

Requirements for reports

1. 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.
2. The answer to a particular question may take a form of a plot, formula etc followed by a brief explanation and **a conclusion**. All your conclusions **MUST** be justified numerically, i.e., by some computed quantities, plots, etc. The answers do not need to be lengthy but, again, they **MUST** be convincing in mathematical and statistical sense, i.e., in terms of some quantitative measures. Note that I pay much attention to the conclusions, so try to make it as clear as possible.
3. Each student **MUST** use a unique data set. It is your responsibility to make sure that no one else is using the same data. **Check the list on my Google Drive and fill in your name and data you are going to use.** Here is a link to that document <https://docs.google.com/document/d/1W0AOzDVsjZTHzIrwQ7ksuYhPUt-beCGamUmRapBMPao/edit?usp=sharing>.
4. When submitting your report, the **subject of your e-mail** must be **Data Analysis:your name**. Otherwise, your report may get lost or not be processed properly and on time!
5. **The due date** is **October 20, 2019**, 23:59pm.
6. **Late submission**: not possible (will not be accepted)
7. The answer to a question **MUST** contain code in R or some other language which can be placed in the appendix of the report.
8. Failures to comply with the above rules may reduce your grade for the assignment.

Lecture slides

See the LMS

Data sources for the assignment

You have at least three options:

1. Use `getSymbols` command of `quantmod` package to download prices for some stock or commodity (oil, gold, wheat, etc) from Federal Reserve Economic Data repository <http://research.stlouisfed.org/fred2/>, Yahoo Finance or Google Finance. 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:
`GOLD=getSymbols('GOLDAMGBD228NLBM', src='FRED', auto.assign=FALSE)`
`GOLD = na.omit(GOLD)`
 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`, `ISwR` or some others. See the summary on p. 24 of [1] for listing and handling the available datasets. You may also look through the Problems following the respective chapters of [1] to make a choice.
 3. Multivariate data:
 - (a) Use the following link <http://www.stat.ufl.edu/~winner/datasets.html> In particular, consider contingency tables in section “Categorical data”.
 - (b) Some built-in datasets, see 2.
 - (c) Go to the JSE archive http://ww2.amstat.org/publications/jse/jse_data_archive.htm.
 - (d) If you are still unsatisfied with the data from the previous sources 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.
 - (e) 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.

1 Assignment on Confidence Intervals and Hypothesis Testing

1. Get a univariate dataset from sources 1 or 2 and briefly describe it. Using these data,
 - (a) Obtain a 97% confidence interval for the population mean.
 - (b) Perform a t-test on whether the population mean is equal to the sample median. Clearly state the null and alternative hypotheses, provide the p-value.
 - (c) Obtain a 95% confidence interval for the population standard deviation.
 - (d) Find some dataset with a categorical variable. For that variable, compute the proportion of some level. Obtain a 99% confidence interval for that proportion.
 - (e) Perform a hypothesis test on whether the population proportion is equal to 1/2. Clearly state the null and alternative hypotheses, provide the p-value.
 - (f) Come up with some data for calculating the confidence intervals between proportions of two populations (in fact, you need just four numbers). Obtain a 99% confidence interval for the difference between proportions.
 - (g) Perform an appropriate hypothesis test for the difference between proportions (perhaps, using imaginary data). Draw a conclusion.

See [1], Chapters 2-8.

2. Pick the time series of **log returns** (make sure you understand what this means) on three securities or commodities from data sources 1. Use `getSymbols` command to download the data and:
 - (a) Perform the Jarque-Bera for normality. State clearly the null and alternative hypothesis.
 - (b) Check whether the (univariate) empirical distribution of log returns for each stock is normal by examining the QQ-plot. Use the command `qqPlot()` from `car` package instead of the built-in function. Discuss whether the observations are within the confidence interval.

See examples in [2], Section 1.6.

