Government of Russian Federation

Federal State Autonomous Educational Institution of Higher Professional Education
«National Research University Higher School of Economics»

Faculty of Social Sciences
Master’s programme 'Politics. Economics. Philosophy'

Syllabus for the course “Data Analysis in the Social Sciences”

Author:
I. Schurov ischurov@hse.ru

Moscow, 2016

1. Scope of Use

This course is an introduction to key quantitative approaches to the analysis of data in social sciences. The course is taught in the form of lectures, seminars and self-running sessions which include individual and group work. All teaching is conducted in English.

2. Learning Objectives

Within this course you will:

- learn about the principal steps of a quantitative research in social sciences;
- learn about the possibilities and limitations of quantitative approaches as applied to different research questions;
- learn to formulate research questions and develop them into testable hypotheses;
- explore the possibilities of data collection and different approaches to sampling;
- learn to evaluate the quality of a quantitative approach;
- learn about the possibilities and limitations of conventional statistical techniques and criteria, as well as some popular contemporary multivariate statistical methods;
- learn to choose and apply in practice a set of appropriate statistical tests for your research question.

3. Learning Outcomes

On completion of the course, the student will be able to:

- account for basic types of data used in social sciences;
- apply basic quantitative methods for analysing data;
- understand research papers that use quantitative analysis;
- critically discuss the limitations of commonly used methods for answering research questions;
- reason on how to interpret the results of quantitative methods, including how to evaluate what kind of information a given method can offer and how to estimate the potential range of variables that can affect results in the research;
- apply different techniques for presenting quantitative data in scholarly writing.

4. Role of the course within the structure of Master program

The course aims to provide students with knowledge and competencies necessary to plan and conduct research projects of their own leading to M.Sc. dissertation and scientific publications.

Prerequisites: "Mathematics" course (adaptation discipline).

5. Schedule

1. Introduction to R. Types of data. Dataframe. User's functions.
    a. Base R objects. Simple plots. Simple functions.
    b. Dataframes. Read and write data. Manipulation with data (R base vs. dplyr)
    c. Base R functions code. User's functions. Loops. Functions in functions. Default arguments. Optional arguments.
2. Visualization.
    a. Base R vs. ggplot2: 2 variables, 3 variables, many variables
    b. RMarkdown. Tables in RMarkdown. Mixing different languages in RMarkdown
3. Descriptive statistics. Confidence intervals. T-tests. P-values. Normality tests
    a. Descriptive statistics, boxplot, violinplot. NAs.
    b. T-tests. P-value.
    c. Z-score. Confidence intervals. CI vs. p-values.
4. Count data.
    a. χ^2-test.
    b. Fisher's exact test.
5. Correlations.
    a. Pearson product-momentum.
    b. Kendall.
    c. Spearman.
6. ANOVA.
7. Regressions.
    a. Simple linear regression.
    b. Polynomial regression.
    c. Dummy variables.
    d. Significance of predictors.
    e. Ridge regression.
    f. Model selection.
8. Binary outcomes.
    a. The likelihood theory.
    b. Logistic regerssion.
9. Mixed-effects models.
10. Classification problems.
    a. Decision tree.
    b. Bootstrap technique.

    c. Decision forest.
11. Clusterization.
    a. Different types of distance matrices. K-means
    b. Hierarchical clusterization. DBSCAN.
12. Dimension reduction.
    a. Principal Component Analysis.

6. Requirements

The student is expected to

- be able to formulate the research problem in formal terms;
- know all the relevant notions;
- understand the theoretical background of methods discussed as well as their limitations;
- be able to use the software to process data;
- be able to give correct interpretation of the output of the software in terms of the research problem.

7. Assessment

The course is examined through continuous assessment of written assignments and the final exam.

Written assignments includes theoretical tests and practical problem-solving. The assignments are published online. The deadline for each assignment is specified upon publishing and will never be postponed.

The assignments should be submitted via an electronic form. The submission after the deadline will lead to penalty: 10% for delay within 1 hour, 20% for delay within 1 week, 50% for delay within 1 month, 90% for delay for more than 1 month.

The grade of every written assignment is a floating point number from 0 to 10. The average of all written assignments (with equal weights) rounded to integer with Google Spreadsheet's ROUND function is student's Cumulative Score.

