**Санкт-Петербургский филиал федерального государственного   
автономного образовательного учреждения высшего профессионального   
образования "Национальный исследовательский университет**

**"Высшая школа экономики"**

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

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

для направления 39.03.01 “Социология”

подготовки бакалавра

3 курс

Автор программы:

Чой Сонгсу, PhD, schoi@hse.ru

Тенишева К.А., ktenisheva@hse.ru

Согласована методистом ОСУП

«\_27\_»\_октября\_2015 г.

Т.Г. Ефимова \_\_\_\_\_\_\_\_\_\_

Утверждена академическим советом ОП «Социология»

«\_27\_»\_октября\_2015 г.

Академический руководитель ОП

Д.А.Александров \_\_\_\_\_\_\_\_\_\_

Санкт-Петербург, 2015

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения кафедры-разработчика программы.*

**Аннотация программы учебной дисциплины**

**Название дисциплины Анализ данных в социологии (преподается на английском языке)\_**

**Направление \_\_\_\_\_\_\_\_39.03.01 «Социология»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Код, название

**Автор программы Чой Сонгсу, PhD, schoi@hse.ru**

**Тенишева К.А., ktenisheva@hse.ru**

1. Краткое описание курса

Курс представляет собой анализ данных в социологии для продолжающих (intermediate-level). Он строится вокруг концепта генерализованных линейных моделей (Generalized Linear Model), включающего линейную регрессию, а также модели для категориальных и счетных (count) переменных, анализ временных рядов и лонгитюдный анализ. Также будут обсуждаться такие вопросы обработки данных, как пропущенные ответы (missing data) и выборки сложного неслучайного дизайна (например, панельные данные). Возможно также включение таких разделов как инструментальные переменные и сопоставление кейсов (matching), являющиеся крайне важными инструментами для статистических выводов в социальных науках.

Курс охватывает базовые темы количественного анализа данных. Его целью 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 исследованиях.

2. План курса

1. Переменные; выборки; статистический вывод и статистические тесты
2. Теория вероятности; распределения; статистические тесты
3. Линейная регрессия.
4. Биномиальная логистическая регрессия (Linear probability model; Logistic regression model)
5. Мультиномиальная логистическая регрессия
6. Порядковая логистическая регрессия
7. Модели для количественных (count) переменных
8. Моделирование временных рядов: discrete-time hazard model
9. Модели с фиксированными и случайными эффектами (Fixed-eﬀects and random-eﬀects models)
10. Веса; пропущенные значения
11. Подготовка и презентация исследовательского проекта.

3. Требования к уровню знаний студентов

Студенты должны иметь базовое знание статистики в социальных науках: вероятность, тестирование гипотез, линейная регрессия.

4. Преподаватель

проф. Сонсу Чой, PhD; Тенишева К.А. преподаватель департамента социологии

5. Тип экзамена

Письменная работа

Англоязычная версия

1. Outline

This course provides an intermediate-level introduction to statistical analysis for quantitative research in sociology. The course will cover advanced topics in linear regression, including a thorough

review of major topics in basic statistics (e.g., probability, estimation and inference, linear regression) and an introductory overview of models for categorical and count data, the analysis of time data, and longitudinal data. We will also discuss data-related issues such as missing data and weighting, and data that are complicated by issues of non-random design (e.g., panel data). All these topics can fall into the concept of `Generalized Linear Model (GLM).' If time allows, we will also discuss instrumental variable and matching, which are critical and widely-used conceptual tools for causal inference in the social science.

The course is designed for third-year undergraduate students in sociology. The course covers the basic 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 interesting in pursuing more advanced methods training or who plan 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.

2. Syllabus

1. Random variable: distribution, types of variables; Sampling and estimation; Inference and statistical testing
2. Probability theory. Distributions. Statistical testing
3. Linear regression model
4. Logistic regression model
5. Multinomial logit model
6. Ordered logit model
7. Models for count outcomes
8. Event history modeling: discrete-time hazard model
9. Fixed-effects and random-effects models
10. Weights and Missing data
11. Presentation of research paper

3. Prerequisite

Students are expected to have taken some sort of basic/introductory statistics

course for social science research or to have at least some experiences and knowledge about basic

social statistics, such as probability, hypothesis testing, linear regression, while they will revisit

those topics in a considerable proportion during the course.

