university of groningen

campus fryslân

Statistics II

Syllabus 2022 - 2023

Course code: CFB012A05
Level: Skills Lab
Credits: 5 EC

Course coordinator: Juan Rivillas – j.c.rivillas.garcia@rug.nl
Teaching Assistants: Maryory Galvis-Pedraza - m.galvis.pedraza@rug.nl
Maret, Sturms – m.m.sturms@student.rug.nl

Content

Introduction

The main purpose of the Statistic II is to ensure that students are confident applying the steps in regression modelling process. This course introduces regression modelling, including data exploration, correlation, regression with a single predictor variable, multiple explanatory variables, confounding factors, model interactions, model checking, model fitting, hands-on examples, explore a model created using R, and interpret results.

This second course on statistics combines an informal lecture, individual assignments, and practical modelling to help students with the understanding and application of statistics I (warming up), covering new topics in linear, logistic, and multiple regressions analysis, and finally, the development of a research project will allow students to apply what they learn and interpretate the output using their own datasets.

The focus of the course is to explain how to apply regression methods to model and test data in terms of one or multiple variables. In this course we discuss how to approach modelling problems and draw conclusions from correlated variables, including k-means cluster classification and Latent Class Analysis (LCA) for the identification of unobservable groups in the data. Additionally, we will focus on the application of regression analysis in a practical session followed by how to prepare data for analysis, interpret, evaluate your model, verify model assumptions and validity, and effectively communicate research findings and results. The practical session will require very basic level use of R, but all necessary scripts will be provided to complete your assignments. This course will allow students to apply previous knowledge and branching out their skills in machine learning techniques.

Throughout the course, students are required to use R for their statistical analysis. In terms of research practices, this course builds on Statistics I in maintaining high standards of transparency in research. In addition, due to the increasing complexity in the models, the students will be encouraged to passively (viewing) or actively engage with online communities (such as stack GitHub and r-bloggers) for support and suggestions for future analyses.

Learning goals

Upon the successful completion of this course, students will be able to:

- **Prepare** data into a suitable format.
- **Identify** the best fit model for the classification of clusters and latent groups in the data.
- **Apply** regressions to model a response variable in terms of a single variable or multiple variables.
- **Identify** the correlation coefficient as a single measure of regression models.
- **Assess** model validity by checking model assumptions.
- **Assess** model fitness by comparing the results produced by the model with your data.
- **Use** effectively reports summarizing results to communicate research findings.

Course structure

The course runs for 9 weeks. There are 2 classes of 2 hours a week, two individual assignments (week 4 and week 8) and one research project and oral presentation (weeks 8 and 9).

Practical Information

Literature

1. Grolemund, G., & Wickham, H. (2017). R for Data Science. Sebastopol, CA: O'Reilly Media Inc. (open-access through http://r4ds.had.co.nz/)
2. Diez, D., et al. (2016) Open Intro Statistics (open access) https://www.openintro.org/stat/
3. Ciaburro, G. (2018). Progression Analysis with R. Design and develop statistical nodes to identify unique relationships within data at scale.
4. Lilja, D., and Linse, G. (2022). Linear Regression Using R. An introduction to data Modelling. Second Edition. https://conservancy.umn.edu/handle/11299/189222
5. Caffo B. (2015) Regression Models for Data Science (open-access) https://leanpub.com/regmods/read
6. Rivillas, JC. (2022). Working paper. Identification of latent adverse childhood groups. Imperial College London.
7. van Zwieten A, et al. (2022). Avoiding overadjustment bias in social epidemiology through appropriate covariate selection: a primer. J Clin Epidemiol. 149:127-136. doi:10.1016/j.jclinepi.2022.05.021
8. Soetewey, A. (2020). The complete guide to clustering analysis: k-means and hierarchical clustering by hand and in R - Stats and R. https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/
9. Collins, L. M., & Lanza, S. T. (2010). Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. 1–295. https://doi.org/10.1002/9780470567333

Essentials library

Next to the Student Service Desk, there is small library where you can find the required books for the course. You can borrow the resources when you are studying in the Beurs or make copies of the parts you want to read at home. Please note that it is not allowed to take resources from the library home.

Brightspace

We use the digital learning environment "Brightspace" as the main platform for communication. Here, you'll find recommended literature, information on assignments and your grades. Announcements regarding schedule -or content changes will also be published in Brightspace. Moreover, you will find quick links to SmartCat and diverse RUG tools (such as Ocasys, Enrollment and Photo and wireless printing).

All essential information about the course can be found in this syllabus. However, as we reserve the right to change the syllabus, please keep track of Brightspace for the most up-to-date information.

