Санкт-Петербургский филиал федерального государственного автономного образовательного учреждения высшего образования "Национальный исследовательский университет “Высшая школа экономики”"

Факультет Санкт-Петербургская школа социальных и гуманитарных наук Национального исследовательского университета «Высшая школа экономики»

Департамент социологии

Рабочая программа дисциплины
Анализ данных в социологии (преподается на английском языке)

для образовательной программы «Социология»
направления подготовки 39.03.01 «Социология»
уровень бакалавр

Разработчик программы:
Широканова А.А., к.социол.н., ashirokanova@hse.ru

Согласована методистом ОСУП

«30» августа 2016 г.

Т.Г. Ефимова _____

Утверждена Академическим советом образовательной программы

«30» августа 2016 г., № протокола_____1_____

Академический руководитель образовательной программы

Д.А. Александров _____

Санкт-Петербург, 2016

Аннотация

<table>
<tr><td>Название дисциплины</td><td colspan="3">Анализ данных в социологии (преподается на английском языке)</td></tr>
<tr><td>Образовательная программа</td><td colspan="3">Социология</td></tr>
<tr><td>Тип дисциплины</td><td colspan="3">Обязательная</td></tr>
<tr><td>Требования к уровню знаний студентов, необходимых для освоения дисциплины (пререквизиты)</td><td colspan="3">Студенты должны иметь базовые знания по теории вероятности, методологии и методам социологического исследования, социологической теории</td></tr>
<tr><td>Объем з.е.</td><td colspan="3">10</td></tr>
<tr><td rowspan="2">Объем в часах</td><td>Аудиторная работа</td><td>Самостоятельная работа</td><td>Всего</td></tr>
<tr><td>152</td><td>228</td><td>380</td></tr>
<tr><td>Краткое описание курса</td><td colspan="3">Дисциплина ориентирована на формирование и развитие умений по формулировке и проверке типичных исследовательских задач, ориентированных на описание и предсказание, при анализе социальных данных в программной среде R. Успешное освоение данной дисциплины предполагает владение основами математической статистики, алгебры и анализа, академического письма. Знание среды R является преимуществом. Дисциплина охватывает круг тем, необходимых для формирования базовых компетенций работы с данными (чтение и интерпретация результатов, создание аналитических отчетов), от вводных тем (типы переменных, проверка статистической и исследовательской гипотез, описательные статистики, меры центральной тенденции, стандартное нормальное распределение) до более сложных (обоснование применения хи-квадрата, t-теста, непараметрических статистик; однофакторный дисперсионный анализ, линейная регрессия, интерактивные эффекты). Далее обучающиеся продолжают углубленно изучать анализ данных. На 3 году обучения студентов курс строится вокруг двух главных тем: факторного анализа и статистического предсказания, включающего линейную регрессию и моделирование структурными уравнениями. Также обсуждаются такие ключевые для статистического анализа вопросы, как создание индексов и определение каузальности на основе полученных результатов. На 4 году обучения студентов курс посвящен многомерным методам категориальных данных и включает как специальные виды предсказательных моделей (бинарная логистическая регрессия), так и методы снижения размерности пространства (анализ соответствий, многомерное шкалирование) и классификации (кластерный анализ).

Целью курса является обучение студентов осознанному использованию возможностей количественных исследований. Данный курс также является отправной точкой для студентов, нацеленных на более углубленное изучение методов статистики или планирующих применение количественных методов в собственных</td></tr>
</table>

	исследованиях.
Образовательные результаты по дисциплине	В результате успешного освоения данной дисциплины студенты могут использовать в профессиональной деятельности методы многомерного анализа данных, используемые для установления связи между переменными, структуры связи переменных, предсказания, сокращения размерности пространства и классификации в программной среде R.
Краткое содержание дисциплины	1. Описательные статистики. 2. Сравнение средних. 3. Введение в обобщенные линейные модели. 4. Введение. Статистический вывод и основы регрессионного анализа 5. Линейная регрессия 6. Эксплораторный факторный анализ 7. Конфирматорный факторный анализ 8. Моделирование структурными уравнениями. Путевой анализ, модели структурных уравнений 9. Логистическая регрессия 10. Многомерное шкалирование и анализ соответствий 11. Кластерный анализ
Образовательные технологии	1. Проблемное обучение: разбор случаев 2. Метод проектов 3. Работа в малых группах
Формы контроля	Письменный экзамен (решение задач, подготовка проектного портфолио).
Литература	Основная 1. Joseph F. Hair et al. (2014). Multivariate Data Analysis, Pearson New International Edition, Pearson Education Limited. 2. R. B. Kline (2015). Principles and practice of structural equation modeling, Guilford publications. Дополнительная 1. Alan Agresti (2013). Categorical Data Analysis, 2nd edition, John Wiley & Sons, Inc. 2. Jorg Blasius and Michael Greenacre (2006). Correspondence Analysis and Related Methods in Practices, in: Greenacre, M., & Blasius, J. (Eds.). (2006). Multiple correspondence analysis and related methods. CRC press. P. 3-40. 3. T.A. Brown (2015). Confirmatory factor analysis for applied research, Guilford Publications. 4. Kurt Taylor Gaubatz (2014). A Survivor's Guide to R: An Introduction for the Uninitiated and the Unnerved, 1st edition, Sage. 5. Todd D. Little (ed.) (2013). The Oxford Handbook of Quantitative Methods. Volume 2: Statistical Analysis, Oxford University Press.
Преподаватель	Широканова А.А., доц. департамента социологии, ashirokanova@hse.ru Волченко О.В., преп. департамента социологии, ovolchenko@hse.ru

