Federal State Autonomous Educational Institution of Higher Education
"National Research University Higher School of Economics"

Faculty of Social Sciences
Department of Higher Mathematics

Course Syllabus
"Data Analysis in the Social Sciences"

For the Master Program 41.04.04 "Politics. Economics. Philosophy"

Author: Alla Tambovtseva
email: atambovtseva@hse.ru

Approved at the meeting of the
Department of Higher Mathematics

Moscow, 2018

This syllabus cannot be used by other university departments and other higher education institutions without the explicit permission of the Faculty of Social Sciences.

1 Course Description

1.1 Pre-requisistes

„Mathematics“ course (adaptation discipline)

1.2 Course Type

This course is compulsory for the master programme “Politics. Economics. Philosophy”.

1.3 Abstract

The course “Data Analysis in the Social Sciences” aims provide students with knowledge necessary to plan and conduct quantitative research in the social sciences. It covers different statistical methods including regression analysis, cluster analysis and methods of dimensionality reduction. The practical part of the course includes mastering data manipulation techniques as well as acquiring basic programming skills useful for data collection and analysis. Particular emphasis will be placed on effective ways to visualize and present the results of academic research. On completion of this course students will be able to process empirical data, use state-of-the-art statistical methods in their master dissertations or in academic publications.

2 Learning Objectives

Students will:

- develop programming skills relevant to data manipulation and analysis;
- conduct statistical analysis using popular statistical software;
- apply statistical concepts and methods to in the context of social science research;
- learn to develop substantial research questions into statistical hypotheses;
- interpret the results of quantitative research in the social sciences and present their own results effectively using suitable techiques of visualization and reporting.

3 Learning Outcomes

On completion of this course students will:

- gain the basic programming skills necessary to collect and process data;
- apply methods of data manipulation using R libraries;
- learn about the main steps of a quantitative research in the social sciences;
- adopt the principles of reproducible research;
- understand which type of visualization should be chosen to illustrate certain patterns or relationships in data.
- apply appropriate statistical methods for their social science research and interpret the results carefully;
- learn about the possibilities and limitations of different quantitative methods in the context of political science.

4 Course Plan

Section 1

Theory

Introduction to the quantitative research design. Population and samples. Types of sampling. Statistical estimates.

Data types. Continuous, interval, ordinal and nominal scales. Descriptive statistics for different types of data. Counts and proportions. Mean, median, quartiles, variance and standard deviation of a sample. Coefficient of variation.

Practice in R

Introduction to the RStudio interface. Opening and saving R files. Markdown as a markup language. RMarkdown and Rmd-files. R libraries: loading and installation. Vectors in R as a principal data structure. Basic data types in R: character, numeric, factor. Operations with vectors. Descriptive statistics in R.

Section 2

Theory

Visualization of data distribution. Histogram and boxplot for numeric data. Normal distribution: recall. Checks for normality: Shapiro-Wilk test and Kolmogorov-Smirnov test.

Practice in R

Data loading in R. Reading and writing csv-files. Loading data files with different extentions (xlsx, dta, sav) with the library `foreign`.

Data cleaning. Handling missing values. Basic operations with datasets: editing and filtering.

Basic plots in R via`plot()` function. Histograms, kernel density plots, boxplots and violinplots.

Section 3

Theory

Statistical estimation. Law of Large numbers. Central limit theorem. Features of estimates: unbiasedness, efficiency, consistency.

Practice in R

Generation of random variables (pseudo-random vectors) in R. Random sampling. Loops in R.

Section 4

Theory

Point estimates vs interval estimates. Confidence intervals. Confidence level and precision of estimates. Confidence interval for proportion, confidence interval for a mean, confidence interval for the difference of means.

Practice in R

Calculation of confidence intervals in R via the library `DescTools`. Visualization of confidence intervals.

Section 5

Theory

Statistical inference and hypotheses testing. Types of statistical hypotheses. Logic of statistical inference. P-value. Errors: I type error and II type error. Power of a statistical criterion. Statistical significance vs practical significance.

Examples of statistical tests. Z-test for the proportion, Student t-test for the mean, Welch test for the mean difference, Sign test.