Assessment

The assessment will be done based on individual computer assignments (35%), final research project (40%), final presentation (15%) and in-class participation (10%). The results of all examinations are given by means of grades ranging from 1.0 to 10.0, with a 5.5 or higher for a passing grade and a 5.5 or less for a failing grade. Each (partial and final) grade is rounded off to one decimal place.

Attendance & Absence

As per the Teaching and Examination Regulations 2022 – 2023:

- Attendance during the first class is mandatory for course participation. If students are not attending the first class of the course unit, the lecturer may require students for alternative efforts to make up for the non-attended class.
- Attendance during all course units is mandatory. If students are not able to attend a class due to extraordinary circumstances (such as sickness), they need to inform the lecturer

and educational secretariat as soon as possible. In the event of absence up to a maximum of three classes per term the instructor may stipulate replacement assignments. Absence of more than three classes per term results in the student being banned from further participation in the course unit and from the final examination. Students have a right to appeal this decision with the individual lecturer. Exceptions to this rule due to extraordinary circumstances only apply if they are known with the Study Advisor.

Cheating and plagiarism

Cheating and plagiarism are academic offences, with severe consequences. They are acts or omissions by students to partly or wholly hinder accurate assessment. Cheating and plagiarism are easy to avoid and prevent by citing the used resources properly. This would enable students to use and share information ethically, with academic integrity and by recognizing other's work. As per the Teaching and Examination Regulations, cases of cheating and plagiarism are reported to Exam Board, that will decide upon the consequences. NB: all assignments are automatically checked on plagiarism.

Contact information

The course coordinator of Statistics II is Juan Rivillas – j.c.rivillas.garcia@rug.nl
Teaching Assistants: Maryory Galvis-Pedraza and Maret Sturms
The office hours are Monday or Wednesdays 11-12m (This will be announced during lectures and via Brightspace).

General questions or suggestions about the course can be addressed to the educational secretariat. Email: cf-sec@rug.nl, phone number: 058-2882132.

Weekly schedule

Week 1 What is exploratory data analysis?

Objectives
Students will be able to:

- **Import** and load datasets.
- **Prepare** datasets for analysis and **getting** data into a usable format.
- **Formulate** questions about a dataset and **generate** useful visualisations for a given question.
- **Identify** potential sources of bias in the data.
- **Explore** distribution, variation, and covariation
- **Analyse** outliers and missing values.

Class
Group A 6 September 2022 10:45 – 12:30
Group B 6 September 2022 15:15 – 17:00
Group C 5 September 2022 10:45 - 12:30

Learning outcomes
The lecture starts with a warming up preparing dataset for analysis. Subsequently, students are introduced to the data, including data importing, data visualisation, data transformation and exploratory data analytics. Exploratory Data Analysis covers distributions, outliers, errors, missing data, variation, and covariation and asking questions about data. Graphics for communication, including principles of data graphics, good and bad practice and grammar of graphics are thoroughly discussed.

Readings
Chapter "Explore" of the R for Data Science and lecture notes.
Chapter 2 "Understand your data" of the Linear Regression Using R.
Chapter 5 "Data preparation using R tools" of the Progression Analysis with R.

Packages in R
Libraries (readxl, dplyr, devtools, Amelia, ggstatsplot, DescTools, table1).

Lab
Group A 9 September 2022 8:45- 10:30
Group B 9 September 2022 10:45- 12:30
Group C 7 September 2022 8:45- 10:30

The lab continues with where the lecture stops and allows students to practice preparing data for analysis using R.

Scripts and slides available in GitHub:
R Markdown "Exploratory data analysis" step by step.
Slides covering basic concepts of the week.

Preparation
Make sure you have access to a computer. Your personal laptop would be best since you can then practise the course material in your own time.

Week 2. What is correlation analysis and testing the significance?

Objectives
Students will be able to:

- **Apply** DAGs to avoid over adjustment, improve variable selection, and modelling accuracy.
- **Apply** basic statistics (Warming up descriptive statistics and normal distributions).
- **Complete** steps to test associations between categorical variables.
- **Interpret** correlograms or correlation matrix.
- **Calculate** correlations for sampled data.
- **Apply** correlations methods to answer questions about a population.

Class
Group A 13 September 2022 10:45 – 12:30
Group B 13 September 2022 15:15 – 17:00
Group C 12 September 2022 10:45 - 12:30

Learning outcomes

The lecture completes the warming up week, including displaying data, interpreting graphical representation of data, and calculating descriptive statistics for data. Understand the importance of establishing whether a linear relationship exists between two things (variables), correlation for all variables (correlalogram or matrix of correlation coefficients), and interpretation of a correlation coefficient. Visualization methods and correlation test are discussed.