Course Syllabus

Title of the course	**Data analysis in Sociology (offered in English) (for 2nd year students)**		
Title of the Academic Programme	Sociology		
Type of the course	Core		
Prerequisites	Algebra and Analysis; Applied Software; Methodology and Methods of Sociological Research; Argumentation Theory and Academic Writing; Theory of Probabilities and Mathematical Statistics.		
ECTS workload	4		
Total indicative study hours	Directed Study	Self-directed study	Total
	60	92	152
Course Overview	Discipline is focused on the formation and development of skills on the formulation and verification of typical research tasks, focused on description and prediction, when analyzing social data in the R software environment. Successful development of this discipline implies mastering the basics of mathematical statistics, algebra and analysis, academic writing. Knowledge of the environment R is an advantage. Discipline covers a range of necessary topics for the formation of basic competences for working with data (reading and interpreting results, creating analytical reports), from basic topics (types of variables, testing statistical and research hypotheses, descriptive statistics, measures of central tendency, standard normal distribution) to more complex (justification of the use of chi-square, t-test, non-parametric statistics; univariate analysis of variance, linear regression, interactive effects).		
Intended Learning Outcomes (ILO)	As a result of mastering the discipline, students will be able to formulate and solve problems on the independence of symptoms, to compare two or more averages, to build a regression with direct and interactive effects and present the results, as well as to master the relevant vocabulary in English.		
Indicative Course Content	Section I. Descriptive statistics. Section II. Comparison of averages. Section III. Introduction to generalized linear models.		
Teaching and Learning Methods	Regular Q&A sessions are needed at the beginning and closure of the sessions, at lectures and lab sessions alike. Engaging students to do their own group projects helps them apply theory to real-life data. Peer-review of their final projects helps students compare their own strong points with those of others, which is especially important given the variety of ways of executing routine processes in R. Master-classes with senior students are possible and welcome		

Content and Structure of the Course

No.	Topic	Academic hours, Total	Direted Study			Self-Directed Study
			Lectures	Seminars	Computer lab	

					sessions	
	Part I. Descriptive Statistics					
1	**Research hypotheses vs. statistical hypotheses**	**18**	**2**	**2**		**14**
1.1	The cycle of research. Posing and testing hypotheses (problem-solving)	9	2			7
1.2	Variable types and their descriptive stats	9		2		7
2	**Central tendency measures. Means as a model**	**20**	**2**	**2**	**2**	**14**
2.1	Mean, median, mode. Standard normal distribution and its use. Z-scores	6	2			4
2.2	APA tables. Moments of distributions. Interpretation of z-scores. Mean as a data model	6		2		4
2.3	Getting to know R. Creating objects, types of objects, basic functions (problem-solving)	8			2	6
	Part II. Means Comparison					
3	**Chi-square**	**18**	**2**	**2**		**12**
3.1	Contingency tables. Independence tests	8	2			6
3.2	Reading and interpreting chi-square	8		2		6
4	**Two means comparison**	**20**	**2**	**2**	**2**	**14**
4.1	t-tests and nonparametric tests for independent and dependent samples	6	2			4
4.2	Reading and interpreting the means	6		2		4
4.3	Means comparison in R	8			2	6
	Part III. Introduction to general linear model					
5	**One-way ANOVA**	**24**	**4**	**4**	**4**	**12**
5.1	Revision test with follow-up discussion	4		2		2
5.2	Assumptions and use of ANOVA	4	2			2
5.3	Reading and interpreting ANOVA	4		2		2
5.4	One-way ANOVA in R	4			2	2
5.5	Multiple comparisons. Post hoc tests	4	2			2
5.6	Presenting the results of ANOVA	4			2	2
6	**Correlation and Linear regression**	**26**	**4**	**6**	**4**	**14**
6.1	Correlations	4	2			2
6.2	Building a linear regression	4		2		2
6.3	Linear regression in R (problem-solving)	4			2	2
6.4	Presenting and interpreting a linear regression	4		2		2
6.5	Assumptions behind linear regression	4	2			2
6.6	Reading and interpreting regressions	4		2		2
6.7	Plotting linear regressions in R (case studies)	4			2	2
7	**Interactions in a linear regression**	**26**	**4**	**8**	**2**	**12**
7.1	Understanding the interaction effects: categorical by categorical, categorical by continuous, continuous by continuous variables	8	4			4
7.2	Reading and interpreting interaction models	6		4		2
7.3	Testing for interactions in R	4			2	2
7.4	Group project presentations	8		4		4

