Course title: **Quantitative methods of data analysis: statistical theory and good practices**

Instructor: Michal Kotnarowski, Ph.D.

General description.

The course will focus on the application of basic and intermediate regression techniques in social sciences. Various regression analyses are among the most commonly used analytical techniques in sociology, political science, and (to a lesser extent) psychology. The critical skill of a scholar in social sciences, regardless of substantive interests, should be knowledge of these techniques. The scholar, on the one hand, should be able to understand the work of other researchers applying these techniques, but on the other, should have an ability to use regression techniques correctly in her/his research.

The course assumes that participants have basic knowledge of descriptive and inferential statistics. During the course, participants' statistical skills will be expanded to an intermediate level. After completing the course, participants will be able to conduct regression analyses by their own on the level allowing for publishing in academic journals. Moreover, the participants will gain statistical foundations required to master more advanced analytical techniques, such as multi-level modeling, structural equation modeling, panel regression, time series analysis, event history analysis, or machine learning.

Goals of the course.

After completing the course, participants will be able to understand scientific texts in which regression techniques have been applied. Participants will get to know how to interpret the published results of regression analyses correctly. They will also gain the ability to critically evaluate the use of regression analyses in the work of other researchers, and recognize when it is not appropriate to use regression techniques in research. Finally, the course participants will be able to conduct their regression analyses correctly on their own, at least at an intermediate level.

Prerequisite Knowledge.

Participants of the course should have a thorough understanding of basic statistical concepts such as mean, median, variance, standard deviation, and standard error. They should be familiar with the fundamentals of inferential statistics such a Central Limit Theorem, confidence intervals, t-tests, Anova, and rules of hypothesis testing. The ability to use statistical programs is not required at the beginning of the course.

Detailed description of the course.

The course will begin with the introduction of linear regression models, also known as ordinary least squares (OLS) models. In these models, the dependent (outcome) variable is a continuous variable defined on the interval scale. Participants will estimate these models, interpret their parameters, and assess the models' fit to the data. The regression models will then be extended by taking into account qualitative exploratory variables and introducing interactions between variables. Next meetings will concern the assumptions of the linear

regression model, such as linearity, multi-collinearity, heteroskedasticity, and autocorrelation. Participants will explore the meanings of these assumptions, the consequences of not meeting them, the methods of diagnosing whether the given assumption is met, and possible remedies for violations of assumptions.

In the second semester, the course will cover regression models in which dependent variables are qualitative. These are situations in which the dependent variable is either:

1. a binary variable, when respondents select one out of two options (e.g., whether they voted in the last election)
2. a nominal variable, when respondents select one out of three or more options (e.g., which party they voted for in the last election)
3. an ordinal variable (e.g., when a respondent chooses an answer on the Likert scale) or
4. a variable counting the number of occurrences of a phenomenon (e.g., how many times a respondent participated in protest actions).

General Linear Models (GLMs), which are an extension of OLS models, will be used to analyze this type of data. In particular, the course will include such kinds of GLM models as binary logistic regression, probit regression, multinomial logit, ordinal logit, Poisson regression, negative binomial model.

The course will focus on the practical application of the introduced statistical techniques. The emphasis will be placed on the presentation of regression analyses results both in tabular form as well as in the form of simple and complex statistical graphics. During the course, theoretical aspects of statistical models, which are crucial to their correct application, will be discussed.

Participants will practice regression techniques on datasets provided by the instructor or on their own data related to their Ph.D. projects. In the practical part of the course, regression techniques will be applied using the R program. No prior knowledge of the R program is required. Course participants will learn how to use the R program during the course.

Detailed schedule of the course.

