****

**Statistics II**

**Syllabus 2022 - 2023**

**Course code**: CFB012A05

**Level:** Skills Lab

**Credits:** 5 EC

**Course coordinator: dr. Dimitrios Soudis – [d](mailto:email@rug.nl).soudis@rug.nl**

# Content

Content 1

Introduction 2

Learning goals 2

Course structure 2

Practical Information 4

Literature 4

Essentials library 4

Nestor 4

Assessment 4

Attendance & Absence 5

Cheating and plagiarism 5

Contact information 5

Weekly schedule 6

Week 1. What is exploratory data analysis? 6

Week 2. What is correlation analysis and test the sugnificance? 7

Week 3. Clustering Analysis and Latent Class Analysis 8

Week 4. Linear Regression 9

Week 5. Generalised Linear Models 10

Week 6. Multi-Variable Regression 11

Week 7. Regression analysis in practice: Open datasets for SDG monitoring 12

Week 8.Putting it all together into your own research project 13

Week 9. Your Turn! 14

Appendices 15

Appendix 1. Assignments and Assessment 15

# 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 (code will be supplied), and interpretate results.

This second course will combine an informal lecture and practical modelling to explain how to apply regression methods to model data in terms of one or multiple variables. We will 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. This course will focus on the application and methods for using regression and will include a practical session followed by how to prepare data for analysis, interpret, evaluate your model, 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 to 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 exchange and r-bloggers) for support and suggestions for future analyses. Knowledge of flexible knowledge resources regarding statistical analysis and data-visualisation will further the students’ abilities to explore more advanced techniques throughout their careers.

## Learning goals

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

* **Prepare** data for regression models.
* **Identify** the correlation coefficient as a single measure of regression models.
* **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 or multiple variables.
* **Assess** model validity by checking model assumptions.

**Assess** model fitness by comparing the results produced by the model with your data.

* **Apply** reproducible research principles to effectively communicate research results.
* **Engage** with online research communities.

## 

## Course structure

The course runs for 9 weeks. There are 2 classes of 2 hours a week (exception week 8 and 9 which has 3 classes).

The course accounts for 5 EC of a total of 140 hrs, distributed across the following elements:

34 hrs computer practical exam and portfolio creation (8 x 3 hour workshop, 10 hour for finalising portfolio).

20 hrs class preparation (1 hr per class)

20 hrs lectures

66 hrs reading and individual exercises (from the mandatory literature, approx. 400 pages at 10 pages per hour, and approx. 26 hours of exercises)

# Practical Information

## Literature

1. Diez, D., et al. (2016) Open Intro Statistics (open access) <https://www.openintro.org/stat/>
2. Ciaburro, G. (2018). Progression Analysis with R. Design and develop statistical nodes to identify unique relationships within data at scale.
3. Lilja, D., and Linse, G. (2022). Linear Regression Using R. An introduction to data Modelling. Second Edition. <https://conservancy.umn.edu/handle/11299/189222>
4. Caffo B. (2015) Regression Models for Data Science (open-access) <https://leanpub.com/regmods/read>
5. Rivillas, JC. (2022). Working paper. Identification of latent adverse childhood groups. Imperial College London.
6. 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
7. 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/
8. 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

At the Student Service Desk, there is small library where you can find a copy of some of the required and recommended literature for the course. You can borrow the resources when you are studying in the Living Lab area 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.

## Nestor

We use the virtual learning environment “[Nestor](https://nestor.rug.nl/)” 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 Nestor. 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 Nestor for the most up-to-date information.

## Assessment

The final grade is compiled based on different elements: active participation, computer exercises, written exam, and group project. The results of all examinations are given in letters on a scale of A – F, expressed as C- or more for a pass and F for a fail, in line with the following grading table:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Letter grades | A+ | A | A- | B+ | B | B- | C+ | C | C- | D+ | D | D- | F |
| Grade point | 4.0 | 4.0 | 3.7 | 3.3 | 3.0 | 2.7 | 2.3 | 2.0 | 1.7 | 1.3 | 1.0 | 0.7 | 0 |

Detailed assignment and assessment descriptions can be found in Appendix 1.

## Attendance & Absence

Attendance of all classes is mandatory. This means that students must actively participate in at least 80% of the classes. In the event of absence of up to 20%, the instructor may stipulate replacement assignments. Absence of more than 20% will result in the student being banned from further participation of the course unit and from the examination. If students do not attend the first class, they will not be able to take the course. If you are unable to attend a class, please inform your lecturer per email and add the educational secretary in cc. All contact details are provided below.