Total	**152**	**20**	**26**	**14**	**92**

Indicative Assessment Methods and Strategy	Written exam (test). The evaluation for the 2nd year students is based on three criteria: class assignments, group projects, and final exam. The cumulative grade is rounded according to the rules of algebra. The final grade is rounded in favour of the student.
Readings / Indicative Learning Resources	Mandatory 1. Alan Agresti (2013). Categorical Data Analysis, 2nd edition, John Wiley & Sons, Inc. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1168529 2. Beh, Eric J., and Rosaria Lombardo. Correspondence Analysis: Theory, Practice and New Strategies, John Wiley & Sons, Incorporated, 2014. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1780725. Optional 1. Chambers, J. (2008). Software for data analysis: programming with R. Springer Science & Business Media. Available from HSE library: http://link.springer.com/book/10.1007/978-0-387-75936-4

Indicative Self- Study Strategies	**Type**	**+/–**	**Hours**
	Reading for seminars / tutorials (lecture materials, mandatory and optional resources)	+	36
	Assignments for seminars / tutorials / labs	+	32
	E-learning / distance learning (MOOC / LMS)	+	4
	Fieldwork	-	
	Project work	+	4
	Other (please specify)	-	
	Preparation for the exam	+	16

Academic Support for the Course	Academic support for the course is provided via LMS, where students can find: guidelines and recommendations for doing the course; guidelines and recommendations for self-study; samples of assessment materials
Facilities, Equipment and Software	Lectures are supported by slide presentations demonstrated with a projector. Seminars and lab sessions are to be held in a fully-equipped computer class with personal computers available to every student in a group (in cases when there are more students than PCs they are welcome to bring their own computers). The necessary software is R (https://www.r-project.org/) and RStudio (https://www.rstudio.com) that are available free of charge.
Course Instructor	Dr. Anna Shirokanova, ashirokanova@hse.ru Olesya Volchenko, ovolchenko@hse.ru

Annex 1

Course Content

Part I. Descriptive Statistics

Topic 1. Research hypotheses vs. statistical hypotheses

The cycle of research. Data analysis as part of the research process. Posing and testing hypotheses. Research hypotheses vs. statistical hypotheses testing. Directed and non-directed hypotheses. Dependent and independent variables. Variable scales: nominal, ordinal, continuous (interval and ratio). Descriptive statistics of a variable depending on its type.

Topic 2. Central tendency measures. Means as a model

Mean, median, mode. Standard normal distribution and its use. Z-scores.

APA tables. Moments of distributions. Distribution plots and reading them. Sources of bias in data. Interpretation of z-scores. Mean as a data model.

Getting to know R. Creating objects, types of objects, basic functions. Descriptive statistics in R.

Part II. Means Comparison

Topic 3. Chi-square

Observed and expected frequencies. Measures of association for categorical variables.

Reading and interpreting chi-square tests. Assumptions of chi-square. Independence. Standardised residuals. Odds ratio.

Chi-square and other association measures in R.

Topic 4. Two means comparison

Independent and paired samples. Assumptions behind the t-test. One-sample t-test. Two-sample t-tests. Nonparametric tests for two samples and for multiple samples.

Reading and interpreting means comparison. Confidence intervals.

Means comparison in R

Part III. Introduction to general linear model

Topic 5. One-way ANOVA

Assumptions and use of ANOVA. Between-group and within-group variance, their ratio. Planned and non-planned comparisons; corrections. Post hoc comparisons for equal and unequal

variances. Reading and interpreting ANOVA. One-way ANOVA in R. Presenting the results of ANOVA

Topic 6. Correlation and Linear regression

Correlations. Research problems for correlational analysis. Correlation coefficients for different types of data. ANOVA, correlation, regression as linear models. Building a linear regression.

Ordinary least squares. Fitting the regression line. Assumptions behind linear regression.

Reading and interpreting regressions. Presenting and interpreting a linear regression.

Categorical predictors in a linear regression. Dummy-coding.

Linear regression in R.

Plotting linear regressions in R (case studies).

Topic 7. Interactions in a linear regression

Understanding the interaction effects: categorical by categorical, categorical by continuous, continuous by continuous variables. Effect coding. Centring. Multicollinearity.

Reading and interpreting interaction models in a linear regression.

Testing for interactions in R.

Reporting and interpreting a linear regression with interactions.

Annex 2

Assessment Methods and Criteria

Assessment Methods

Types of Assessment	**Forms of Assessment**	**Modules**			
		1	**2**	**3**	**4**
Interim	Group projects			*	*
	Class assignment				*
Final exam	Exam				*

Assessment Criteria

Group projects. Students form groups of three at the first class and work together on the hypotheses and code during the whole period of the course. Each group selects one country from the 7th round of the European Social Survey, then picks the topic of interest within the scope of available survey questions (e.g. Emotions, Democracy, Family, Morals, Religion, etc.) and performs all the tests covered in class on this data, if possible. One day before each computer lab, the due piece of code with its interpretation is to be submitted and blindly peer-reviewed by two other groups in the shared folder. The instructors would assign reviewers and grade for them, while students might not know who would be their reviewers next time.