4. Author

Prof. Seongsoo Choi, PhD; Ksenia Tenisheva, teaching assistant

5. Examination type

Written assignment

# 1. Area of Application and Regulatory References

The program intends to lay the basic foundation of knowledge and determine the content and forms of educational activities and assessment.

The program is designed for the instructors of the Data Analysis in Sociology course, teaching assistants, and students of sociology (Bachelor Program 39.03.01 "Sociology").

The program of the Data Analysis course has been developed in accordance with:

* Educational standard of NRU HSE for Bachelor level education. URL: <http://www.hse.ru/data/2014/02/03/1329546127/ОС%20БАК%20ГУ-ВШЭ%20социология%20%202010.pdf>
* Educational Program of NRU HSE for Bachelor level education, field of study 39.03.01 “Sociology”.
* Curriculum of NRU HSE – Saint Petersburg for Bachelor level education, field of study 39.03.01 “Sociology”.

# 2. Course Goals

The goal of the course is to train students to be informed consumers of quantitative research. This course is also the starting point for students interesting in pursuing more advanced methods training or who plan to use quantitative methods in their own research. This course is more applied and comprehensive than the basic statistics course that students should have taken earlier.

# 3. Students' Competencies to be developed by the Course

As a result of studying the discipline a student is supposed to be able to:

* Conduct simple statistical analysis in RStudio environment;
* Deal with basic types of sampling bias;
* Choose appropriate methods for certain types of variables and certain aims of the analysis;
* Give meaningful interpretation of statistical results: regression coefficients, tables, plots and diagrams (produced in R);
* Perform data transformations for normality correction and variables’ standardization;
* Graphically represent results of the analysis;
* Create analytical reports describing all the stages of the analysis and interpreting its results

The Course develops the following competencies:

|  |  |  |  |
| --- | --- | --- | --- |
| Competencies | NRU-HSE Code | Descriptors –  the learning outcomes  (the indicators of achievement) | Forms and methods  of studies that contribute to  the development of  a competence |
| **The graduate should have such general cultural competences (GC),** as: | | | |
| The preparedness to cooperate with colleagues and to work in a team. | GC-3 | Students should be able to communicate in English with other students, he/she is expected to demonstrate cooperation and contribute to group work.  Professional competence in oral and written English.  Students will be required to use English literary sources to develop their language skills and communicate their arguments in English. | Lectures, seminar and practical sessions in computer classes, homework assignments. Elaboration and public defense of individual analytical projects. Participation in group presentations at seminars. Each student will be expected to take turns in representing their teams in group projects. |
| The ability of applying fundamental ideas and methods of liberal arts, social and economic sciences in order to solve professional tasks. | GC-9 | To use English literary texts in the preparation of lectures, seminars, practical classes, and home assignments.  Develop written and oral communication skills in English.  Students are taught to use the information that he/she got from English-language literary sources to express themselves.  Students are trained to work with different sources of sociological and statistical information: web-pages, data sets, secondary sociological data, academic and study literature sources.  Students should be able to solve research problems applying the knowledge of modern theory in social sciences and comparative methodology.  Students should be aware of existing methods of statistical analysis applied in sociology and use them in his or her professional research. | Lectures, seminar and practical sessions in computer classes, reading, homework assignments, designing own analytical project, performing statistical analyses, and writing analytical reports of own project research. |
| The ability to analyze socially important issues and processes. | GC-10 | Students should be able to prepare for lectures, seminars, practical classes and home assignments in English.  Develop professional competency in oral and written communication skills in English.  Students should be able to work with different sources of sociological and statistical information: web-pages, data sets, secondary sociological data, academic and study literature sources.  Students should be able to use the information that he/she got from English literary sources to express and develop their arguments.  Students should have the ability to solve research problems by applying recent theoretical approaches in the social sciences and comparative methodology.  Students should be aware of existing methods of statistical analysis applied in sociology and able to use them in his or her professional research. | Lectures, seminar and practical sessions in computer classes, reading, homework assignments, writing analytical reports on own project research. |
| The ability to use the basic laws of the natural sciences in professional research, apply methods of mathematical analysis and modeling, theoretical and experimental research. | GC-11 | To use English texts in the preparation of lectures, seminars, practical classes, and home assignments.  Develop written and oral communication skills in English.  Students are taught to use the information that he/she got from English-language literary sources to express themselves.  Students are trained to work with different sources of sociological and statistical information: web-pages, data sets, secondary sociological data, academic and study literature sources.  Students should be able to solve research problems applying the knowledge of modern theory in social sciences and comparative methodology.  Students should be aware of existing methods of statistical analysis applied in sociology and use them in his or her professional research. | Lectures, seminar and practical sessions in computer classes, reading, homework assignments, performing statistical analyses and writing analytical reports of own project research. |
| The ability to work with information in the global computer networks. | GC-14 | Students should be able to use web resources.  To use English texts in the preparation of lectures, seminars, practical classes, and home assignments.  Develop written and oral communication skills in English.  Students should be able to work with different sources of sociological and statistical information: web resources, data sets, secondary sociological data, and other traditional databases and sources.  Students should be able to identify the gaps in existing research and lack of sources and explore new avenues of information to recuperate the same.  Students should be able to explore open and archived sources for sociological surveys.  Students should be able to perform statistical analyses both using on-line built-in facilities of sociological data archives web-sites and PC-installed SPSS. | Lectures, seminar and practical sessions in computer classes, reading, homework assignments, designing an own analytical project, performing statistical analysis for analytical projects, writing analytical report on own project research. |
| **The graduate should have such general professional competences (PC),** as: | *in research activity:* | | |
| The ability to independently formulate objectives, set specific tasks of research in various fields of sociology and solve them with the help of modern research methods, applying the latest Russian and international experience and modern facilities, gadgets, information technologies. | PC-2 | To use English texts in the preparation of lectures, seminars, practical classes, and home assignments.  Develop written and oral communication skills in English.  Students should be able to work with different sources of sociological and statistical information: web-pages, data sets, secondary sociological data, academic and literary sources.  Students are taught to use the information that he/she got from English-language literary sources to express themselves and defend their arguments.  Students should have the ability to solve research issues by applying a comparative methodology and modern theoretical approaches in the social sciences.  Students should be aware of existing methods of statistical analysis applied in sociology and use them in his or her professional research. | Lectures, seminar and practical sessions in computer classes, reading, homework assignments, elaboration and fulfillment of an own research project. |
