Instructors:	Dr Zbigniew Karpiński, Dr Marta Kołczyńska, Dr Małgorzata Mikucka	
Course title:	**Quantitative methods of data analysis: statistical theory and good practices.**	
	WINTER Semester	SPRING Semester
HOURS	45	30
FREQUENCY	3 x 45 minutes every week Mondays, 10 AM – 12.15 PM First class: October 15, 2018	3 intensive sessions in March, April, and May (in each session 2 consecutive days of 4 x 45 of instruction + 2 x 45 individual consultations)

Zbigniew Karpiński
e-mail: zbigniew.karpinski@startmail.com
Office hours: Wednesday, 12 – 2 PM

Marta Kołczyńska
e-mail: mkolczynska@gmail.com
Office hours: Monday, 12.30 – 2.30 PM

Małgorzata Mikucka
e-mail: mikucka.m@gmail.com
Office hours: upon appointment

THE COURSE CORRESPONDS TO TOPICS: Statistical analysis of data, regression analysis

ASSESSMENT

- Course open to all students
- Form of the course: lecture + computer lab + individual consultations
- The course will be conducted in English
- Credit requirements: homeworks and group research project

COURSE AIMS AND CONTENT

Data on individual opinions, attitudes, values, and behaviours, as well as quantitative "indices" describing countries and regions, and increasingly also automatically generated digital trace data, are now easily available to every interested user. Powerful computers allow researchers and analysts to process complex and large datasets, comprising hundreds of thousands of records, in a matter of seconds. Specialised statistical software, both proprietary and open-source, enables students and scholars to "mine", model and visualise the data in novel and innovative ways. While these processes seem to make it easier for individuals and organisations, including governments, to make more informed choices and shape "evidence-based policies", there is also a relatively large margin for misusing or abusing the data, which may result from a poor understanding of the underlying concepts and methods (e.g., using *p*-value as a rule for deciding between valid and invalid findings) or ill-shaped incentive structures (e.g., publication bias, *p*-hacking).

The purpose of this course is to teach students about the foundations of statistical analysis in the social sciences, beginning with the very basic concepts of statistical description, such as distribution and measures of central tendency and dispersion. While these concepts are rarely the sole focus of analysis, they are necessary building blocks for understanding the more complex tools and approaches, including linear regression and its extensions. The course covers two main extensions of linear regression. The first is the multilevel modelling, which is currently the method of choice in cross-national analyses and allows estimating the "effects" of macro-level variables, e.g. unemployment rate, on individual-level outcomes. The second extension are regression models for panel data (fixed and random effects models) to analyse change over time. During the course, the focus is less on the theory behind the methods and more on their correct application and on the interpretation of their results.

Class time will combine (1) a lecture presenting elements of statistical theoretical underpinnings of the discussed methods, and (2) a computer lab which allows students to acquire practical experience with statistical analysis and modelling of the data.

EDUCATIONAL OUTCOMES

Knowledge:

- Knowledge of the basic notions used to describe data (e.g. variable and level of measurement, sample, population, distribution, mean, variance etc.);
- Familiarity with the concepts describing relationships among variables (e.g. statistical independence, correlation, covariance);
- Understanding of linear regression analysis, knowledge of its assumptions and applicability;
- Knowledge of selected extensions of linear regression (logistic regression, multilevel models, panel data analysis) and understanding when and why they should be used;
- Awareness of the importance of replicability in quantitative research;

Skills:

- Preparing the survey data for analysis and understanding their limitations;
- Correctly applying a particular statistical model to the data using STATA statistical software;
- Correctly interpreting the results of such analysis;
- Justifying the choice of model in an analysis; discussing the limitations and advantages of alternative statistical tools;
- Basic understanding of data visualization;
- Ability to search for and find data for secondary analysis in data archives and repositories; ability to use data documentation;
- Orientation in the types of data offered by cross-national survey projects;
- Effective organization of work on the project to ensure replicability of results.

Social Competence:

- Working in a team: organizing work, allocating tasks, achieving objectives using varied competences and skills of team members
- Providing feedback to other students on research projects
- Presentation skills: structuring a presentation, public speaking

READINGS

First semester

A recommended reading for the topics listed above is the text *Statistical Methods for the Social Sciences* by Alan Agresti and Barbara Finley.

Second semester

Readings are provided in the spring semester schedule below.

COURSE REQUIREMENTS

Class participation. This course has an intensive schedule and it is vital that students come to all classes prepared, having read the assigned readings and reviewed the material from previous sessions. Students should actively participate in all class activities.

This course will increase in difficulty as the semester goes on. Thus, it is important to develop good study habits early on. While some of the material at the beginning of the semester may seem remedial, this knowledge is the basis for much more difficult analysis later on.

Homework assignments (60% of the final grade). The homework assignments (4 in the first semester, 3 in the second semester) are designed to facilitate your understanding of the concepts and problems addressed during the class. Homework assignments must be turned in by the due date specified in the assignment instructions and announced in class. Late assignments will not be graded. While studying with another student is permitted, the homework write-up must be your independent work.