Project presentation is the final group project presentation that involves two steps. At the first stage, the group submits the whole of their code with interpretations to the shared folder. After it, they present the findings and procedures in class for the rest of the group in the last two seminar sessions. Students are expected to choose and perform correctly the ways to analyse and interpret the data, as well as to demonstrate their knowledge and skills in presenting these results to the audience. The contribution of each student will be graded. Projects themselves could be submitted as R-codes or R Markdown objects, while in-class presentations should adapted for the .ppt-like presentations (e.g. Prezi, LibreOffice Impress, etc.). The first stage is 60% of this grade, the presentation is 40%.

Class assignment *Revision test is* a standard close-ended questions test that runs for 80 minutes in class and covers necessary topics, from probabilities to means comparisons, checking students' knowledge and understanding of basic topics in data analysis. Some forms of training tests will be available.

An *additional point* to the cumulative grade can be gained by *participation in the seminars* (presenting the material for the others) or by preparing a computer lab script for the whole group after one of the labs. The number of additional points is limited to the number of seminars and labs available.

The **final exam** presents five problems to solve on the spot, including means comparison, linear regression, ANOVA, and chi-square. The maximum final grade is 10 points. To pass the course, the grade should be above or equal to 4.

Your cumulative grade for the course is calculated as a weighted sum of grades for each type of monitoring control tasks in the following way:

$$G_{cumulative} = 0.2* G_{group\ project1} + 0.4* G_{group\ project2} + 0.4\cdot* G_{class\ assignment,}$$ whereby

$G_{group\ project1}$ is the mean grade for your group research project tasks in module 3; $G_{group\ project2}$ is the mean grade for your group research project tasks in module 4; and $G_{class\ assignment}$ is the grade for the interim test at the beginning of module 4.

The cumulative grade can be raised by up to one point for the students who deliver a good seminar presentation or computer lab script (not obligatory).

The exam grade for this course is calculated in the following way:

$G_{final} = 0.6*G_{cumulative} + 0.4*G_{exam}$, whereby

$G_{cumulative}$ is the cumulative grade for this course; and G_{exam} is the exam grade.

The cumulative grade is rounded according to the rules of algebra. The final grade is rounded in favour of the student.

Assessment Criteria

Project Work

Grades	Assessment Criteria
«Excellent» (8-10)	A well-structured, analytical presentation of project work. Shows strong evidence and broad background knowledge. In a group presentation all members contribute equally and each contribution builds on the previous one clearly; Answers to follow-up questions reveal a good range and depth of knowledge beyond that covered in the presentation and show confidence in discussion.
«Good» (6-7)	Clearly organized analysis, showing evidence of a good overall knowledge of the topic. The presenter of the project work highlights key points and responds to follow up questions appropriately. In group presentations there is evidence that the group has met to discuss the topic and is presenting the results of that discussion, in an order previously agreed.
«Satisfactory» (4-5)	Takes a very basic approach to the topic, using broadly appropriate material but lacking focus. The presentation of project work is largely unstructured, and some points are irrelevant to the topic. Knowledge of the topic is limited and there may be evidence of basic misunderstanding. In a group presentation, most of the work is done by one or two students and the individual contributions do not add up.
«Fail» (0-3)	Fails to demonstrate any appropriate knowledge.

Written Assignments (Class assignment, Exam)

Grades	Assessment Criteria
«Excellent» (8-10)	Has a clear argument, which addresses the topic and responds effectively to all aspects of the task. Fully satisfies all the requirements of the task; rare minor errors occur;
«Good» (6-7)	Responds to most aspects of the topic with a clear, explicit argument. Covers the requirements of the task; may produce occasional errors.
«Satisfactory» (4-5)	Generally addresses the task; the format may be inappropriate in places; display little evidence of (depending on the assignment): independent thought and critical judgement include a partial superficial coverage of the key issues, lack critical analysis, may make frequent errors.
«Fail» (0-3)	Fails to demonstrate any appropriate knowledge.

Samples of assessment material

For each task on your group project, remember to self-check on the following points:

1) Are your variables of correct types suitable for this analysis?
2) What is your statistical hypothesis?
3) What is your research hypothesis? Does it make sense?
4) Were the assumptions of the method you are going to use met? If not, what can you do?
5) What conclusions can you draw based on your results?
6) What should you report in a paper using this method?
7) What would be the most effective way to deliver your results to others around?

To download the data for your country, go to the European Social Survey data archive and locate a data set of your interest. Register and download it. Check your data for normality and other assumptions about variables. Formulate the hypotheses. Create a model using your data set and fit it. Report your findings based on the output. Interpret on your results and report them.

In your final presentation, please make sure all of your team members know what you are going to tell the others during the time allocator for presentation. Try to tell the story behind the data.

To prepare for the interim class assignment, one should revise the topics dealing with variable types, error types, standard normal deviation and z-scores, distribution plots and box plots, their meanings; and central tendency measures. It is good to keep a vocabulary with new words, e.g. confounding variable, outliers, cumulative frequency, kurtosis, etc.