3. Use a built-in set from 2 to perform the χ^2 -test for homogeneity (uniform distribution). Describe the data and discuss the result. See lecture slides and Section 9.1.2 of [1].
4. Get a two-way contingency table from sources 3. Conduct a χ^2 -test for association (independence) between the variables. See lecture slides and Section 9.2 of [1]

2 Assignment on Regression and Classification

1. Simple regression. Get a univariate dataset from sources 2.
 - (a) Build a simple regression model (command `lm`). Provide the estimates of the model's parameters. Draw the scatter plot and the regression line.
 - (b) Analyze the summary statistics (command `summary()`) focusing on:
 - i. The t-test for the slope. Explain.
 - ii. The F-test. Explain.
 - iii. R^2 coefficient. Explain.
 - (c) Plot the residuals against fitted values and comment on the model's adequacy. Examine the qq-plot for the residuals.
 - (d) Make predictions for several new values of the independent variable. For each predicted value, compute and plot the confidence intervals for the mean and single value.
2. Multivariate regression. Get a multivariate dataset (at least 3 variables) from 3.
 - (a) Choose the response and explanatory variables.
 - (b) Build a multivariate linear model (command `lm`). Provide the estimates of the model's parameters.
 - (c) Analyze the summary statistics (command `summary()`) with the emphasis on:
 - i. t-test for slopes. Explain.
 - ii. Overall F-test. Explain.
 - iii. R^2 and adjusted R^2 coefficients. Explain.

- (d) Plot the residuals against fitted values and comment on the model's adequacy.
 - (e) Play with your model by adding or removing the explanatory variables. Alternatively, add a non-linear term(s) to your model:
 - i. Choose the best one by the AIC criterion (command `stepAIC`), see p. 295 of [1].
 - ii. For each model, watch the value of the adjusted R^2 . Explain.
3. Logistic regression. Get a binary response regression dataset from 2 or 3. Briefly describe the data.
- (a) Build a logistic regression model (command `glm`). Comment on the significance of the coefficients.
 - (b) Use `stepAIC` command to select the best model.
 - (c) Make a prediction based on the entire dataset. State the threshold of acceptance. Compare the forecast with the actual observations. Comment on the results.
 - (d) Divide the entire set into training and test subsets. Rebuild the model using only the training subset. Make predictions for the test subset. Comment.
4. Discriminant analysis. Use the same dataset as for the logistic regression.
- (a) Conduct the linear discriminant analysis (command `lda`, package `MASS`) using training and test subsets. Compare the forecast with the actual observations. Comment on the results.
 - (b) Conduct the quadratic discriminant analysis (command `qda`). Comment.
5. The KNN classifier. Use the same dataset as for the logistic regression and discriminant analysis.
- (a) Conduct the KNN classification (command `knn()`, package `class`) using training and test subsets. Compare the forecast with the actual observations. Comment on the results.
 - (b) Play with the number of nearest neighbors K .
6. Compare the quality of classification obtained by algorithms 3-5 for the test subset.

For tasks 3-5, see [3], Chapter 4.

3 Assignment on Principal Component Analysis

1. Get the multivariate data. You have several options:
 - (a) Built-in datasets (packages `UsingR`, `car`, `ISwR`, etc).
 - (b) Go to the JSE archive http://ww2.amstat.org/publications/jse/jse_data_archive.htm.
 - (c) 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.

- (d) 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** 4 quantitative (continuous, not binary or integer!) variables (the more, the better).
- In addition, there may be some categorical variables.
- The number of individuals should not be too small ($> 6 - 7$).
- You may consider time periods (e.g., years) as individuals, i.e., time series will work.

Clearly specify the data you have chosen in your report.

2. Use **FactoMineR** package to study individuals:
 - (a) Plot the individuals in the plane corresponding to the first two principal components (PCs), see [4], p.31. Comment on the resulting cloud.
 - (b) Justify the choice of the PCs by plotting the eigenvalues, [4], p.32. Calculate how much of the total variability is explained by the first two PCs.
 - (c) Discuss the quality of the PCA representation: provide \cos^2 and the contributions for each individual, [4], p.34.
 - (d) If there are categorical variables, paint the individuals with different colors according to the categories. Draw the confidence ellipses and interpret them, [4], p. 36.
3. Study cloud of variables, [4], pp. 36-44.
 - (a) Using the graphical output of **pca** command, discuss correlation between the variables including presence of groups of variables that are closely related.
 - (b) Discuss the quality of the PCA representation: provide \cos^2 and the contributions for variables.
 - (c) Plot the correlations between variables using **pairs** function. Compare the result with that of 3a.