Confidence intervals and hypotheses testing.

Practice in R

Hypotheses testing in R. Modifications of t-test, sign test.

Section 6

Theory

Measures of association between quantitative variables. Scatterplot. Correlation. Pearson and Spearman correlation coefficients. Correlation matrix. Correlation and causation.

Association between qualitative variables: contingency tables and χ^2-test (Fisher's exact test).

Practice in R

Visualization of association between quantitative variables: scatterplot and scattermatrix in R. Visualization of association between two qualitative variables: contingency tables and mosaic plot.

Calculating correlation coefficients in R and cheking their statistical significance. Correlation matrix. Visualization of a correlation matrix via the library `car`.

Section 7

Theory

Examples of exploratory analysis. Statistical inference on real data. Replication of research. Reproducibility and validity of results.

Practice in R

Data processing with the library `dplyr`. Grouping and aggregation of data. Advanced visualization via the library `ggplot2`: layers, aesthetics, grouping and facets.

Section 8

Theory

Analysis of variance (ANOVA): finding differences between several groups. ANOVA vs multiple t-tests for several groups. Bonferroni correction and the idea of corrections in statistics.

Practice in R

ANOVA in R: `aov()` function and the summary of the output. Multiple t-tests with corrections in R and reporting p-values.

Section 9

Theory

Regression vs correlation analysis. Simple linear regression. Ordinary least squares (OLS). OLS estimates and their features. Coefficient of determination R^2. Interpretation of regression results.

Practice in R

Linear regression in R. Reporting the results of regression analysis. Exporting regression tables to Word and pdf-files (LaTeX) via the library `stargazer`. Including tables in Rmd-reports.

Section 10

Theory

Multiple linear regression. Interpretation of regression results. Tests for model quality. Possible problems: multicollinearity, heteroskedasticity and influensive observations. Omitted variable bias. Information criteria: AIC and BIC.

Practice in R

Multiple linear regression in R. Tests for multicollinearity and heteroskedasticity. Types of stardard errors of regression coefficients. Detection of influensive observations: dfbetas and Cook's distance. Model comparison in R.

Section 11

Theory

Linear regression with dummy variables as predictors. Including qualitative predictors in linear models. Sets of dummy variables and the problem of multicollinearity. Interaction effects in linear models. Marginal effects. Regression with dummy variables and ANOVA.

Practice in R

Handling text variables in R. Factor variables and sets of dummy variables. Including interaction terms in linear models. Interpretation of coefficients in interaction models. Marginal effects: calculation and visualization.

Section 12

Theory

Nested data. Cross-sectional and time-series data. Time-series cross-section data (TSCS). Fixed and random effects in linear models. Challenges of nested data analysis. Social networks data and nested data.

Practice in R

Fixed and random effects models in R. Visualization of associations by groups in nested data. Comparing OLS models with fixed-effects models.

Section 13

Theory

Models with a binary outcome. Logistic regression. Quality of a logisitic model: precision, sensitivity and specificity. ROC curve and its interpretation. Analogues of R^2 for logistic models.

Practice in R

Generalized linear models in R via `glm()` function. Logistic and probit regression in R. Assessing the quality of binary models. Coefficients, exponents of coefficients and odds. Marginal effects in the context of logistic regression.

Section 14

Theory

Cluster analysis. Hierarchical cluster analysis: measures of distance and methods of aggregation. Dendrograms. Validating the results of cluster analysis. Comparing the results of clustering in R.

K-means clustering: idea and implementation. K-means vs hierachical clustering: advantages and disadvantages. Clustering in machine learning.

Clustering of binary data and measures of similarity.

Practice in R

Basic hierarchical cluster analysis via `hclust()` function. Plotting dendrograms and its variations. Handling dendrograms for large datasets. K-means algorithm in R. Comparing the results of clustering in R. Introduction to geospatial data in R. Visualization and clustering of geospatial data.

Section 15

Theory

The idea of dimensionality reduction. Principal component analysis (PCA). The variance of components. PCA and singular value decomposition.

Practice in R

PCA in R. Revision of correlation analysis. Application of PCA as a way to avoid multicollinearity in linear models.