Readings
Correlation coefficient and correlation test in R and lecture notes.

Packages in R
Libraries (readxl, dplyr, DescTools, table1, ggstatsplot, ggplot2, compareGroups, BioAge).

Lab
Group A 16 September 2022 8:45- 10:30
Group B 16 September 2022 10:45- 12:30
Group C 14 September 2022 8:45- 10:30

The lab continues where the lecture stops and allows students to further practise their skills in correlation analysis and testing the significance.

Scripts and slides available in GitHub:
R Markdown "Correlation coefficients" step by step.
Slides covering basic concepts of the week.

Preparation
Review the material in the previous lectures. Make sure you have access to a computer. Your personal laptop would be best since you can then practise the course material in your own time.

Week 3. Clustering Analysis and Latent Class Analysis (LCA)

Objectives
Students will be able to:

- **Explain and apply** assumptions and the mechanics behind Cluster Analysis (CA).
- **Explain and apply** assumptions and the mechanics behind Latent Class Analysis (LCA).
- **Perform and visualise** the outcome.
- **Explain** fit and selection criteria and advantages and disadvantages of both methods.
- **Interpret** results and selection best fit model.

Class
Group A 20 September 2022 10:45 – 12:30
Group B 20 September 2022 15:15 – 17:00
Group C 19 September 2022 10:45 - 12:30

This lecture introduces the students to the most popular exploratory method: cluster analysis (CA) and the most reliable: LCA. CA is introduced and motivation for its use is given. Concepts such as dimensionality reduction, data summarization, and variance explained are introduced through practical examples. Next, two clustering algorithms are discussed and their use is exemplified through several datasets. In this lecture, we will explain assignment 1.

Readings
The complete guide to clustering analysis: k-means and hierarchical clustering by hand and in R.
Latent Class and Latent Transition Analysis.
Identification of latent groups and lectures notes.

Packages in R
Libraries (readxl, cluster, factoextra, poLCA, ggplot2, table1, ggstatsplot, scatterplot3d).

Lab
Group A 23 September 2022 8:45- 10:30
Group B 23 September 2022 10:45- 12:30
Group C 21 September 2022 8:45- 10:30

The lab continues where the lecture stops, and students practise Clustering Analysis and Latent Class Analysis (LCA) using R.

Scripts and slides available in GitHub:
R Markdown "identification of latent groups" step by step (cluster analysis and latent class analysis).
Slides covering with basic concepts and assumptions.

Assignment 1
Individual computer assignment 1 is handed out in this week. The deadline for the assignment is Week 5 of the course, at 12pm on the day of your Group's lecture.

Preparation

Make sure you have access to a computer. Your personal laptop would be best since you can then practise the course material in your own time.

Week 4. Exploratory data analysis in practice: Open datasets for SDG monitoring, definition of research question and prepare analysis for data.

Objectives
Students will:

- **Explore** major open datasets for social science research.
- **Use** datasets to collect information for their research projects.
- **Prepare** datasets for analysis and **getting** data into a usable format to monitor the achievement of Sustainable Development Growth.
- **Complete** three steps approach applying DAGs for modelling accuracy.
- **Complete** assignment 1: identification of the latent groups in dataset.
- **Open** GitHub account and publish assignment 1.
- Have a happy coding!

Class
Group A 27 September 2022 10:45 – 12:30
Group B 27 September 2022 15:15 – 17:00
Group C 26 September 2022 10:45 - 12:30

This lecture of the course aims at introducing the students to the major open datasets that most social science research is based on. We will discuss the databases from the World Bank, WHO, OECD, IMF, and the World Value Surveys. Students will be able to use these databases to extract data for their research projects and understand how the goals of the Sustainable Development program are measured and monitored.

Readings and p**reparation**
Lecture slides and online resources provided by the lecturer.
Chapter "Explore" of the R for Data Science and lecture notes.
Chapter 2 "Understand your data" of the Linear Regression Using R.
Chapter 5 "Data preparation using R tools" of the Progression Analysis with R.

Packages in R
Libraries (readxl, dplyr, devtools, Amelia, ggstatsplot, DescTools , table1sjPlot, ggplot2, table1, ggstatsplot).

Lab
Group A 30 September 2022 8:45- 10:30
Group B 30 September 2022 10:45- 12:30
Group C 28 September 2022 8:45- 10:30

The lab continues where the lecture stops, and students collect data and prepare for the final research project.