Date	**Topic**	**Readings**
Oct 8	1. Introductory session –Statistical models in social sciences.	ARAGLM – Ch.1
Oct. 15	2. Regression analysis – what is it?	ARAGLM – Ch.2, Field Ch. 4
Oct. 22	3. Examining data	ARAGLM – Ch.3; CAR Ch. 3
Nov 5	4. Transforming data	ARAGLM – Ch.4, Field Ch. 5
Nov 12	5. OLS regression - estimation, parameters and goodness of fit measures	ARAGLM – Ch.5, CAR - Ch. 4.1-4.4
Nov 19	6. OLS regression - statistical inference	ARAGLM – Ch.6; CAR – Ch. 5.1-5.2
Nov 26	7. Regression with dummy variables	ARAGLM – Ch.7; CAR – Ch. 4.5-4.9
Dec 3	8. Regression with interaction terms	Brambor, Clark, Golder

		2006;
Dec 10	9. Analysis of variance	ARAGLM Ch. 8; Field – Ch. 10
Dec 17	10. Outliers and influential cases	RD Ch. 4; CAR – Ch. 8
Jan 7	11. Regression assumptions –non-linearity	RD Ch. 7 & 8
Jan 14	12. Regression assumptions – collinearity	RD Ch. 3
Jan 21	13. Regression assumptions – heteroscedasticity	HiR Ch. 1 & 2
Jan 28	14. Regression assumptions - autocorrelation	ARAGLM Ch. 12
Feb 4	15. First semester overview, presentations of students' first semester reports	
Feb 25	16. Introduction to General Linear Models – linear model vs. general linear model	Long Ch. 3
Mar 3	17. Introduction to General Linear Models – linear predictor, link function	ARAGLM Ch. 14.1
Mar 10	18. Maximum Likelihood Estimation	
Mar 17	19. Binary Logistic Regression vs. Probit models	ARAGLM Ch. Ch. 15.1
Mar 24	20. Binary Logistic Regression – interpretation of parameters, predicted probabilities.	Long Ch. 4
Mar 31	21. Binary Logistic Regression - goodness of fit measures.	Long Ch. 4
Apr 7	22. Binary Logistic Regression – interaction terms.	Fox (2003)
Apr 21	22. Binary Logistic Regression – interpretation using tools of statistical graphics I.	CAR – Ch. 6
Apr 28	23. Binary Logistic Regression – interpretation using tools of statistical graphics II.	
May 5	24. Multinomial logit – interpretation of the model parameters, interaction terms	Long Ch. 4
May 12	25. Multinomial logit – predicted probabilities, goodness of fit measures	ARAM Ch. 14.2, Fox & Hong (2009)
May 19	26. Conditional logit	Liao Ch.7
May 26	27. Ordinal logit	Long Ch. 5
Jun 2	28. Poisson regression and negative binomial model	Long Ch. 8
Jun 9	29. Overview of the second semester	
Jun 16	30. Presentations of students' first semester reports.	

References:

ARAGML: Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. Third Edition. Los Angeles: SAGE.

CAR: Fox, John, and Harvey Sanford Weisberg. 2011. *An R Companion to Applied Regression*. Second Edition. Sage Publications, Inc.

Field: Field, Andy P., Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. London ; Thousand Oaks, Calif: Sage.

HiR: Kaufman, Robert L. 2013. *Heteroskedasticity in Regression: Detection and Correction*. Thousand Oaks, California: SAGE Publications.

Long: Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. 1st ed. Sage Publications, Inc.

RD: Fox, John. 1991. *Regression Diagnostics*. Newbury Park, Calif: Sage Publications.

***

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14(1): 63–82.

Liao, Tim Futing. 1994. *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*. Thousand Oaks, Calif: Sage.

Fox, John. 2003. "Effect Displays in R for Generalised Linear Models." *Journal of Statistical Software* 8(15). http://www.jstatsoft.org/v08/i15/ (July 13, 2017).

Fox, John, and Jangman Hong. 2009. "Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the Effects Package." *Journal of Statistical Software* 32(1): 1–24.

Students' duties during the course:

Course participants are required to read the assigned readings before each meeting (approx. 30 pages per week). Additionally, participants will have to prepare homework assignment for every second meeting. The assignment will consist of exercises similar to the ones discussed previously during the class. At the end of each semester, the participants will prepare a research paper. Techniques discussed during the semester will be applied in the research papers.