Syllabus on the course “Data analysis”

Approved by Programme Academic Council

Process-verbal 2 from April 10, 2018

Author	Sergey V. Petropavlovsky
Credits	3
Academic Hours	114
Year of study	1
Mode of study	Full-time

1. Pre-requisites

Programming (R is a plus but not essential), mathematics (algebra and calculus), probability theory and statistics. Good command of English.

2. Course Type

The course Data Analysis is mandatory for master students enrolled for the program “Big Data Systems” other than those having the Bachelor Degree from HSE. For HSE graduates the course is optional.

3. Abstract

"Data Analysis" is taken in the first module of the Master’s program. The course is designed to refresh basic statistical and data analysis techniques and thereby prepare students for the subsequent more advanced disciplines. In the first part of the course we review basic methods of descriptive and inferential statistics including data visualization and point estimation, interval estimates, hypothesis testing, linear regression (univariate and multivariate) and analysis of variance. The second part of the course covers selected modern approaches such as dimension reduction methods (principal component analysis and extensions), classification algorithms (cluster analysis, linear discriminant analysis, logistic regression) and foundations of statistical learning. The emphasis is put on practical implementation of the algorithms in the first place, so a brief theoretical exposure to each topic is accompanied by multiple examples. The students are supposed to use the R language for doing analysis throughout the course (but not limited to), so a brief introduction to R is done at the very beginning. The duration of the course is one module.

4. Learning Objectives

The course provides a review of major data analysis techniques and aims at developing practical skills of data acquisition, processing and interpretation.

5. Learning Outcomes

At the end of the course, students are supposed to:

Be aware of:

- the need, applicability and basic concepts of data analysis;

- principles of data collection and pre-processing;
- basic algorithms and techniques of modern data analysis;
- details of implementing the algorithms using modern software.

Be able to:

- select the appropriate algorithms of data analysis as per the research goals;
- collect and pre-process the raw data;
- apply a variety of methods for explaining, summarizing and presenting data and interpreting results clearly.

Learn how to:

- search for, select and download data for the subsequent analysis;
- conduct the data analysis using modern software;
- present the results of the analysis.

6. Course Plan

Topic 1. Introduction to R.

Data objects in R: data frames, matrices and arrays, factors, lists. Indexing of data frames, conditional selection. Installing and using packages. Loading data from local files and on-line databases. Handling missing values. The R environment: session management, the graphics subsystem. Plotting data in R. Time series objects in R. Overview of basic statistical functions in R. Major programming constructs: conditional operators, loops, functions.

Topic 2. Descriptive Statistics

Types of data. Graphical presentation of univariate data: bar chart, histogram, stem-and-leaf display. Kernel density estimators. Measures of central tendency: mean, median. Assessing shape of distribution. Measures of variability: range, sample standard deviation, interquartile range. Quartiles and quantiles. Numerical summaries of the data in R. Boxplots. Bivariate data: side-by-side boxplots, quantile-to-quantile plots, scatter plots. Pearson's and Spearman's coefficients of correlation. Kendall's τ measure. Bivariate boxplots. The convex hull of the bivariate data. Bubble and glyph plots for multivariate data. Scatter plot matrix.

Topic 3. Inferential Statistics

Interval estimates. The idea behind confidence intervals (CI). CI for a parameter of single population (mean, proportion, variance). CI for difference between means and proportions.

Hypothesis testing. The general concept of hypothesis testing. Type I and Type II errors. Test statistic. Decision rules: rejection regions, p-values. One-sample tests: z- and t-test for a single mean, test for a single proportion. Two-sample tests: test for difference between means and

proportions. Paired samples in hypothesis testing. Wilcoxon rank-sum test for two independent samples.

Chi-squared goodness-of-fit test for association between the categorical variables. The chi-squared test for homogeneity.

Goodness-of-fit tests for continuous distributions. Kolmogorov-Smirnov test and the Shapiro-Wilk test for normality.

Topic 4. Linear Regression

The simple linear regression model. Estimating the parameters in simple linear regression. Linear regression in R. Statistical inference for simple linear regression. R-squared and adjusted R-squared coefficients. Model diagnostics: assessing normality of the residuals. The model goodness-of-fit statistics: F-test. t-test for model coefficients, confidence intervals for model coefficients. Prediction intervals. Confidence intervals for the correlation coefficient.