Example of a question from the class assignment: "List three specific features of the standard normal distribution: _____."

Annex 3

Recommendations to the Instructor

Try using as many ways of approaching students as possible. Since the knowledge and readiness to learn in English is non-equivalent among the students, be always prepared to stratify exercises as well as theory for different levels. Find and use the youtube.com tutorials on R. Small groups and project work are always encouraged whenever possible and reasonable.

Recommendations to the Students

The purpose of the course is to equip you with necessary understanding of the methods when doing data analysis of some kind. At this stage you are offered an array of basic methods of data analysis which you are likely to use while staying in the social sciences and beyond. Textbooks are now accompanied by useful websites with the data sets and additional learning materials provided. Make full use of them.

Try using your own words when describing the methods or covering new material. Do not hesitate to pose your questions to the instructor but please refrain from blind copying from the book even when it seems a good idea.

In doing your group projects, allocate the time to work together and talk to the instructor or teaching assistant at least once a week. Your results should be logical and rather easy to comprehend, whether it is you talking about them or another person learning about your project.

When required to do peer review for other teams, try to provide helpful and timely feedback and suggest possible improvements to your group mates' projects. While your comments would not be decisive in grading other project, they will be crucial to fostering discussion and coordination within the groups.

Course Syllabus

Title of the course	**Data Analysis in Sociology (offered in English) (for 3rd and 4th years students)**		
Title of the Academic Programme	Sociology		
Type of the course	Core		
Prerequisites	Algebra and Analysis; Applied Software; Methodology and Methods of Sociological Research; Argumentation Theory and Academic Writing; Theory of Probabilities and Mathematical Statistics.		
ECTS workload	6		
Total indicative study hours	Directed Study	Self-directed study	Total
	92	136	228
Course Overview	This course provides an intermediate-advanced statistical analysis for quantitative research in sociology. In the 3rd year of study, the course covers two main topics - factor analysis and statistical prediction, including linear regression and structural equation modeling. We also discuss key issues in statistical analysis, such as creating indices and identifying causality based on the results of the analysis. The 4th year of study focuses on multivariate analysis of categorical data. It includes special types of prediction models (logistic regression), techniques of dimension reduction (correspondence analysis, multidimensional scaling) and classification (cluster analysis). The course is designed for senior students in sociology. The course covers the building blocks of quantitative data analysis with the goal of training students to be informed consumers of quantitative research. This course is also the starting point for students interested in pursuing advanced methods training or planning to use quantitative methods in their own research. This course is more applied and comprehensive than the basic statistics course that you might have taken earlier.		
Intended Learning Outcomes (ILO)	Upon completion of this course students will be able to apply for professional purposes multivariate data analysis methods used to establish relationships between variables, and test variable structures, to predict, to reduce data dimensions and to classify data in the R software.		
Teaching and Learning Methods	1. Problem-solving in case studies 2. Project portfolio 3. Work in small groups		

Content and Structure of the Course

№	**Topic / Course Chapter**	**Total**	**Directed Study**		**Self-directed Study**
			Lectures	**Tutorials**	
3rd year of study					
1	Introduction	4	0	2	2
2	Linear regression: OLS. Diagnostics	14	0	4	10

3	Linear regression: Interaction effects	14	0	4	10
4	Exploratory factor analysis	18	4	4	10
5	Confirmatory factor analysis	20	4	6	10
6	Introduction in SEM	18	4	4	10
7	SEM: model specification	16	2	4	10
8	Path analysis	16	2	4	10
9	SEM with latent variables	16	2	4	10
10	Putting it all together	16	2	4	10
4th year of study					
1(11)	Overview of categorical data analysis	8	2	2	4
2(12)	Binary logistic regression	18	4	4	10
3(13)	Multidimensional scaling	16	2	4	10
4(14)	Correspondence analysis	16	2	4	10
5(15)	Cluster analysis	18	4	4	10
Total study hours		228	34	58	136

Indicative Assessment Methods and Strategy	Written exam (problem solving, project portfolio preparation). The evaluation for the 3rd year students is based on four criteria: activity, home works, mid-term exam, and final exam. The evaluation for the 4th year students relies on homework projects, in-class tests, grade for the first year, and the final exam.
Readings / Indicative Learning Resources	Mandatory 1. Denis, Daniel J. (2015). Applied Univariate, Bivariate and Multivariate Statistics, John Wiley & Sons, Inc. https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4338227 2.Kline, R. B. (2015). Principles and practice of structural equation modeling, Guilford publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4000663 Optional 1. T.A. Brown (2015). Confirmatory factor analysis for applied research, Guilford Publications. http://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1768752 2. Stowell, Sarah (2014). Using R for Statistics. http://proxylibrary.hse.ru:2099/toc.aspx?bookid=66684 3. Todd D. Little (ed.) (2013). The Oxford Handbook of Quantitative Methods. Volume 2: Statistical Analysis, Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199934898.001.0001/oxfordhb-9780199934898. http://proxylibrary.hse.ru:2089/view/10.1093/oxfordhb/9780199934898.001.0001/oxfordhb-9780199934898?rskey=QWL2FA&result=2