Group project (40% of the final grade). At the beginning of the course you will be divided into groups of 3-4 persons. Each group will consist of students of possibly similar research interests and possibly varied initial competence in statistical analysis. During the year, each group will work on a research question related to their research interests and will attempt to answer this question by analyzing survey data. We encourage consultations with the instructors during this work. At the end of the first semester you will be asked to present your project’s outline, including the research question, data, and preliminary analyses using the methods covered so far. During the last session of the course you will present your group’s project and discuss the projects of other groups.

Schedule for the Winter Semester

Dates	1	2	3
Week 1: Oct 15 – 19	Organisation. Introductions. Screening test.	Basic concepts: population and sample, variable and measurement, levels of measurement, description and inference, types of quantitative research	
Week 2: Oct 22 – 26	Distributions and their visualisation: univariate distributions, cumulative distributions, joint distributions, conditional distributions. Discrete and continuous distributions.		Theory – Methods – Data triad
Week 3: Oct 29 – Nov 2	Measures of central tendency. Positional parameters: median and other quantiles.		Sources of secondary data: data archives, repositories. Cross-national surveys
Week 4: Nov 5 – 9	Measures of dispersion. Variance and standard deviation. Measures of variation for non-metric variables. Standardized scores (*z*-scores).		Mini presentations by teams: Research topic, question, ideas for datasets
Week 5: Nov 12 – 16	Probability distributions for continuous and discrete random variables. Normal distribution. Properties of distributions: critical values, degrees of freedom. Statistical tables.		Intro to Stata: do-files, log files, data formats, opening data
Week 6: Nov 19 – 23	Random sampling, sampling from a distribution. Central Limit Theorem. Sampling distributions.		Intro to Stata: generating variables, recodes, subset
Week 7: Nov 26 – 30	Statistical inference about a population parameter: estimators and their properties, point estimates, confidence intervals. Type I and Type II Error. Effect sizes, power		Stata lab
Week 8: Dec 3 – 7	Null hypothesis significance testing: one-sample t-test. Comparison of two groups: independent and dependent samples.		Stata lab
Week 9: Dec 10 – 14	Variance decomposition. ANOVA. Analysing association between categorical variables: Contingency tables, Chi-squared test of independence.		
Week 10: Dec 17 – 21	Replicability and reproducibility. Preregistration, publication bias, *p*-hacking, multiple comparisons problem (and corrections)		Stata lab
Week 11: Jan 7 – 11	Correlation: Pearson's *r*, Spearman's ρ: assumptions, interpretation, visualization.		Stata lab
Week 12: Jan 14 – 18	Linear regression: assumptions, interpretation, and diagnostic. Univariate and multivariate regression.		Stata lab
Week 13: Jan 21 – 25	Logistic regression: assumptions, interpretation, and diagnostic.		Stata lab
Week 14: Jan 28 – Feb 1	Review	Project presentation	

Schedule for the Spring Semester

Homework assignment: Linear regression	
March session	**Multi-level models.** Why use multilevel models? Examples of problems which require multilevel models. Assumption of independence of observations. Hierarchical data and problems related to higher-level sampling. Random intercept and random slopes. Interpretation of coefficients. Cross-level interactions. How to present the results of multi-level analysis in oral presentation and in an article? **Reading:** Luke, D. A. (2004). *Multilevel modelling* (Vol. 143). A Sage University Papers Series: Quantitative Applications in the Social Sciences. Selected chapters.; Hox, J. (2010). *Multilevel analysis: Techniques and applications* (edition 2). Routledge. Selected chapters.
Homework assignment: Multilevel models	
April session	**Models for panel data.** Why use longitudinal data? Longitudinal vs. cross-sectional data. Which problems require using longitudinal data? Balanced and unbalanced panels; attrition and refreshment samples; individuals and households. Modelling of within person changes (fixed effects models, first difference models) vs. modelling between person differences of the dependent variable (between effects). Random effects models, their strengths and limitations. How to present the results of panel data analysis in oral presentation and in an article? **Reading:** Longhi, S. and Nandi, A. (2015). A Practical Guide to Using Panel Data. Sage. Selected chapters.
Homework assignment: Models for panel data	
May session	**Working with quantitative projects.** Advantages of using script files (do files in Stata) for replicability of the results. Use of working folders and paths. Automated export of results in the form of tables and figures. Documenting your work: versions, folders and subfolders, comments. **Good practice in regression analysis.** Which variables to include as controls and which to exclude from a model? Omitted variable bias. How to understand and discuss the size of effects? When do we need interaction terms? Various types of interaction terms and their interpretation. Non-response in survey data and the use of weights.
Presentation and discussion of research projects	

Note: The dates provided here are tentative and may change depending on how the class proceeds. Some topics may take a bit more time than we have allowed for, and others may take somewhat less time. Any changes in dates will be announced in class.