Undergraduate Program in International Relations

Data Analysis in R
(Fall/Winter 2017-18)

Course objectives:

To provide a comprehensive overview of R and cover the essential exploratory techniques for summarizing data and creating model simulations in R.

Learning Outcomes:

- Mastering R language fundamentals and basic syntax
- Becoming familiar with major R data structures
- Applying basic techniques of data analysis to real-world data sets

Course outline

1. Introduction to Data

Data basics. Role of quantitative data analysis in decision making process. Common pitfalls in drawing conclusions from data. Introduction to RStudio.

2. Fundamentals of R Syntax

Basic data types in R. Assigning a variable. Creating, naming and selecting elements from vectors. Basic operations with vectors, matrices, data frames, tibbles and lists. R base graphics.

3. Summarizing Data

Measures of location: mean, median, mode. Measures of spread: standard deviation, interquartile range, range. Percentiles. Robust statistics. Data transformation.

4. Getting and Cleaning Data

Importing data to R. Various data sources: text files, web, APIs. Raw and processed data. Working with dates. ZOO and XTS data formats. The principles of tidy data.

5. Probability Distributions

Probability mass functions. Probability density functions. Conditional Probability. Expected values. Data variance. Standard error of the mean. Frequencies and the mode. Empirical distribution function.

6. Linear Regression

Univariate linear regression. Interpreting coefficients. Residuals. Linear regression for prediction.

7. Statistical Inference and Hypothesis Testing

Confidence intervals. T tests. P values. Calculating power of tests. Bootstrapping procedure.

8. Binary Choice Model

Probit and Logit models. Underlying latent variable. Maximum likelihood estimation procedure. Goodness-of-fit measures. Confusion matrix. ROC-curve. Application of binary choice model to estimate ability of monetary policymakers to adjust exchange rates.

9. Support Vector Machines with Applications in Text Mining

Separating hyperplane. Soft margin and hard Margin. Linear vs nonlinear classification. Processing text in R. Assembling a text-mining corpus. Building a word cloud. Estimating the sentiment of the State of the Union addresses.

10. Social Network Analysis

Using Vkonakte API. Getting basic information about users. Building and analyzing a social graph.

Duration: Fall/Winter 2017-18 (Modules 1-3)

Course Materials:

- Instructor's Handouts
- An Introduction to Statistical Learning: with Applications in R http://www-bcf.usc.edu/~gareth/ISL/index.html
- Tan, P. N., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. 1st.
- Quick-R: http://www.statmethods.net
- FRED, Quandl, World Bank Open Data

Course Structure: The course revolves around the most essential structural elements of R, which will be illustrated through case studies.

Forms of Final Assessment: home assignments+group project

Module Grade: 50% - home assignments, 50% - group project

(96-100% - 10, 90-95% - 9, 80-89% - 8, 75-79% - 7, 65-74% - 6, 55-64% - 5, 45-54% - 4, 35-44% - 3, 25-34% - 2, 0-24% - 1)

Instructors: Mikhail Vladimirovich Kamrotov (kamrotov@gmail.com) and Nikolai Igorevich Korzhenevsky (nick@rbk.ru)

Office hours: by appointment

Classroom policies:

- Hand-in assignments policy: All home assignments should be submitted electronically via instructor's email on the due date. No deadline extensions are possible.
- Cheating policy: In case of any kind of plagiarism (with the detected source), the assignment is evaluated as zero without the chance to make up for it. In case of two written assignments with the similarity index of 50% and higher from two students, both get a zero for the assignment.