Scripts and slides available in GitHub:
All published R Markdowns.
Lecture's slides covering basic concepts and assumptions through weeks.

Preparation
Review the material in the previous lectures. Make sure you have access to a computer. Your personal laptop would be best since you can then practise the course material in your own time.

Week 5. Linear Regression.

Objectives
Students will be able to:

- **Explore** classification of the regressions based on assumptions, purposes, and uses.
- **Identify** the correlation coefficient as a single measure of linear association.
- **Apply** linear regression with a single predictor variable.
- **Assess** model validity by checking model assumptions.
- **Assess** model fitness by comparing the results produced by the model.
- **Explain** main purposes of regression - residuals and least squares.
- **Interpret** the weaknesses of Linear Models for binary and count outcomes.

Class
Group A 4 October 2022 10:45 – 12:30
Group B 4 October 2022 15:15 – 17:00
Group C 3 October 2022 10:45 - 12:30

Students will be taught the pros and cos of Linear Models for the analysis of binary choice models and count data. Hypothesis testing and marginal effects interpretation are discussed. Students deepen their knowledge about linear regression discussing how to avoid over adjustment, improve selection of the variables, ensure model accuracy. Interpretation of coefficients, standard error, statistical significant test, R square, exploring model results, and diagnostic plots are provided.

Reading
Avoiding overadjustment bias in social epidemiology through appropriate covariate selection.
Chapter 2 "Basic concepts -Simple Linear Regression" of the Progression Analysis with R.
Chapter 3 "Simple Linear Regression" of the Linear Regression Using R.

Packages in R
Libraries (readxl, sjPlot, dplyr, ggplot2, table1, ggstatsplot).

Lab
Group A 3 October 2022 8:45- 10:30
Group B 10 October 2022 8:45- 10:30
Group C 5 October 2022 8:45- 10:30

The lab continues where the lecture stops, and students practice Linear Regression modelling using R.

Assignment 1
Deadline for assignment 1.

Scripts and slides available in GitHub:
R Markdown "Linear regression modelling" step by step.
Slides covering basic concepts and assumptions.

Preparation
Review the material in the previous lectures. Make sure you have access to a computer. Your personal laptop would be best since you can then practise the course material in your own time.

Week 6. Generalised Linear Models (GLM).

Objectives
Students will be able to:

- **Interpret** the weaknesses of GLM for categorical predictor variables.
- **Understand** and use Logistic regression for binary dummy variables.
- **Understand** and use Poisson regression.
- **Contact** hypothesis testing using GLMs.
- **Analysis** of Variance (ANOVA) models.
- **Complete** assignment 2: regression analysis.
- **Publish** assignment 2 on GitHub.
- Have a happy coding!

Class
Group A 11 October 2022 10:45 – 12:30
Group B 11 October 2022 15:15 – 17:00
Group C 10 October 2022 10:45 - 12:30

Learning outcomes
Students will be taught the limitation of Linear Models for the analysis of binary choice models and count data. The logistic and Poisson regression models are introduced as an alternative. Hypothesis testing and marginal effects interpretation are discussed. In this lecture, we explain instructions for assignments 2.

Reading
Chapter 4 "Logistic regression" of the Progression Analysis with R.
Practical Regression and Anova using R.

Packages in R
Libraries (readxl, sjPlot, dplyr, ggplot2, table1, ggstatsplot).

Lab
Group A 14 October 2022 8:45- 10:30
Group B 14 October 2022 17:45- 19:00
Group C 12 October 2022 8:45- 10:30

The lab continues where the lecture stops, and students practise logistic regression modelling using R.

Scripts and slides available in GitHub:
R Markdown "Logistic regression modelling" step by step.
Slides covering with basic concepts and assumptions.

Assignment 2
Individual computer assignment 2 is handed out in this week. The deadline for the assignment is Week 8 of the course, at 12pm on the day of your Group's lecture.

Preparation
Review the material in the previous lectures. Make sure you have access to a computer. Your personal laptop would be best since you can then practise the course material in your own time.

Week 7. Multivariate Regression.

Objectives
Students will be able to:

- **Explore** Assumptions and steps in multiple regression modelling process.
- **Test** multiple hypotheses using multiple explanatory variables.
- **Explore** confounding factors.
- **Interpretate** model interactions, model checking, and model fitting.
- **Assess** violations of regression assumptions.
- **Identify** potential pitfalls in multiple regression and how to deal with them.
- **Interpretate** outcomes with multiple variables.
- **Assess** model quadratic relationships and interactions.