| The ability and readiness to participate in the processing of the scientific and technical documentation, research reports, present the results of research taking into account the demands of a potential audience. | PC-3 | To use English texts in the preparation of lectures, seminars, practical classes, and home assignments.  Develop written and oral communication skills in English.  Students should be able to work with different sources of sociological and statistical information: web-pages, data sets, secondary sociological data, academic and literary sources.  Students are taught to use the information that he/she got from English-language literary sources to express themselves and defend their arguments.  Students should be aware of existing methods of statistical analysis applied in sociology and use them in his or her professional research.  Students should be able to apply methods of statistical analysis to social and economic problems using SPSS.  Students should be able to interpret the results of the statistical analysis and make conclusions.  Students should be able to make tables and graphs presenting the results of statistical analysis.  Students should be able to write academic and commercial analytical reports. | Lectures, seminar and practical sessions in computer classes, reading, homework assignments, writing an analytical report of own project research. |
|  | *in project activity:* | | |
| The ability to prepare and present scientific research and analytical projects in accordance with the guidelines of normative documents. | PC-7 | To use English texts in the preparation of lectures, seminars, practical classes, and home assignments.  Develop written and oral communication skills in English.  Students should be able to work with different sources of sociological and statistical information: web-pages, data sets, secondary sociological data, academic and literary sources.  Students are taught to use the information that he/she got from English-language literary sources to express themselves and defend their arguments.  Students should be able to solve research problems applying the knowledge of modern theory in social sciences and comparative methodology.  Students should be aware of existing methods of statistical analysis applied in sociology and use them in his or her professional research.  Students should be able to write both academic and commercial analytical reports.  Students should be able to meet the requirements for information security in his research work. | Lectures, seminar and practical sessions in computer classes, reading, homework assignments, performing statistical analysis and writing analytical report of own project research. |
| The ability to process and analyze data to produce analytical solutions, expertise, and solutions. | PC-8 | To use English texts in the preparation of lectures, seminars, practical classes, and home assignments.  Develop written and oral communication skills in English.  Students should be able to work with different sources of sociological and statistical information: web-pages, data sets, secondary sociological data, academic and literary sources.  Students are taught to use the information that he/she got from English-language literary sources to express themselves and defend their arguments.  Students should be aware of existing methods of statistical analysis applied in sociology and use them in his or her professional research.  Students should be able to apply methods of statistical analysis to social and economic problems using SPSS.  Students should be able to write both academic and commercial analytical reports.  Students should be able to meet the requirements for information security in his research work. | Lectures, seminar and practical sessions in computer classes, reading, homework assignments, performing statistical analysis and writing analytical report of project research. |
|  | *in the organizational and administrative activity:* | | |
| The ability to use basic theoretical knowledge as well as practical skills in research, analytical and consulting activities. | PC-10 | To use English texts in the preparation of lectures, seminars, practical classes, and home assignments.  Develop written and oral communication skills in English.  Students should be able to work with different sources of sociological and statistical information: web-pages, data sets, secondary sociological data, academic and literary sources.  Students are taught to use the information that he/she got from English-language literary sources to express themselves and defend their arguments.  Students should be able to set research goals and implement the research design to study various social problems through a comparative perspective.  Students should be aware of existing methods of statistical analysis applied in sociology and use them in his or her professional research.  Students should be able to apply methods of statistical analysis to social and economic problems using SPSS. Students should be able to write both academic and commercial analytical reports.  Students should be able to meet the requirements for information security in his research work. | Lectures, seminar and practical sessions in computer classes, reading, homework assignments, performing statistical analysis and writing analytical report of own project research. |
| The ability of using methods of gathering, processing and interpretation of the complex social data in order to solve managerial and organizational issues including those outside a direct sphere of activity. | PC-11 | To use English literary texts in the preparation of lectures, seminars, practical classes, and home assignments.  Develop written and oral communication skills in English.  Students should be able to work with different sources of sociological and statistical information: web-pages, data sets, secondary sociological data, academic and literary sources.  Students are taught to use the information that he/she got from English-language literary sources to express themselves and defend their arguments.  Students should be aware of existing methods of statistical analysis applied in sociology and use them in his or her professional research.  Students should be able to apply methods of statistical analysis to social and economic problems using SPSS.  Students should be able to write both academic and commercial analytical reports.  Students should be able to meet the requirements for information security in his research work. | Lectures, seminar and practical sessions in computer classes, reading, homework assignments, performing statistical analysis and writing analytical report on own project research. |

