Course title: Logistic Regression Analysis for Social Scientist.

Instructor: Michał Kotnarowski.

Michał is an Assistant Professor at the Institute of Philosophy and Sociology of the Polish Academy of Sciences. He specialises in voting behaviour, comparative politics and political methodology. He has contributed a number of articles to journals including, *Party Politics*, *Communist and Post-Communist Studies*, *Acta Politica*, and the *International Journal of Sociology*. Michał is also a fan of Moravian wine, Czech literature and film. He likes spending holidays in the Moravian region, cycling slowly and tasting wine.

Prerequisite Knowledge

Course participants should have at least basic knowledge in the following topics:

1. Understand the logic of inferential statistics. Participants should be familiar with hypothesis testing and concepts such as confidence intervals and significance level. If you would like to refresh or complete your knowledge, I recommend the following reading: Wheelan 2014 – Chapters. 8, 9, 10 (more intuitive approach) or Healey 2011 – Chapters 6, 7, 8, 9 (more technical approach, but still easy to understand).
2. Participants should be familiar with the rudiments of the linear regression model estimated using the Ordinary Least Squares method. Course participants should have at least a basic understanding of the following topics: logic of linear regression analysis, assumptions of the linear regression model and regression with dummy variables. To refresh or complete your knowledge, I recommend: Wheelan 2014 – Chapters. 11, 12 or Healey 2011 – Chapters 14, 16.
3. Basic skills in R statistical environment. Participants should be able to import into R a dataset written in the SPSS format and run a linear regression model. They should also know how to conduct a set of elementary data manipulations in R, such as: selecting observations, selecting variables and computing new variables using existing variables. If you new to R or would like to refresh your R skills, I recommend: Navarro 2020: Chapters 3, 4 and R for Data Science e-book – Chapter 5 (available at https://r4ds.had.co.nz/transform.html)

Course Outline

Researchers working in broadly defined social sciences often have to deal with analyses in which the dependent variable is not a continuous variable defined on the interval scale. 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),
4. a variable counting the number of occurrences of a phenomenon (e.g., how many times a respondent participated in protest actions).

For this type of dependent variable, it is not appropriate to use Ordinary Least Square (OLS) regression models but General Linear Models (GLMs), which are estimated in a different way from linear regression models.

In this course, we will focus on logistic regression models, which are special cases of GLMs. Although logistic regression models are often used in social sciences, their use and correct interpretation still give researchers difficulties. During the course, participants will learn the foundations of logistic regression models as well as their advanced applications. The course will address both (1) practical problems in applying the logistic regression models and (2) statistical theory necessary to understand how these models work.

Day to day schedule and readings.

Day 1 Linear and non-linear probability models.
We start by discovering why it is inappropriate to use OLS models in case of a binary dependent variable. In particular, we will indicate which OLS model assumptions are not met and the negative consequences of using OLS models for this type of data. Next, I will show you how to generalise a linear model to be applied to models with a limited dependent variable. You will learn the linear predictor and the link function.

Readings:
Long 1997: Sections 3.1, 3.4.
Fox 2016: Sections 14.1 and 15.1.
Hosmer, Lemeshow, and Sturdivant 2013: Sections 1.1, 1.2.

Day 2 Latent variable model, Maximum Likelihood Estimation, Significance of goodness of fit.
On the second day, participants will learn how the logistic regression model can be understood as a latent variable model. Then I will introduce the Maximum Likelihood Estimation as a technique for estimating logistic regression model parameters. Finally, we discuss the goodness of fit measures, including various versions of pseudo-R-squared.

Readings:
Long 1997: Sections 3.2, 3.3, 3.5, 3.6; Chapter 4.
Hosmer, Lemeshow, and Sturdivant 2013: Sections 1.3-1.5, Chapter 2, Chapter 5.

Day 3 Interpretation of the model: coefficients, odds ratios, predicted probabilities and marginal effects.
Participants will develop practical skills related to the interpretation of the binary logistic regression model, i.e. the interpretation of regression coefficients, odds ratios, and the interpretation using predicted probabilities, in particular through statistical graphics.

Readings:
Long 1997: Sections 3.7-3.8.
Hosmer, Lemeshow, and Sturdivant 2013: Chapter 3,
Fox 2003,
Mood 2010.

Day 4. Models for polytomous dependent variable - multinomial logistic regression

On the last day, participants will learn multinomial logistic regression models - an extension of binary logistic regression models to cases where the dependent variable contains more than two categories. We will practice how to correctly estimate the multinomial logistic regression model, evaluate its goodness of fit, and interpret the results.

Readings:
Long 1997: Chapter 6
Hosmer, Lemeshow, and Sturdivant 2013: Section 8.1,
Fox and Hong 2009.

Software Requirements: the newest version of R and R Studio

Literature.

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).

———. 2016. *Applied Regression Analysis and Generalized Linear Models*. Third Edition. Los Angeles: SAGE.

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.

Healey, Joseph F. 2011. *Statistics: A Tool for Social Research*. 9th ed. Belmont, CA: Cengage Learning/Wadsworth.

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. Third edition. Hoboken, New Jersey: Wiley.

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

Mood, Carina. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It." *European Sociological Review* 26(1): 67–82.

Navarro, Danielle. 2020. "Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners." http://compcogscisydney.org/learning-statistics-with-r.

Wheelan, Charles J. 2014. *Naked Statistics: Stripping the Dread from the Data*. 1. publ. as a Norton paperback. New York: Norton.