See pp. 44-58 of [4] for the examples.

4 Assignment on Correspondence Analysis

1. Get the multivariate data. You have at least three options:
 - Consider the datasets at <http://www.stat.ufl.edu/~winner/datasets.html>, section “Categorical data”.
 - Suggest your own two-way cross tabulation (contingent table) from some other source. Some free sources of data are listed here <http://guides.emich.edu/data/free-data>. Some research is required.
 - Generate some meaningful two-way contingency tables on your own. Describe these (imaginary) data in detail.

Remember that:

- There MUST be **at least** 3 categories for each variable.

2. Use **FactoMineR** package to:

- (a) Do the χ^2 test for independence and interpret it, see Section 2.2.2 of [4].
- (b) Perform the CA, get the 2D representation of row and column profiles
 - separately
 - in the same graph

See p.87 of [4]

- (c) Analyze the patterns obtained in item 2b. Focus on the total variability, similarities/dissimilarities and the conclusions that can be made from the simultaneous representation of rows and columns. See examples, [4], pp. 92-125.
- (d) Provide the table and graph of eigenvalues, justify the choice of principal components.
- (e) Discuss the quality of the CA representation based on \cos^2 for rows and columns, [4], p.87.

See pp. 92-125 of [4] for more.

5 Assignment on Multidimensional Scaling

Get the distances between 10-12 Russian cities. For example, you can retrieve this information at <https://www.avtodispetcher.ru/distance/table/c172-rossiya/>

1. Do the classical multidimensional scaling using command `cmdscale` from **MVA** package:
 - (a) Plot a two-dimensional MDS configuration representing the cities. Compare the result with the actual geographical location of the cities across the country.
 - (b) Based on the computed eigenvalues, discuss the quality of representation in the 2D space.
 - (c) Plot the Shepard diagram and discuss it.
 - (d) Check whether the MDS configuration you obtained does restore the original distances in a sufficiently high dimensional space.
2. Get the airline distances between 10-12 cities/towns worldwide or, alternatively, a rank-order dissimilarity matrix such as Tables 4.4 or 4.5 in [2], e.g., <http://prtools.org/disdatasets/>.
 - Do the same analysis as for task 1.
 - Pay special attention to the negative eigenvalues, if any. In that regard, discuss which dimensions you choose to represent the MDS solution and why.

See examples in [2], Section 4.4.2.

6 Assignment on Clustering

Get the built-in data with at least 4 quantitative (continuous) variables.

1. Do the hierarchical clustering (preceded by the PCA) using command HCPC from FactoMineR package:
 - (a) Clearly name the recommended (by HCPC) clusters.
 - (b) Explain the meaning of the barplot in the upper-right corner of the output.
2. Perform the K -means clustering, choosing K according to the results of hierarchical clustering.
 - (a) Plot the results
 - (b) Compare distribution of points over clusters with that of hierarchical approach

See examples in [4], Chapter 4 and [3], Section 10.5

References

- [1] J. Verzani, [Using R for Introductory Statistics, Second Edition](#), Chapman & Hall/CRC The R Series, Taylor & Francis, 2014.
URL <https://books.google.ru/books?id=086uAwAAQBAJ>
- [2] B. Everitt, T. Hothorn, [An introduction to applied multivariate analysis with R](#), Springer, New York, 2011.
URL <http://dx.doi.org/10.1007/978-1-4419-9650-3>
- [3] G. James, D. Witten, T. Hastie, R. Tibshirani, [An Introduction to Statistical Learning: With Applications in R](#), Springer Publishing Company, Incorporated, 2014.
- [4] F. Husson, S. Le, J. Pagès, [Exploratory Multivariate Analysis by Example Using R, Second Edition](#), Chapman & Hall/CRC Computer Science & Data Analysis, CRC Press, 2017.
URL <https://books.google.com/books?id=nLrODgAAQBAJ>