Introduction to multiple linear regression. Multiple regression assumptions, diagnostics, and efficacy measures. Fitting the multiple regression model in R. Interpreting the regression parameters. Model selection: paired F-test, the Akaike information criterion. Problems with many explanatory variables: multicollinearity. Remedies for multicollinearity. Confusion between explanatory and response variables.

Topic 5. Dimension Reduction Methods.

Geometrical view on data. Optimal projecting onto a low-dimensional space. Principal component analysis (PCA). Data for PCA. Different approaches to PCA. PCA through the singular value decomposition. PCA via diagonalization of the covariance matrix. Coordinates of individuals and variables in the reduced basis. The transition formulae. Quality of projecting and individual contributions into construction of new dimensions. Interpretation of PCA output. Simultaneous analysis of individuals and variables. PCA implementation in R.

Correspondence analysis (CA). Data for CA, chi^2 metric. Profiles of rows and columns. CA implementation in R. Quality of projecting and individual contributions into construction of new dimensions. Link between row and column representations. Multiple CA. Indicator matrix.

Multidimensional scaling (MDS). Dissimilarity matrices. Goals of multidimensional scaling. Computing dissimilarities: Euclidean and non-Euclidean distances. Classical multidimensional scaling. Metric and non-metric MDS. Goodness-of-fit measures for the metric MDS. Shepard's diagrams. Distance scaling. Issues of the non-metric MDS. Interpretation of the MDS analysis.

Embedding external variables.

Topic 6. Classification Methods

Logistic regression. Estimating the regression coefficients and making predictions. Logistic regression with several variables. Case-control sampling and logistic regression. Logistic regression with more than two classes.

Linear discriminant analysis (LDA). Using Bayes' theorem for classification. Discriminant functions. Fisher's discriminant plots. Advantages and downsides of LDA. Naive Bayes approach. Quadratic discriminant analysis. K-nearest neighbor algorithm.

Cluster algorithms. Distances between clusters (linkage). Agglomerative hierarchical clustering (AHC). Constructing an indexed hierarchy. Ward's algorithm. Quality of partition. Agglomeration according to inertia. Properties of the agglomeration criterion. Impact of different linkage type on the performance of the AHC. Direct search for partitions: K-means and K-medoids approaches. Probabilistic clustering: Gaussian mixture model (GMM). Expectation maximization algorithm. Clustering and principal component methods.

7. Reading List

Required

1. Husson, François. Exploratory Multivariate Analysis by Example Using R / François Husson, Jérôme Pagès, Sébastien Lê.–Taylor & Francis, 2017. – URL: http://web.a.ebscohost.com/ehost/detail/detail?vid=0&sid=fb7294f3-d029-4c54-b73a-c14351e4cac1%40sdc-v-sessmgr02&bdata=#AN=1516055&db=nlebk – ЭБС EBSCO eBooks 2017
2. Dramis, James. An Introduction to Data Analysis using Aggregation Functions in R.– Springer, 2016–URL: https://link.springer.com/book/10.1007%2F978-3-319-46762-7 – ЭБС Springer eBooks (Complete Collection 2016)

Optional

1. Shmueli, Galit. Data Mining for Business Analytics: Concepts, Techniques, and Applications in R / Galit Shmueli, Kenneth C. Lichtendahl Jr, Nitin R. Patel, Inbal Yahav, Peter C. Bruce .– Wiley, 2018– URL: https://library.books24x7.com/toc.aspx?bookid=142547 – ЭБС Books 24x7 IT Pro Collection

2. Mailund,Thomas. Beginning Data Science in R: Data Analysis; Visualization; and Modelling for the Data Scientist.– Apress, 2017– URL: https://link.springer.com/book/10.1007%2F978-1-4842-2671-1 – ЭБС Springer eBooks (Complete Collection 2017)

8. Grading System

The formula for the final grade O_{fin}

$$O_{\text{fin}} = 0.7 \times O_{\text{accm}} + 0.3 \times O_{\text{exam}}$$

is made up of the grade O_{accm} accumulated over the module and the grade O_{exam} for the final exam. The accumulated grade O_{accm} is calculated as follows:

$$O_{\text{accm}} = 0.6 \times O_{\text{HA}} + 0.4 \times O_{\text{MT}}$$

where O_{HA} and O_{MT} are the grades for the home assignments and the in-class tests, respectively.