5 Reading List

5.1 Compulsory

1. Diez, Daniel et al. 2015. OpenIntro Statistics.

2. Wheelan, Charles. 2013. Naked Statistics: Stripping the Dread from the Data. W. W. Norton & Company; 1 edition.

5.2 Optional

1. Gohil, Atmajitsinh. 2015. R Data Visualization Cookbook. Packt Publishing.

2. Wooldridge, Jeffrey. 2009. Introductory Econometrics: A Modern Approach. 4th edition. South-Western College Pub.

6 Grading System

The course is examined through continuous assessment of written home assignments, written quizes in class, the midterm and the final exam.

Written home assignments include theoretical tests, practical problem-solving and labs in R. 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 a Google form or uploaded via Dropbox (in some cases works in writing are allowed). The submission after the deadline will lead to penalty: 10% for delay within 1 hour, 20% for delay within 1 day, 50% for delay within 1 week. Assignments submitted one week later are not graded, but a grader can provide a feedback on demand.

Written quizes are 15-minutes theoretical tests in class that cover the topics discussed at the previous lectures. Students are notified in advance about such quizes. During the quiz it is not permitted to use any information sources (closed book policy). Quizes include multiple-choice questions and open questions with a short answer. Quizes cannot be re-written or written later in case of absence at the quiz (regardless the circumstances).

Midterm is conducted in class and consisits of a theoretical part (test questions and statistical problems to solve) and a practical part (lab in R). It takes 2 academic hours. During the theoretical part of the midterm students cannot use any sources except their own A4 list with information in a handwritten form prepared before the midterm. During the practical part of the midterm students may use any information sources (open book policy) except communication with other persons. The midterm can be written on another day than announced as long as a student provides a medical evidence to the study office.

The final exam is conducted in class in the form similar to written assignments and the midterm. During the exam student may use any information sources (open book policy) except communication with other persons.

Academic ethic policy. Any attempt to communicate (including online communication) during the midterm/exam will be considered as a violation of academic ethic policy and will result in the score 0 for the midterm/exam. If plagiarism is detected in any graded submissions (including home assignments), all works involved will get the score 0. The administration of programme will be notified in a written form about any cases of academic misconduct.

Cumulative and final scores. Score for home assignments (Home assignments) is calculated as the average of scores for all assignments rounded to the closest integer. Score for quizes (Quizes) is calculated as the average of scores for all quizes rounded to the closest integer. Final score is rounded to the closest integer as well.

$$\text{Cumulative Score} = 0.4 \times \text{Midterm} + 0.4 \times \text{Home assignments} + 0.2 \times \text{Quizes}$$

$$\text{Final Score} = 0.6 \times \text{Cumulative Score} + 0.4 \times \text{Exam}$$

7 Guidelines for Knowledge Assessment

7.1 Criteria

All theoretical assignments are checked to fulfil the following requirements: correct usage of statistical concepts, understanding key features of data types and distributions, ability to identify a scope of use and limitations of different quantitative methods, and adequate interpretation of results.

All practical assignment in R are checked to fulfil following requirements: correctness and reproducibility of code, adequacy of methods applied, correspondence between data types and chosen methods of analysis.

If serious mistakes are detected in a student's answer (examples of such mistakes: negative variance or standard deviation, correlation coefficient more than 1 by absolute value, negative probabilities or probabilities more than 1), the answer can get the score 0. If a student's R code cannot be reproduced due to syntax errors, the code and the interpretation of the statistical output can be graded as 0.

7.2 Examples of exam tasks

Problem 1. Match the corresponding purpose with the method/model that should be used to get desirable results. Mind the assumptions that lie behind each method and than make your choice.