Indicative Self- Study Strategies	**Type**	**+/–**	**Hours**
	Reading for seminars / tutorials (lecture	+	40

	materials, mandatory and optional resources)		
	Assignments for seminars / tutorials / labs	+	40
	E-learning / distance learning (MOOC / LMS)	+	6
	Fieldwork	-	
	Project work	+	30
	Other (please specify)	-	
	Preparation for the exam	+	20
Academic Support for the Course	Academic support for the course is provided via LMS, where students can find: guidelines and recommendations for doing the course; guidelines and recommendations for self-study; samples of assessment materials		
Facilities, Equipment and Software	Lectures are supported by slide presentations demonstrated with a projector. Seminars and lab sessions are to be held in a fully-equipped computer class with personal computers available to every student in a group (in cases when there are more students than PCs they are welcome to bring their own computers). The necessary software is R (https://www.r-project.org/) and RStudio (https://www.rstudio.com) that are available free of charge.		
Course Instructor	Dr. Ksenia Tenisheva, ktenisheva@hse.ru Dr. Anna Shirokanova, ashirokanova@hse.ru		

Annex 4

Course Content

Topics 1-2. Introduction. Linear regression: OLS. Diagnostics

Covariance and correlation, basic concept and logics of linear regression, OLS estimator of linear regression, interpretation and statistic test of OLS estimators, fitted values and residuals, R-squared, addressing nonlinearity in linear regression framework, standardized coefficients, drawing plots, practice in R.

Topic 3. Linear regression: Interaction effects

Main and multiplicative effects in regression models. Interaction effects, additive effects. Interpreting results. Choosing best model. Practice in R.

Topic 4. Exploratory factor analysis

Dimensionality reduction. Manifest and latent variables. Factors, graphical representation of factors. Exploratory factor analysis. Factor scores, factor space, types of rotation. Optimal number of factors. Interpretation of the results. Creating indices based on factor analysis. Practice in R.

Topic 5. Confirmatory factor analysis

Difference between exploratory and confirmatory factor analyses. Factor structure. Testing your (or somebody else's) scales. Types of latent variables. Constructing factor model in lavaan package. Calculation of degrees of freedom, minimal number of cases. Non-correlated and correlated latent factors. Interpreting results. Model diagnostics. Cronbach's alpha. Practice in R.

Topic 6. Introduction in SEM

Structural equation modeling as extension of confirmatory factor analysis. Exogenous and endogenous variables. Testing causal assumptions. Partial correlation, heterogeneous correlations (polychoric, tetrachoric and polyserial correlations). Practice in R.

Topic 7. SEM: model specification and identification

Formulating theory-based causal hypotheses. Causal inference. Specification concepts. Mediation and moderation effects. Measurement error: correlated and uncorrelated. Practice in R.

Topic 8. Path analysis

Concept of "path". Path analysis: only observed variables. Graphical representation. Identification of path model. Estimation of structural equation model. Model fit. Degrees of freedom, number of cases. Meaning of the indices. Corrected chi-square measures. Interpreting the results. Practice in R.

Topic 9. SEM with latent variables

Introducing latent factors in the model. Identification of SEM. Estimation of structural equation model. Model fit. Meaning of the fit indices. Model modification. Interpreting the results. Practice in R.

Topic 10. Putting it all together

Implementing all the methods to the real-life research. Combining factor analysis and regression analysis. Using SEM to test theoretical assumptions about causality. Advantages and disadvantages of the methods.

Topic 1(11). Overview of Categorical Data Analysis

Models for categorical outcome variables. Variety of goals of analysis with categorical data. Examples of empirical research for various methods, e.g. Poisson regression (count variable), binary logistic regression, ordinal regression, multinomial regression, correspondence analysis, conjoint analysis, multidimensional scaling, and cluster analysis. Typical goals of analysis and interpretation of results.

Topic 2(12). Binary Logistic Regression

Logistic regression. Objectives of logistic regression. Logistic curve. Assumptions of logistic regression. Transforming a probability into odds and logit values. Maximum likelihood estimation. Goodness-of-fit measures for logistic regression. Interpretation of results (linear and dichotomous predictors). Stepwise model building. Diagnostics. Procedures in R.

Topic 3(13). Multidimensional Scaling

Dimension reduction as an objective of data analysis. Idea of MDS. Perceptual map. MDS vs. factor and cluster analyses. Objectives of MDS. MDS algorithms. Decompositional and compositional approach. The number and selection of objects. Nonmetric vs. metric methods. Similarity data and preference data. Assumptions of MDS analysis. Selecting dimensionality. Ideal point. Goodness-of-fit measures. Interpreting MDS results. Procedures in R.