Class
Group A 18 October 2022 10:45 – 12:30
Group B 18 October 2022 15:15 – 17:00
Group C 17 October 2022 10:45 - 12:30

This lecture will demonstrate a full round of multiple regression analysis with more than two predictors, including exploratory visualisation, hypotheses testing, estimation, and interpretation of results. Students deepen their knowledge about multiple regression discussing how potential pitfalls might affect the quality of inference. Multicollinearity, endogeneity, omitted variables are discussed.

Reading
Chapter 4 "Multiple Linear Regression" of the Using Linear Regression.
Chapter 3 " More than just one Predictor – MLR" and Chapter 4 " Multiple Logistic Regression" of the Regression Analysis in R.

Packages in R
Libraries (readxl, sjPlot, dplyr, ggplot2, table1, ggstatsplot).

Lab
Group A 21 October 2022 8:45- 10:30
Group B 21 October 2022 17:45- 19:00
Group C 19 October 2022 8:45- 10:30

Scripts and slides available in GitHub:
R Markdown "Multiple regression modelling" step by step.
Slides covering with basic concepts and assumptions.

Assignment 2
Deadline for assignment 2.

Preparation
Review the material in the previous lectures. Make sure you have access to a computer. Your personal laptop would be best since you can then practise the course material in your own time.

Week 8. Regression in practice: Putting it all together into your own research project.

Objectives
Students will:

- **Reflect** on the material covered in the course.
- **Select** a research question to test in the selected dataset.
- **Develop** a research project (pairs)
- **Apply** new machine learning skills in your own dataset: Exploratory data analysis, preparing data for analysis, assumption testing, hypothesis testing, model accuracy and model building, graphics for communication, and R Markdown formats.
- **Use** feedback from the lecturer and each other to come full circle on what they have learned and how to use it.
- Have a chance to ask questions about their final project.
- Have a happy coding!

Class

Group A 25 October 2022 10:45 – 12:30
Group B 25 October 2022 15:15 – 17:00
Group C 24 October 2022 10:45 - 12:30

This lecture is reserved for students to reflect upon what they have learned, actively. This will be done in the form of a general discussion based on the questions that the student brings to the classroom. It is also a second opportunity to receive feedback and diagnose any unusual issues about their final project.

Preparation
Review as much material as possible from the course and collect questions that you would like to clear.

Lab

Group A 28 October 2022 8:45- 10:30
Group B 28 October 2022 17:45- 19:00
Group C 26 October 2022 8:45- 10:30

The lab sessions for this week focus on rapping up the final project of the course. Students should use this time to receive more feedback on their project, particularly, from their fellow students. The lecturer will act as a mediator making sure no obvious mistakes are made.

Week 9. Your Turn!

Objectives
Students will:

- **Present** the results of their final research project, gaining exposure to presenting data driven reports.
- **Publish** your research project and results in the GitHub (slides and R Markdown document).

Class I
Group A 1 November 2022 10:45 – 12:30
Group B 1 November 2022 15:15 – 17:00
Group C 31 October 2022 10:45 - 12:30

Students will deliver a 15-minute presentation of their final project followed by questions by their fellow students and the lecturer. They will also submit a written version of their report and R code used to produce the analysis via Nestor.

Preparation

- Instruction for the format of the report will be uploaded to Nestor.

Class II
Group A 4 November 2022 10:45 – 12:30
Group B 4 November 2022 15:15 – 17:00
Group C 2 November 2022 10:45 - 12:30

This class will be used as extra time for the presentation sessions.

Appendices

Appendix 1. Assignments and Assessment

Computer assignments (35%)
There are two assignments. Assignment 1: identification of observed groups (20%) and Assignment 2: regression analysis (15%). Both computer assignments will consist of a data problem from real world data focussed on the lecture content discussed in the previous 1 or 2 weeks. Students will work in small groups (<=2) and deliver a written report. Instructions will be published on Brightspace.

Final Research Project (40%)
The final research project will involve the students forming groups (2 students) and conducting a small research simulation where they will statistically test hypotheses that they define based on the data. Instructions about the data and the format of the project's report will be published on Brightspace.

Final Presentation (15%)
Students will be evaluated on the quality of their final presentation as part of their grade. Instruction on how to structure the presentation will be given on Brightspace.

Active Participation and Preparation (10%)
Attendance is mandatory in this course. Participation and preparation comprise 10% of the grade. You are expected to be well-prepared for the lectures by means of reading the allocated study material before the lectures as well as doing the individual or group assignments. You are expected to facilitate discussions with peers on topics relevant for the lectures, challenge each other's ideas in a constructive way, and provide each other with feedback when needed, such as during the presentations. Attendance is graded by way of a participation rubric.