1. construct integral index of economic prosperity based on different social and economic indicators
2. evaluate the effect of freedom of the press operationalized as the Freedom of the Press Index (Freedom House) on the level of corruption represented by Corruption Perception Index
3. decide how students can be grouped based on their abilities of different kinds (set of numeric scores for many tests)
4. evaluate the effect of age, gender and political preferences (place on the left-right scale) on people's propensity to participate in actions of civil unrest (0 – not participate, 1 – participate)
5. evaluate the effect of economic development measured as the GDP per capita on the political stability operationalized as the Political Stability Index taking into account ethnic heterogeneity and the level of social inequality (suppose that we have the panel data and want to somehow capture individual effect of predictors in each country)

a. cluster analysis

b. principal component analysis

c. simple linear regression

d. logistic regression

e. linear regression with fixed effects

Problem 2. A researcher wants to organize countries in groups based on the indicators he has in the dataset, but he has no idea how many clusters he want to obtain. He wants to see the whole picture and than decide how deep classification needed should be. Which approach should the researcher use: K-means or hierarchical clustering? Explain your answer.

Problem 3. The Russian party “Yabloko” proclaim themselves as the social liberal party. The electorate of this party is mainly concentrated in big cities. One student decided to check statistically whether the share of urban population in a region positively influenced the share of votes for this party at the parliamental elections in 2011. He performed the bivariate linear regression and got the following output:

```
##
## Call:
## lm(formula = Yabloko ~ urban, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4835 -0.6980 -0.1776  0.4906  6.7773
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.33607    0.87116  -4.977 3.84e-06 ***
## urban        0.10489    0.01233   8.505 1.08e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.394 on 77 degrees of freedom
## Multiple R-squared:  0.4844, Adjusted R-squared:  0.4777
## F-statistic: 72.34 on 1 and 77 DF,  p-value: 1.078e-12
```

1. How does the vote share of “Yabloko” change when the share of urban population increases by one percentage point?

2. Can we conclude that the share of urban population really affects the support of the party “Yabloko”?

3. Provide the interpretation of the determination coefficient (R-squared) from this output, i.e. decide based on its value whether the quality of the model is acceptable.

4. What is the Pearson correlation coefficient between the share of urban population and the vote share of “Yabloko”?

Problem 4. One political science student decided to compare GDP per capita in free and not free countries (in the Freedom House notation). He wants to prove that the mean value of GDP per capita differs in these two types of states.

1. What is the null hypothesis in this research?

2. What can be considered as the Type I error in this research?

3. What can be considered as the Type II error in this research?

The student chose 90 countries, performed t-test and got the p-value of 0.015.

1. Interpret the p-value in the context of this study.

2. What conclusion should be made at the 5% significance level?

3. Considering the conclusion in 1.4, what error, Type I or Type II, might the student make?

Problem 5. You are suggested to conduct a small research on the political self-identification of the Spanish people. Your question of interest is the following: which factors affect the people's propensity to identify themselves as advocates of right-wing policy? You are provided with a dataset with the results of the survey conducted in 2014 in Spain. It contains the following variables:

- ideolog : respondents's position on the ideological spectrum, 1 – right, 0 – left
- age : respondent's age
- ident_reg : equals 1 if a respondent indentifies themselves with a region (province in Spain)
- ident_cntr : equals 1 if a respondent indentifies themselves with a country (Spain as a whole political unit)
- male : equals 1 if a respondent is male, 0 – female
- educ : respondent's level of education (1 – primary school, 4 – higher education)
- unempl : equals 1 if a respondent is unemployed, 0 – otherwise

Make a regression model that would help you to decide which factors mentioned above affect the people's position on the ideological spectrum.

1. What is the dependent variable in your model?
2. What are the independent variables in your model?
3. What type of the regression you are going to use? Explain your choice.
4. Make the model. Provide the R code you used to perform the model.
5. Which of the factors significantly affect the people's position on the ideological spectrum? At what level of significance?
6. Interpret the coefficient of the variable age, i.e. explain what happens when the age of a respondent increases by one year.
7. Interpret the coefficient of the variable male, i.e. explain what happens when we move from a female respondent to a male one.
8. How does the logarithm of odds differ if we compare its value for unemployed and employed people?

8 Methods of Instruction

The course is taught in the form of lectures and seminars including computer labs. All teaching is conducted in English. Materials for the course including files with data and R code are published online, on the website *math-info.hse.ru* and on Github.

9 Special Equipment and Software Support

For this course students need the statistical package R (http://r-project.org) and RStudio (http://rstudio.com). All lectures and seminars should take place in a computer class with a projector. A projector is necessary for programming labs since students have to replicate lecturers' code during classes.