# 4. How the Course Fits in with the Curriculum

The present course relates to professional disciplines, the block of basic disciplines. It is an essential element of study for a comprehensive and in depth understanding of the majority of professional disciplines and is crucial for developing research competency.

The “Data Analysis” is an obligatory course within the Basic Syllabus for the area of studies 39.03.01 "Sociology" for the B.A. program at the Higher School of Economics.

The Course is based on the acquisition of the following Courses (at the bachelor level):

* Algebra and analysis
* Probability Theory and Mathematical Statistics
* Applied Software
* Sociological Theory
* Methodology and Methods of Sociological Research

The Course requires the following students' competencies and knowledge:

* Major social theories
* Basic statistical concepts
* Methods of quantitative social research

The main provisions of the Course should be used for further studies of the following Courses:

* Scientific Research Seminar

# 5. Course Schedule

This course is worth 7 credits.

A usual weekly procedure of this course will be a combination of a three-hour lecture given by the instructor and a one-hour group lab practice, during Modules 2 and 3. In Module 4, students will be required to work on group projects, in which students conduct their own quantitative research, using any methods they have learned during this course or courses taken earlier.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Topic | Total Academic Hours | Classroom Activities | | | Self-Study |
| Lectures | Seminars | Practical Classes |
| 1 | An overview of the course | 12 | 2 | 2 | - | 8 |
| 2 | Random variable: distribution, types of variables | 12 | 2 | 2 | - | 8 |
| 3 | Probability theory | 12 | 2 | 2 | - | 8 |
| 4 | Linear regression model | 14 | 2 | 4 | - | 8 |
| 5 | Logistic regression model | 16 | 4 | 4 | - | 8 |
| 6 | Multinomial logit model | 22 | 4 | 4 | 2 | 12 |
| 7 | Ordered logit model | 22 | 4 | 2 | 4 | 12 |
| 8 | Models for count outcomes | 20 | 2 | 2 | 4 | 12 |
| 9 | Event history modeling: discrete-time hazard model | 22 | 4 | 4 | 2 | 12 |
| 10 | Fixed-eﬀects and random-eﬀects models | 16 | 2 | 2 | 2 | 10 |
| 11 | Weights and Missing data  A quick review of the course | 16 | 2 | 2 | 2 | 10 |
| 12 | Writing and presenting own research project | 32 | - | 4 | 6 | 22 |
|  | Total: | 216 | 30 | 34 | 22 | 130 |

# 6. Forms and Types of Testing

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type of control | Form of control | Modules | | | | Requirements |
| 1 | 2 | 3 | 4 |
| Current | Homeworks |  | x | x | x | 1. Problem sets will be given once in 2 weeks. Students may work together to discuss and ﬁnd solutions, but they must write their own answers rather than team or group answers; identical answer sets will be considered as cheating. One problem set with the lowest score will be excluded after averaging scores for each student, which means students will have a one-time second chance even after a poor performance.  2. Research Project. During the last module (Module 4), students will be required to do their own research project, making use of knowledge and skills they would have learned over the previous modules. Projects can be either individual or collaborative. Students will choose their own data sets, propose their own research questions and hypotheses they intend to test in order to address the questions, and design statistical models, perform analyses preferably using R, and write a research paper, which describes their ideas and work, reports results, and provides discussions. A paper should not be longer than 20 pages with double space. |
| Current | Activity |  | x | x | x | Students are expected to ask meaningful questions and participate in the discussions, as well as helping other students during practices. |
| Midterm exam | Take-home written exam |  | x |  |  | At the end of Module 2 students are given written assignment to prove the skills they developed during Module 2. Students will be given two days (48 hours sharp) to complete and submit their answers after an exam is posted at the class website. Students will be supposed not to discuss or work with other students. Students should submit the hard copies of their answers to instructor’s oﬃce by the due time. |
| Final exam | Exam (take-home written assignment) |  |  |  | x | At the end of Module 4 students are given written assignment to prove the skills they developed during the entire course. Students will be given two days (48 hours sharp) to complete and submit their answers after an exam is posted at the class website. Students will be supposed not to discuss or work with other students. Students should submit the hard copies of their answers to instructor’s oﬃce by the due time. |

## 7. Grading criteria

The evaluation is based on five criteria: activity, home works, mid-term exam, final exam, and the research project.

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

*Home works.* 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; correct graphics – proper type of plot and formatting should be chosen. Homeworks are graded from 0 (“extremely poor”=”fail”) to 10 (“perfect”=”pass”) each. Level of English language does not affect the grade.

*Research project* is aimed to show that students are ready to apply the skills they got in this course to the real data and real goals. To get high grade, students should excel in doing the following: 1) formulate and reason research question and hypotheses; 2) choose and describe appropriate variables; 3) choose and reason correct method to achieve the research goals; 4) provide the analysis and correctly interpret the results; 5) summarize study results and give it a possible explanation. Oral presentation is optional and allows getting one extra point. If project is made in group, each group member gets the same grade.

*Mid-term and final exams* are aimed to check the skills students should have gotten 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 for the first module are calculated by the following scheme:

*Cumulative score1*= 0,9\*mean(home works) + 0,1\*activity

*Midterm grade* = 0,6\*cumulative score1 + 0,4\*midterm exam

The grades for the fourth module are calculated by the following scheme:

*Cumulative score2* = 0,1\*activity + 0,6\*mean(home works) + 0,3\*project

*Final grade* = 0,8\*cumulative score2 + 0,2\*final exam

# 8. The Course Content

**Topic 1. Random variable: distribution, types of variables. Sampling and estimation. Inference and statistical testing**

Random variable (RV): distribution, types of random variables. Estimating a RV (or its distribution): mean, median, variance. Skewed distribution. Interquartile range. Types of distribution. Sample and sampling error. Statistical inference and testing.