9. Guidelines for Knowledge Assessment

Sample concept questions for final exam

1. Basic statistical measures of univariate data. Robustness to outliers.
2. Plotting data in R.
3. Discuss different approaches to assess correlation between data.
4. Outline the concept of interval estimates.
5. Describe the hypothesis testing process.
6. One-sample and two-sample statistical tests.
7. Describe the idea behind the principal component analysis.
8. What is the difference between principal component analysis and correspondence analysis?
9. Estimation and diagnostics of linear regression model.
11. Estimation and diagnostics of multiple regression model.
12. Linear discriminant analysis: applications, advantages and drawbacks.
13. Logistic regression as a classification tool.
14. Describe the difference between the hierarchical clustering and direct search algorithms.
15. Ward's algorithm.
16. K-nearest neighbor and naïve Bayes approaches for classification.

17. The concept of statistical learning and its limitations.

18. Support vector machines and their use in practice.

Sample practice assignments for final exam

1. Download the univariate market data from the on-line databases. Provide the summary statistics and discuss it.

2. Download the univariate market data from the on-line databases. Provide and describe a graphical summary of the data (plot, boxplot etc).

3. Download the univariate market data from the on-line databases. Detect the outliers, remove them from the data set and re-compute the summary statistics.

4. Download the bivariate market data from the on-line databases. Compute several measures of correlation and visualize the cloud of observations.

5. Download the univariate market data. Construct a 95% confidence interval for the mean and perform the appropriate t-test.

6. Download the univariate market data. Construct a 95% confidence interval for the proportion of data that exceed the mean and perform the appropriate t-test.

7. Download the univariate market data. Construct a 95% confidence interval for standard deviation and perform the appropriate F-test.

8. Download the bivariate market data. Construct a 95% confidence interval for the difference between the means and perform the appropriate t-test.

9. Perform a chi^2 test for association between two categorical variables.

9. Using the suggested dataset, build and test the simple linear regression model.

10. Using the suggested dataset, build and test the multiple linear regression model.

11. Using the suggested dataset, classify the data by means of hierarchical clustering.

12. Using the suggested dataset, classify the data by means of K-means algorithm.

13. Using the suggested dataset, classify the data using linear discriminant analysis.

14. Using the suggested dataset, perform the principal component analysis. Discuss the quality of representation.

15. Using the suggested dataset, perform the multidimensional scaling of data. Interpret the result.

Sample home assignment

Data sources for the assignment

There are at least three options:

1. Use getSymbols command of quantmod package to download prices for some stock or com-

modity (oil, gold, wheat, etc) from Federal Reserve Economic Data repository http://research.stlouisfed.org/fred2/ . You may want to try the following commands if download does not start:

options(download.file.method="libcurl") or options(download.file.method="wget")

or options(download.file.method="wininet")

- Clearly state in your report what kind of data you are using (daily, monthly etc).
- Check for the missing data and remove the respective entries from the dataset, if any.

You may use the following script as an example:

```
getSymbols(`GOLDAMGBD228NLBM', src='FRED')
idx <- c(1:nrow(GOLD))[is.na(GOLD)]
GOLD <- GOLD[-idx]
```

See also Section 1.3.3 of [1].

- If you did find the missing data, add a comment on that.

2. Use the built-in datasets provided by packages UsingR, MASS or ISwR. See the summary on p. 24 of [1] for listing and handling the available datasets. You may also look through the Problems in [1] to make a choice.

3. Multivariate data. You may try the following:

- Go to the JSE archive http://ww2.amstat.org/publications/jse/jse_data_archive.htm .

If are unsatisfied with the data from the previous source or these have been already picked up by your classmates, visit https://www.census.gov/data/tables/2015/econ/asm/2015-asm.html or, more generally, https://www.data.gov/ . However, some minor research and preprocessing may be needed here to get a meaningful and compact dataset.

Suggest your own dataset from some other source. Free sources of data are listed here http://guides.emich.edu/data/free-data . Some research is needed.

Remember that:

- There MUST be at least 20 observations (the more, the better).
- There MUST be at least 3 variables.

You may consider time periods (months, years, etc) as observations, i.e., time series will work.

Assignment on Descriptive Statistics

1. Univariate data. Get a univariate dataset from sources 1 or 2 and briefly describe it.

Using these data,

(a) Construct a stem-and-leaf and histogram. Impose the empirical density estimate on the histogram. Discuss the results focusing on the shape of the plots and number of modes.

