute 20% to the final grade. The remaining 40% of the final grade consists of in-class activity, which includes participating in in-class exercises.

[4] If a student attends only this class, s/he will be awarded 2 ECTS. In case of attendance of both classes, the resulting number of ECTS for a student is 4.

[5] Postdoctoral Fellow at the V-Dem Institute, The University of Gothenburg, Sweden. Email: juraj.medzihorsky@gmail.com

Components	
Participation	40%
Exercise 1	20%
Exercise 2	20%
Exercise 3	20%

Scale		
credit	minimum	maximum
pass	60%	100%
fail	0%	59%

Reading list

Day 1: Introduction to computer-assisted text analysis (room 41, December 18, 9:45-16:45)

- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267-297.

Recommended:

- Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons.
- Roberts, C. W. (2000). A conceptual framework for quantitative text analysis. *Quality & Quantity*, *34*(3), 259-274.

Day 2: Text scaling (room 41, December 19, 9:45-16:45)

- Lowe, W. (2016). Scaling things we can count. Available online.

Recommended:

- Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705-722.
- Lowe, W., Benoit, K., Mikhaylov, S., & Laver, M. (2011). Scaling policy preferences from coded political texts. *Legislative studies quarterly*, *36*(1), 123-155.
- Lowe, W. (2008). Understanding wordscores. *Political Analysis*, *16*(4), 356-371.

Day 3: Text categorization and annotation (room 41, December 20, 9:45-16:45)

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

- Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278-295.

Recommended:

- Roberts, M. E., et al. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064-1082.
- Carlson, D., & Montgomery, J. M. (2017). A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts. *American Political Science Review*, 1-9.