*Core reading*

Andy Field, Jeremy Miles, Zoe Field (2012). Discovering statistics using R. - Sage

*Additional reading*

Роберт И. Кабаков. 2014. R в действии. Анализ и визуализация данных на языке R, ДМК Пресс.

Alan Agresti & Barbara Finlay. 2007. Statistical Methods for the Social Sciences, Fourth Edition, Pearson Prentice Hall.

**Topic 2. Probability theory. Distributions. Statistical testing**

Relationship of two random variables: covariance and correlation. Probability theory. Normal distribution. Properties of normal distribution. Transformations of random variable: log transformation; standardization. Central limit theorem. Confidence intervals. Statistical testing: assumptions; hypotheses; test statistic; p-value; conclusion and interpretation.

*Core reading*

Andy Field, Jeremy Miles, Zoe Field (2012). Discovering statistics using R. - Sage

*Additional reading*

Alan Agresti & Barbara Finlay. 2007. Statistical Methods for the Social Sciences, Fourth Edition, Pearson Prentice Hall.

Paul Teetor. 2011. R Cookbook (O'Reilly Cookbooks), 1st Edition, O'Reilly Media.

**Topic 3. Linear Regression Model**

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, interactions, standardized coefficients, drawing plots, practice in R.

*Core reading*

John Fox & Sanford Weisberg. 2011. An R Companion to Applied Regression, Second Edition,

Sage Publications.

*Additional reading*

Andrew Gelman & Jennifer Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press.

Kurt Taylor Gaubatz, 2014. A Survivor's Guide to R: An Introduction for the Uninitiated and the Unnerved, 1st edition, Sage.

**Topic 4. Logistic regression model**

Introduction to models with categorical outcomes, models predicting probabilities of dichotomous outcomes, linear probability model, basic logic of binary logit model, interpretation of logit coefficients and odds ratios, computing predicted probabilities and marginal effects, drawing plots of marginal effects, statistical tests, model selection, practice in R

*Core readings*

John Fox & Sanford Weisberg. 2011. An R Companion to Applied Regression, Second Edition,

Sage Publications.

*Additional reading*

D. W.Hosmer Jr, S.Lemeshow, R. X.Sturdivant. 2013. Applied logistic regression, John Wiley & Sons.

John Fox. 2008. Applied Regression Analysis and Generalized Linear Models, Second Edition, Sage Publications.

**Topic 5. Multinomial logit model**

Basic logic and functional form of multinomial logit model, interpretation of logit coefficients and odds ratios, computing predicted probabilities, marginal effects and drawing plots of marginal effects, statistical tests, model selection, practice in R.

*Core reading*

Andy Field, Jeremy Miles, Zoe Field (2012). Discovering statistics using R. - Sage

*Additional reading*

John Fox. 2008. Applied Regression Analysis and Generalized Linear Models, Second Edition, Sage Publications.

D. W.Hosmer Jr, S.Lemeshow, R. X.Sturdivant. 2013. Applied logistic regression, John Wiley & Sons.

**Topic 6. Ordered logit model**

Basic logic and functional form of ordered logit model, interpretation of logit coefficients and odds ratios, computing predicted probabilities, marginal effects and drawing plots of marginal effects, statistical tests, model selection, practice in R

*Core readings*

D. W.Hosmer Jr, S.Lemeshow, R. X.Sturdivant. 2013. Applied logistic regression, John Wiley & Sons.

*Additional reading*

Andrew Gelman & Jennifer Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press.

John Fox. 2008. Applied Regression Analysis and Generalized Linear Models, Second Edition, Sage Publications.

**Topic 7. Models for count outcomes**

Understanding the nature of count variable, basic logic and functional form of poisson regression model and negative binomial model, interpretation of coefficients, computing predicted probabilities, marginal effects and drawing plots of marginal effects, statistical tests, other models for count outcomes, practice of poisson regression model and negative multinomial model in R

*Core readings*

Joseph Michael Hilbe. 2014. Modeling Count Data, 1st Edition, Cambridge University Press.