Data Analysis in R – Handout 1

R download: https://cran.r-project.org

RStudio download: https://www.rstudio.com/products/rstudio/download/#download

Useful links:

1) Calling Bullshit, highly recommended online course at the University of Washington
   http://callingbullshit.org/syllabus.html#Introduction

2) Cumulative vs quarterly sales example
   https://www.statschat.org.nz/2013/09/11/cumulative-totals-tend-to-increase/

3) Super Bowl indicator
   https://en.wikipedia.org/wiki/Super_Bowl_indicator

4) Spurious correlations
   http://www.tylervigen.com/spurious-correlations

5) Simpson's paradox
   http://vudlab.com/simpsons/

6) Reproducibility in economic research
   http://www.financial-math.org/blog/2013/10/reliability-reproducibility-and-the-reinhart-rogoff-error/

7) A brief note on the value of economic theory
   http://www.financial-math.org/blog/2016/07/where-are-the-billionaire-financial-academics/

Homework 1 (due date – October 29, 2017)

Homework 1 may be completed in groups of 4 students max. Only one copy of the homework per group will be accepted and graded. Each participating member of the team receives the same grade.

So far, we've learnt two basic ways of importing data to R:

- downloading data from the web (using Quandl or any other API)
- importing csv files.

Your first assignment is to create your own dataset that you will use for the course project. Since we're a little bit off-schedule, you don't have to finish your project this module.
The main goal of the project is to draw simple conclusions from your data with statistical methods we've used: summary statistics (mean, median, mode, etc.), histogram plots, frequency polygons, kernel density estimations, empirical cumulative distributions.
Keeping this in mind, complete the following steps:

1) choose any problem you can gain insight into using quantitative data;
2) look for appropriate datasets on Quandl or any other source that allows you to access data using R or simply to download data in the text format;
3) elaborate on the issue you would like to address using this data;
4) get your data into R and clean it: remove missing observations, check data type, convert data to xts format if necessary, etc. Explain each step of the data cleaning process.

Your assignment should be submitted as an R script with detailed comments. I should be able to run your script and reproduce your results. Names of group members should be listed at the beginning of the script.

Homework 2

Deadline: January 12, 23:59.

Homework 2 may be completed in groups of 4 students max. Only one copy of the homework per group will be accepted and graded. Each participating member of the team receives the same grade.

Model selection is one of the most important steps in data analysis. This process is generally based on a set of model performance measures. Last part of the SVM script that we've discussed shows how to compute four of them: precision, true positive rate, false positive rate, F-1. It also includes several examples of how to build a ROC-curve. Information on the measures mentioned above can be found in Chapter 5.7 of "Introduction to Data Mining" (**https://www.sendspace.com/file/dxk5by**). In this homework, you will compare several classifiers and evaluate their out-of-sample performance.

1) Read Chapter 5.7 of the book
2) Pick any dataset you are interested in. If you don't have your own dataset, you can use one of two options:
    a. Data on the US and Eurozone interest rates and EUR/USD exchange rate (ir_fx.RData). In this case your classification model will be based on the uncovered interest rate parity idea, you will try to predict the direction of EUR/USD change using both interest rates as two separate variables instead of conventional interest rates differential.
    b. Cross-national voting survey data (singh_2014.RData, singh_2015.RData). For descriptions of these datasets please see two papers attached.
3) Divide your dataset into two parts: train data (approx. 2/3 of all observations) and test data (remaining 1/3). Complete steps 3-7 using **only train data**.
4) Choose variables to be used as features for classification and estimate a binary choice model.
5) Build a ROC-curve for the derived classifier. By default, in binary choice model predictions are computed using 0.5 threshold for probabilities. To build a ROC-curve you should vary this threshold (from 0 to 1) and compute TPR and FPR for each value of the threshold.
6) Build an SVM model for the same dataset using linear and radial kernels.
7) Select optimal values for *cost* and *gamma* using 10-fold cross-validation
8) Compare your binary choice model, SVM with linear kernel and SVM with radial kernel on the basis of accuracy measures and ROC-curves and pick the best classifier.
9) Evaluate performance of your best classifier on the test data.

Your assignment should be submitted as an R script with detailed comments. I should be able to run your script and reproduce your results. Names of group members should be listed at the beginning of the script.