Topic 4(14). Correspondence Analysis

Objectives of correspondence analysis. Assumptions of correspondence analysis. Perceptual mapping. Principal components analysis. The row and column problems. Correspondence analysis displays. Correspondence analysis biplots. Multiple correspondence analysis. MCA maps. Measures of fit for MCA. Interpreting correspondence analysis results. Canonical correspondence analysis. Procedures in R.

Topic 5(15). Cluster Analysis

Objectives of cluster analysis (taxonomy description, data simplification, and relationship identification). Necessity of conceptual framework. Similarity measures. Proximity matrix. Decision-process in cluster analysis. Dendrograms. Cluster profiles. Distance measures for various types of variables. Assumptions of cluster analysis. Measures of overall fit. Between- and within-cluster variation. Hierarchical and non-hierarchical clustering algorithms. Determining the number of clusters. Interpretation of clusters. Cross-classification from several solutions. Procedures in R.

Annex 5

Assessment Methods and Criteria

Assessment Methods

3rd year of education

Types of Assessment	**Forms of Assessment**	**Modules**			
		1	**2**	**3**	**4**
	Project		*	*	
	In-class Participation		*	*	
Summative Assessment	Exam			*	

Project. There are three basic features assessed: correct calculations and correct code (syntax); correct interpretations – students must describe trends properly, assess significance of the results, and predict values of dependent variable correctly; and produce correct graphics, with proper types of plots and formatting applied. Homework is graded from 0 (“extremely poor”=”fail”) to 10 (“perfect”=”pass”) each. Proficiency in the English language does not affect the grade.

In-class participation during lectures and seminars. Students are expected to ask meaningful questions and participate in discussions, as well as help other students during practices. Regular active participation in the classes is graded as perfect (10); no participation is graded as 0.

Exam is aimed at checking the skills students should have obtained during the course. Its structure is close to the structure of home assignments, though it covers all the topics studied. Criteria for the assessment of the exam are the same as for home works: correct calculations, correct interpretation and correct graphics.

The grades are calculated by the following formula:

Cumulative score = 0.2*activity + 0.8*mean(projects)

Final grade = 0.8*cumulative score + 0.2*exam

4th year of education

Types of Assessment	**Forms of Assessment**	**Modules**			
		1	**2**	**3**	**4**
	Project			*	
	In-class Participation			*	
Summative Assessment	Exam			*	

Project. Two individual projects are due. A project applies one of the methods covered in the course (binary logistic regression or cluster analysis) and presents the results as a report. Two projects sum up to a student's portfolio. Specific project requirement are available in the LMS.

In-class participation. Every second seminar there is an in-class test on interpretation of binary logistic regression, multidimensional scaling, correspondence analysis, and cluster analysis.

Exam consists of two problems involving the methods covered in this course. Criteria for the assessment of the exam are the same as for home projects: correct specification, interpretation or results, and conclusions.

The grades are calculated by the following scheme:

Cumulative score = 0.2* exam grade in sophomore year + 0.2* exam grade in junior year + 0.2*project 1 + 0.2*project 2 + 0.2*activity

Final grade = 0.8*cumulative score + 0.2*final exam

The method of rounding the final grade: arithmetic.

Assessment Criteria

In-class Participation

Grades	Assessment Criteria
«Excellent» (8-10)	A critical analysis which demonstrates original thinking and shows strong evidence of preparatory research and broad background knowledge.
«Good» (6-7)	Shows certain evidence of preparatory research and background knowledge, a reasonable standard of expression.
«Satisfactory» (4-5)	Satisfactory overall, showing a fair knowledge of the topic. Significant hesitation in answering follow-up questions and/or incomplete or partly irrelevant answers.
«Fail» (0-3)	Very limited to insufficient evidence of relevant knowledge and skills in addressing the topic. Unable to offer relevant information or opinion in answer to follow-up questions.

Project Work

Grades	Assessment Criteria
«Excellent» (8-10)	A well-structured, analytical presentation of the project. Student shows strong evidence and broad background knowledge. In a group presentation, all members contribute equally and each contribution builds on the previous one clearly. Answers to follow-up questions reveal a good range and depth of knowledge beyond that covered in the presentation and show confidence in discussion.
«Good» (6-7)	Clearly organized analysis, showing evidence of good overall knowledge of the topic. The presenter of the project work highlights key points and responds to follow-up questions appropriately, making several minor mistakes. In group presentations, there is evidence that the group has met to discuss the topic and is presenting the results of that discussion, in an order previously agreed but lacks some knowledge to address the necessary points.
«Satisfactory» (4-5)	Takes a very basic approach to the topic, using broadly appropriate but suboptimal material, lacks focus. The presentation of project work is largely unstructured, and some points are irrelevant to the topic. Knowledge of the topic is limited and there may be evidence of basic misunderstanding. In a group presentation, most of the work is done by one student and the individual contributions do not add up.
«Fail» (0-3)	Fails to submit the project or to demonstrate enough knowledge to meet the project's requirements.

