Quantitative methods 2 - Linear and generalized linear models.

Instructor: Michal Kotnarowski, PhD (IFIS PAN and GSSR).

Academic year: 2023/2024

Planned course timetable:

Tuesdays 12.00-14.00 (until March 12 2024)
Mondays 14.30-16.30 (from March 18 2024)
Office hours: Thursdays 14.00-16.00 (by prior appointment).
Teaching period: 27 February 2024 -18 June 2024.

Format of the course: onsite class, room 268.

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 scholars in social sciences, regardless of substantive interests, should be an understanding 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 the ability to use regression techniques correctly in their research.

The course assumes that participants have basic knowledge of descriptive and inferential statistics. During the course, participants will expand their statistical skills to an intermediate level. After completing the course, participants will be able to conduct regression analyses on their own at the level allowing for publication in academic journals. Moreover, the participants will gain the statistical foundations required to master more advanced analytical techniques, such as multi-level modelling, structural equation modelling, 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 academic 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 as the Central Limit Theorem, confidence intervals, and rules of hypothesis testing. The class will be carried out in R. Therefore, participants should have a basic knowledge of R as a statistical programming language and of RStudio.

A 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. The next meetings will concern the assumptions of the linear regression model, such as linearity, multi-collinearity and heteroskedasticity. 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.

In the second part of the semester, the course will cover regression models in which dependent variables are categorical. 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 analyse this type of data. In particular, the course will include 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 PhD projects. In the practical part of the course, regression techniques will be applied using the R program.

Students' duties during the course:

Course participants are required to read the assigned readings before each meeting, and actively participate in the classes. Additionally, participants will have to prepare homework assignments and a research paper at the end of the course.

Detailed schedule of the course.

Date	Topic	Readings
Feb 27	1. Introductory session	
Mar 5	2. –Statistical models in social sciences. Regression analysis – what is it? Examining data. Transforming data.	ARAGLM – Ch.5-6, CAR - Ch. 4.1-4.4, 5.1-5.2
Mar 12	3. OLS regression - estimation, parameters and goodness of fit measures,	ARAGLM – Ch.5-6, CAR - Ch. 4.1-4.4, 5.1-5.2
Mar 18	4. OLS regression - estimation, parameters and goodness of fit measures,	ARAGLM – Ch.5-6, CAR - Ch. 4.1-4.4, 5.1-5.2
Mar 25	5. OLS regression - statistical inference	ARAGLM – Ch.5-6, CAR - Ch. 4.1-4.4, 5.1-5.2
Apr 8	6. Regression with dummy variables interaction terms	ARAGLM – Ch.7; CAR – Ch. 4.5-4.9, Brambor, Clark, Golder 2006;
Apr 22	7. Outliers and influential cases,	RD Ch. 4; CAR – Ch. 8
May 6	8. Regression assumptions –non-linearity	RD Ch. 7 & 8
May 13	9. Regression assumptions – collinearity, heteroscedasticity	RD Ch. 3, HiR Ch. 1 & 2
May 20	10. Introduction to General Linear Models – linear model vs. general linear model, linear predictor, link function.	Long Ch. 3 ARAGLM Ch. 14.1
May 21	11. Maximum Likelihood Estimation, Binary Logistic Regression vs. Probit models, Binary Logistic Regression – interpretation of parameters, predicted probabilities.	ARAGLM Ch. Ch. 15.1 Long Ch. 4
Jun 2	12. Binary Logistic Regression - goodness of fit measures. Binary Logistic Regression – interaction terms, interpretation using tools of statistical graphics.	Long Ch. 4, Fox 2003, CAR – Ch. 6
Jun 10	13. Multinomial logit – interpretation of the model parameters, interaction terms, predicted probabilities, goodness of fit measures.	Long Ch. 4, ARAM Ch. 14.2, Fox & Hong (2009)
Jun 17	14. Ordinal logit, Poisson regression and negative binomial model	Long Ch. 5, Ch. 8

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.