*Additional reading*

Daniel A. Powers & Yu Xie, 2000. Statistical Methods for Categorical Data Analysis, Second Edition, Academic Press.

Роберт И. Кабаков (2014). R в действии. Анализ и визуализация данных на языке R – ДМК Пресс.

**Topic 8. Event history modeling: discrete-time hazard model**

Introduction to longitudinal, panel data, basic logic of event history modeling, conceptual understanding of various event history models, introduction to discrete-time hazard rate model as the simplest model of the event history family, understanding time-varying hazard rates and drawing plot, interpreting coefficients and statistical tests, computing predicted probabilities (hazard rates) and drawing plots, model selection, practice in R

*Core reading*

Jeﬀrey M. Wooldridge. 2013. Introductory Econometrics: A Modern Approach, Fifth Edition,

South-Western Cengage Learning.

*Additional reading*

Kurt Taylor Gaubatz, 2014. A Survivor's Guide to R: An Introduction for the Uninitiated and the Unnerved, 1st edition, Sage.

Paul Teetor. 2011. R Cookbook (O'Reilly Cookbooks), 1st Edition, O'Reilly Media.

**Topic 9. Fixed-eﬀects and random-eﬀects models**

Extending linear regression to panel models, logics of random-effects and fixed-effects models, interpreting the models, interclass correlation, practice in R

*Core reading*

Jeﬀrey M. Wooldridge. 2013. Introductory Econometrics: A Modern Approach, Fifth Edition,

South-Western Cengage Learning.

*Additional reading*

Andrew Gelman & Jennifer Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press.

Paul Teetor. 2011. R Cookbook (O'Reilly Cookbooks), 1st Edition, O'Reilly Media.

**Topic 10. Weights and Missing data**

Basic logic of weighting, sampling weights and analytic weights, issues of missing data, options for handling missing data, conceptual approach to multiple imputation methods

*Core reading*

Trivellore Raghunathan (2015) Missing Data Analysis in Practice - CRC Press

*Additional reading*

Kurt Taylor Gaubatz, 2014. A Survivor's Guide to R: An Introduction for the Uninitiated and the Unnerved, 1st edition, Sage.

Роберт И. Кабаков (2014). R в действии. Анализ и визуализация данных на языке R – ДМК Пресс.

# 9. Educational Technologies

The course syllabus, reading texts, presentations, practical tasks and home assignments will be available in the Google group “HSE Data Analysis class” and on the course website. Students are expected to log in to the course web-site on a regular basis.

9.1. Methodical recommendations for students.

Students are recommended to read lots of scientific articles to learn how to apply statistical methods and, most importantly, describe the results. They are also expected to adopt the knowledge the got during these classes in their course papers and to discuss with the teaching stuff any issues arising during this process.

# 10. Diagnostics for Monitoring Students' Work and Progress

To assess students’ progress they will be given home tasks each two weeks. Each home task is based on the materials from the last lecture and includes some theoretical questions and practical tasks that have to be done using R.

Example:

* *What is the null hypothesis? What is the alternative hypothesis?*
* *Find the test statistic in the result reported by R. Show how the resulting test statistic could be computed.*
* *What is the p-value and what does the p-value mean?*
* *Should we reject the null hypothesis? Can we conclude that the estimated relationship between household size and satisfaction would hold true in the population?*
* *Regress satisfaction on gender, race, age, education (years), high school degree, and household income. Explain gender gap and race gap in life satisfaction among South Africans.*

Besides home tasks, students have to pass midterm and final exam. Their structure is similar to home tasks, though it covers all topics studied:

* during the second module – for midterm exam
* during modules 2-4 – for final exam

Example of exam questions:

* *Compute the correlations for the pairs of the continuous variables below, by computing the covariances and standard deviations (Do not use R functions that yield the correlation directly)*
* *Fit a linear regression model examining (a) whether the association between political orientation and family income is curvilinear and (b) whether the relationship differs by education (high school vs. college or more). Draw a plot that describes the fitted pattern over family income. (Note: You do not have to include other control variables.)*
* *Fit a logit model predicting attitude toward abortion by race and gender, holding education constant (no interaction between race and gender). Interpret the result of race and gender using odds ratios. Which group do you think is most outstanding in their attitude toward abortion?*
* *Fit the model predicting abortion by gender, race, educ, south, rural and age. Compute the marginal effects of educ, south, rural, and age for white men (hint: female=0, all race dummies=0) with all other variables are fixed at their means (educ: 0.386, south: 0.376, rural: 0.209, age: 49.429). Interpret the marginal effects.*

The last form of diagnostics of students’ performance is research project. It is assigned in the last module and might be accomplished individually or in small groups. The topic of the project is chosen by the student. Students will get several datasets prepared by the teaching stuff to choose from. They might as well take another dataset. The final project should include following sections:

* Introduction. Statement of the problem. Null and alternative hypotheses
* Descriptive analyses of the data
* Methodology: brief description of the statistical method chosen
* Analysis
* Results

# 11. Information basis for the course

# 11.1. Core Textbook

John Fox & Sanford Weisberg. 2011. An R Companion to Applied Regression, Second Edition,

Sage Publications.

**11.2. Internet resources**:

# John Verzani, Simple R: Using R for Introductory Statistics. Available at <http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>

# UCLA Stats Lab, <http://www.ats.ucla.edu/stat/r>

## 11.3. Supplementary Reading

Andrew Gelman & Jennifer Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press.

John Fox. 2008. Applied Regression Analysis and Generalized Linear Models, Second Edition, Sage Publications.

John Fox & Sanford Weisberg. 2011. An R Companion to Applied Regression, Second Edition,

Sage Publications.

Daniel A. Powers & Yu Xie, 2000. Statistical Methods for Categorical Data Analysis, Second Edition, Academic Press.

Jeﬀrey M. Wooldridge. 2013. Introductory Econometrics: A Modern Approach, Fifth Edition,

South-Western Cengage Learning.

Alan Agresti & Barbara Finlay. 2007. Statistical Methods for the Social Sciences, Fourth Edition, Pearson Prentice Hall.

Andy Field, Jeremy Miles, Zoe Field (2012). Discovering statistics using R. – Sage

Роберт И. Кабаков. 2014. R в действии. Анализ и визуализация данных на языке R, ДМК Пресс.

Paul Teetor. 2011. R Cookbook (O'Reilly Cookbooks), 1st Edition, O'Reilly Media.

Kurt Taylor Gaubatz, 2014. A Survivor's Guide to R: An Introduction for the Uninitiated and the Unnerved, 1st edition, Sage.

D. W.Hosmer Jr, S.Lemeshow, R. X.Sturdivant. 2013. Applied logistic regression, John Wiley & Sons.

Joseph Michael Hilbe. 2014. Modeling Count Data, 1st Edition, Cambridge University Press.

Trivellore Raghunathan (2015) Missing Data Analysis in Practice - CRC Press

## 11.4. Program packages

# 1. R: R is a software that is widely used for data analyses among social scientists in these days, it has been getting popularity increasingly because of its several merits over other stats package tools. First, it is free and easy to get. You can download a recent version of R on the web (https://www.r-project.org/). Second, self-teaching materials, manuals or references for solving problems you may come across are abundant on the web, all free of charge. Third, it is a programming tool, so learning R, even at a very basic level, might provide an interesting experience of getting over any inner barriers in mind and ﬁnding a programming no scary thing with students who are afraid of programming or anything like that. Lastly, R will be extremely useful if you are considering to be a more serious researcher.

# 2. R-Studio: R-Studio is a software proving a more user-friendly environment when you run R. It is also available free of charge at https://www.rstudio.com.

# 12. Technical support

Each lecture is supported by Power Point presentations shown by OHP projector. Seminars and practical sessions are held in a fully-equipped computer class with personal computers available to every student in group (in cases when there are more students than PCs they are welcome to bring their own computers).

# 13. Academic Integrity

Each student in this course is expected to abide by the Higher School of Economics’ Academic Honesty Policy. For this course, collaboration is allowed for pair and group work during seminar classes (Modules 2 – 4) and preparation of group presentation of analytical work (Module 4).