Written Assignments (Written Exam)

Grades	Assessment Criteria
«Excellent» (8-10)	All problems are solved correctly, all results are properly interpreted. The student makes a clear argument that responds effectively to all aspects of the problem. Few minor errors may occur.
«Good» (6-7)	The answer addresses most aspects of the topic with a clear argument which is not always correct. The response covers part of the task requirements.
«Satisfactory» (4-5)	The student addresses a minor part of the task which solves at least some part of the problems correctly. The answer demonstrates certain skills of correct problem-solving and interpretation of results.
«Fail» (0-3)	The student fails to demonstrate enough appropriate knowledge to be able to

	solve at least part of the problems correctly.

Samples of Assessment Material

To assess their progress, students will be given home projects each two weeks. Students are expected to replicate analysis from a given scientific article using the methods discussed during the classes or to produce their own model following the lecture. Each project is based on the materials from preceding lectures and includes both theoretical questions and practical tasks that ought to be answered using R.

Example:

- *Replicate the regression table given in the article. Interpret the results.*
- *What is the null hypothesis? What is the alternative hypothesis in this test?*
- *Find the test statistic in the R output. Show how the resulting test statistic could be computed.*
- *What is the p-value and what does the p-value mean?*
- *Should the null hypothesis be rejected here? Can one conclude that the estimated relationship between household size and satisfaction would hold true in the population?*

Exam structure is similar to home tasks, though it covers all topics studied. Sample of exam questions (3rd year, students are expected to answer at home):

- *Read the article carefully. Identify the measures used in the analysis, find them in the dataset.*
- *Were the variables transformed in any way? Prepare the data for analysis*
- *Replicate the regression table given in the article.*
- *Interpret the results. Do your results fit those presented in the article? What might be the reason of this difference?*

Sample of exam questions (4th year):

"You have 60 minutes to solve two problems. Hand in your solutions as an R script titled "YourLastName_Option.R". You may use your own or others' scripts. You may not use the help of other people by talking to them, texting them, or otherwise. Should you be noticed doing so, the exam is over and you get zero points for it.

1. Estimate and describe a logit model. File: exam1.txt
You have the data about religiosity (0=not religious, 1=religious), sex (1=male, 2=female), age (years), and marital status of the respondent (married, divorced, widowed, or single).
Build a model predicting whether the respondent is religious or not.
Report your full model and describe the coefficients.

2. Identify and describe clusters. File: exam3.csv
You have the data about trust to political parties, to the European Parliament, and to the United Nations Organizations in 26 countries around Europe. Using cluster analysis, identify which countries are closer to each other in how their citizens trust these political institutions.
Report the final solution, how you reached it, and describe the clusters."

Annex 6

Guidelines and Recommendations

Recommendations to Course Instructors

The goal of this course above and beyond teaching specific methods is to enable students to use the methods covered in the course on a stand-alone basis whenever they need this in the future. Therefore, every reasonable effort should be made to make the material understandable and comprehensible, depending on the level of the student. Encouraging those students who have already understood new material to share their understanding with the others has demonstrated rewarding results. Regular Q&A sessions are needed at the beginning of each session, at lectures and labs alike. The general recommendation is to put emphasis on training the skills to perform the same types of analysis autonomously; therefore, the more time students get to practise their data analysis skills on different data sets, the more reliable the success of the course. Try using as many ways of approaching students as possible. Since the knowledge and readiness to learn in English is non-equal among students, be always prepared to stratify exercises as well as theory for different levels. Find and use the youtube.com tutorials. Small groups are always encouraged whenever possible and reasonable.

Recommendations to Students for Doing the Course

Completing this course is not like any other discipline you have studied. Here, the purpose of the course is to equip you with necessary methods when doing data analysis of some kind. You are offered an array of methods of data analysis which you are likely to use while staying in the social sciences and beyond. Try using your own words when describing the methods or covering new material. Do not hesitate to pose your questions to the instructor but please restrict from blind copying from the book even when it seems a good idea. This course is meant for your better understanding of what goes on when you call a function in R. Additionally, try keeping a vocabulary on each topic covered, supplying the most important terms with examples. This will help you at the exam and in the future as your personal reference book.

Recommendations for Self-Study

Read and watch as much about the topic as possible. Having read on the same topic from several textbooks usually improves your understanding substantially. Watch and learn extra beyond the classes. Complete the excellent introductory methods courses on Coursera or DataCamp to recap on basic statistical concepts if you feel you could benefit from it. Try to use systematically the power of community at stackoverflow. Thick methods textbooks are now accompanied by useful websites with the data sets and additional learning materials provided (www.mvstats.com).

Annex 7

Special conditions for organization of learning process for students with special needs

The following types of comprehension of learning information (including e-learning and distance learning) can be offered to students with disabilities (by their written request) in accordance with

their individual psychophysical characteristics:

1) *for students with visual impairment:* a printed text in enlarged font; an electronic document; audios (transferring of learning materials into the audio); an individual advising with an assistance of a sign language interpreter; individual assignments and advising.
2) *for students with hearing impairment: a* printed text; an electronic document; video materials with subtitles; an individual advising with an assistance of a sign language interpreter; individual assignments and advising.
3) *for students with physical impairment:* a printed text; an electronic document; audios; individual assignments and advising.