The exam is conducted in class in the form similar to assignments. During the exam student may use any information sources (open book policy) except communication with other persons. Any attempt to communicate will be considered as violation of academic ethic policy and lead to permanent ban from the exam with exam score '0'.

The Final Score is obtained from the following formula:

Final Score = 0.6 × (Cumulative Score) + 0.4 × (Exam Score).

Rounding with Google Spreadsheet's ROUND function is applied.

8. Course Description

Introduction to R. Types of data. Dataframe. User's functions.

Basic R objects. Simple plots. Simple functions. Dataframes. Converting data. Reading and writing data. Attaching data. Working with Unicode and other encodings.

Manipulation with data (R base vs. dplyr). Basic R functions code. User's functions. Loops. Functions in functions. Default arguments. Optional arguments.

R Studio.

Visualization.

Base R vs. ggplot2: two variables, three variables, many variables. RMarkdown. Tables in RMarkdown. Mixing different languages in RMarkdown.

Descriptive statistics. T-tests. P-values. Normality tests. Confidence intervals.

Descriptive statistics, boxplot, violinplot. NAs.

Null and alternative hypotheses. Statistical significance. Significance level. T-tests. P-value. Misconceptions about p-value. One-tailed and two-tailed tests. Wilcoxon and Mann-Whitney tests.

Z-score. Confidence intervals. Standard error. CI vs. p-values.

χ^2-test. Fisher's exact test.

Contingency tables. Observed and expected frequencies. The $\chi 2$ -test of independence. Mosaic plots. $\chi 2$ -test and big data. Effect size metrics. Fisher's exact test.

Correlation

Relationship between two quantitative variables. The Pearson product-moment correlation coefficient. Correlation vs. paired t-test. Spearman's *rho* and Kendall's *tau* rank correlation. Correlograms.

Regressions: linear and polynomial

Linear regression with several explanatory variables. Organizing data and making hypotheses. Selecting the explanatory variables. Checking for overfitting. Linear regression with categorical explanatory variables. Polynomial regressions.

ANOVA

Analysis of variance (ANOVA): finding differences between several groups. Reporting results of ANOVA. Repeated-measures and mixed ANOVA. Other types of

ANOVA: analysis of covariance (ANCOVA), multivariate analysis of variance (MANOVA).

Logistic regressions

Binomial logistic regression modeling. Logistic function. Intercept and slope. Checking for overfitting. Polytomous logistic regression modeling.

Mixed-effects models

Fixed and random effects. Research design with random-effect variables. Random intercepts. Random slopes. Linear mixed-effect modeling. Logistic mixed-effect modeling.

Bootstrap. Decision trees. Decision forests.

Bootstrapping. When bootstrap?

Classification and regression trees. Bias and overfitting. Random forests. Importance of explanatory variables.

Clusterization

Distances in a multidimensional space. Distance metrics (Euclidean, Manhattan, Maximum). Bottom-up and top-down approach (agglomerative and divisive clustering). Methods of clustering (complete, single/nearest neighbors, average, ward). Which number of clusters is optimal? Validation of cluster solution. AU/BP values.

Dimension reduction

Multidimensional Scaling.

Principal Component Analysis. Contributions of components of PCA to explaining variance. Variables factor map. Individuals factor map.

9. Bibliography

1. Charles Wheelan. 2013. Naked Statistics: Stripping the Dread from the Data. W. W. Norton & Company; 1 edition.
2. Fox, John. 2008. Applied Regression Analysis and Generalized Linear Models. 2nd edition. Sage.
3. King, Gary. 1998. Unifying Political Methodology. Ann Arbor: University of Michigan Press.
4. Wooldridge, Jeffrey. 2009. Introductory Econometrics: A Modern Approach. 4th edition. South-Western College Pub.
5. Matloff, Norman. 2011. The Art of R Programming. San Francisco: No Starch Press.

10. Software

We will use open source statistical package R (http://r-project.org/) and its libraries. As a GUI, we will use RStudio (http://rstudio.com/), Rcmdr (http://www.rcommander.com/) and Deducer (http://deducer.org/). All the software used is free, open source and available on all major platforms.