(b) Compute the mean and median. Based solely on that, conclude whether the distribution is skewed. Find the proportion of the data which are less than the mean value.

(c) Compute the 1st and 3rd quartiles, the 90th quantile and the mode. Explain the meaning of the obtained quantities. Find the value that cuts off the top 25% of the data.

(d) Compute the range, the sample standard deviation and the IQR. Construct the boxplot of the data. Comment on the boxplot including skewness, outliers etc.

(e) Check whether the empirical distribution is normal by examining the QQ-plot. See examples in [1], Section 2.2.

2. Bivariate data. Get the bivariate data by: (i) retrieving prices for two stocks from data source 1 (of the equal length and periodicity!) or (ii) finding the appropriate bivariate built-in set 2

(a) Create side-by-side boxplots. Compare the centers and spreads.

(b) Draw the scatter plot. Comment on the possible dependence and presence of outliers.

(c) Compute Pearson's and Spearman's coefficient of correlation. Interpret and compare these values. Are these values consistent with the scatter plot?

(d) Add the marginal distributions to the scatter plot. For that purpose, use histogram and box plot.

(e) Depict the bivariate box plot. Comment on the outliers. Remove the outliers, if any, and re-compute the Pearson correlation coefficient.

(f) Create the convex hull. Remove the observations lying on the hull and re-compute the correlation coefficient.

For items 2a-2c, see [1], Section 3. For items 2d-2f, see [2], Section 2.2.

3. Multivariate data:

(a) Pick up a dataset which has three variables (from source 2 or 3) and create the bubble plot. Interpret the result. See [2], Section 2.3.

(b) Use data source 2 or 3. Create the glyph plot of all observations, Section 2.3. Do any stars look alike?

(c) Use data source 2 or 3. Create the scatter plot matrix and analyze it. See [2], Section 2.4.

10. Methods of Instruction

In general, lectures should give insight into the concepts and ideas underlying the topic under review. The theoretical core of presentation should be preceded and followed up by clear examples. The lecture slides may contain pieces of (quasi) code illustrating implementation of the algorithms in some programming language (presumably, in R). It is highly recommended to provide students with the lecture slides prior to the lecture so that they could familiarize themselves with the material in advance and prepare some questions. The lecturer should refer

the students for technicalities to the recommended textbooks, reviews and papers as needed throughout the presentation.

Practice classes play the key role in providing the course. The instructor should focus on the implementation of data analysis algorithms on computers. The difficult tasks should be discussed and worked out together with students. The tasks being discussed should be close to those of home assignment so as students could solve similar problems on their own. The students are supposed to prepare a report on a particular home assignment and submit it to the instructor electronically or in paper form. Some requirements for these reports may be set, e.g.:

- The questions should be addressed in the same order they appear in the assignment. The text of the question must be retained and placed before each answer. The working language is English.

- The answer to a particular question may take a form of a plot, formula etc followed by a brief explanation and a conclusion. All conclusions must be justified numerically, i.e., by some computed quantities, plots, etc. The answers do not need to be lengthy but they must be convincing in mathematical and statistical sense, i.e., in terms of some quantitative measures.

- Each student must use a unique data set. It is the student's responsibility to make sure that no one else is using the same data. To facilitate the distribution of datasets among the students, the instructor can create an editable shared check-in list on Google Drive or some other cloud resource.

- The deadlines for the reports should be clearly specified.

- The instructor should notify the students about the penalties for late submission of the reports.

- The solutions should normally contain code in R or some other language.

It is good practice to suggest the students some datasets for the home assignments. For example, a great amount of market data can be found at Yahoo Finance, Google Finance, Federal Reserve Economic Data repository http://research.stlouisfed.org/fred2/ and so on. Other possible data sources include the JSE archive http://ww2.amstat.org/publications/jse/jse_data_archive.htm, a huge repository at https://www.data.gov/ and a list of freely available sources at http://guides.emich.edu/data/free-data. Remarkably, most of these data can be downloaded in R directly by using the respective functions which should be pointed out to students.

11. Special Equipment and Software Support

#	Title	Terms of access

1.	Microsoft Windows 7 Professional RUS	Internal HSE network
2.	Microsoft Office Professional Plus 2010	Internal HSE network
3.	R programming language	Internal HSE network
4.	R Studio IDE	Internal HSE network