## Cheating and plagiarism

Cheating and plagiarism are academic offences, with severe consequences, but they are easy to avoid and prevent it citing properly used resources They are acts or omissions by students to partly or wholly hinder accurate assessment. This course also aims to equip students with a working knowledge of plagiarism and how to avoid it. This enables students to use and share information ethically, with academic integrity and in accordance with recognize other’s work. Where plagiarism is detected in group work, members of that group may be deemed to have collective responsibility for the integrity for work submitted by that group and may be liable for any penalty imposed, proportionate to their contribution. 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 Title Course is dr. Dimitrios Soudis – [d](mailto:email@rug.nl).soudis@rug.nl.

The office hours are:

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

# Weekly schedule

## Week 1 What is exploratory data analysis?

**Objectives**

* Preparing dataset for analysis.
* Getting data into a usable format.
* Importing and working with data.
* Visualising distributions and graphics for communication.
* Dealing with outliers and missing values.
* Exploring variation and covariation.

Students will be able to:

**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**

* **Format** research data ready for analysis.
* **Formulate** questions about a dataset.
* **Select** a suitable visualisation for a given question.
* **Generate** useful visualisations from your data.
* **Evaluate** the effectiveness of a data visualisation.
* **Identify** potential sources of bias in the data.

The lecture starts with a warming up preparing dataset for analysis. Subsequently, students are introduced to understanding 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 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” with step by step.

Slides covering with 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 test the significance?

**Objectives**

Students will be able to:

* Basic statistics
* Steps to test associations between categorical variables
* Correlation test between variables using R packages.
* Interpretate correlograms or correlation matrix.

**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

* **Warming up** descriptive statistics and normal distributions.
* **Understand** the need to check association before assess causation.
* **Interpret** graphical correlations.
* **Calculate** correlations for sampled data.
* **Apply** correlations methods to answer questions about a population.

The lecture completes warming up week, including displaying data, interpretating 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 test the significance.

**Scripts and slides available in GitHub:**

R Markdown “Correlation coefficients” step by step.

Slides covering with basic concepts of the week.

**1**

13

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

**Objectives**

Students will be able to:

* Understand the mechanics and motivation behind Cluster Analysis (CA).
* Understand the mechanics and motivation behind Latent Class Analysis (LCA).
* Perform CA and LCA on datasets and interpret the outcome.
* Visualise the outcome of CA and LCA.

**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

Learning outcomes

* **Understand** advantages and disadvantages of both methods.
* **Identify** the best fit model in cluster analysis and Latent Class Analysis (LCA).
* **Formulate** comparisons CA and LCA in the data.
* **Interpretate** results and selection best fit models.

This lecture introduces the students to the most popular exploratory method: cluster analysis 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 used is exemplified through several datasets.

**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**

Deadline for assignment 1.

## Week 4. Linear Regression.

**Objectives**

Students will be able to:

* Apply Directed Acyclic Graph (DAGs) for model accuracy.
* Understand purposed of regression - residuals and least squares.
* Understand the use of regression with a single predictor variable.
* Interpret the weaknesses of Linear Models for binary and count outcomes.

**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

**Learning outcomes**

* **Apply** DAGs to avoid over adjustment, improve variable selection, and modelling accuracy.
* **Identify** the correlation coefficient as a single measure of linear association.
* **Apply** linear regression to model a response variable in terms of a single or multiple variables
* **Assess** model validity by checking model assumptions.
* **Assess** model fitness by comparing the results produced by the model.

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 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 practice Linear Regression modelling using R.

**Assignment 2**

Assignment 2 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.

**Scripts and slides available in GitHub:**

R Markdown “Linear regression modelling” step by step.

Slides covering with basic concepts and assumptions.

## Week 5. 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.

**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

**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.

**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 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 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**

Deadline for assignment 2.

## Week 6. Multi-variable Regression.

**Objectives**

Students will be able to:

* Challenges and opportunities of multiple regression.
* Assumptions and steps in multiple regression modelling process
* How to test multiple hypotheses using multiple explanatory variables.
* Confounding factors.
* Interpretation with multiple variables.
* Model interactions, model checking, and model fitting.
* Assess violations of regression assumptions.
* Identify potential pitfalls in multiple regression and how to deal with them.
* How to model quadratic relationships and interactions

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

* **Understand** the challenges and opportunities of multiple regression.
* **Understand** how to test multiple hypotheses using multiple regression.
* **Assess**model fitness by comparing the results produced by the model.

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 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

**Scripts and slides available in GitHub:**

R Markdown “Multiple regression modelling” step by step.

Slides covering with basic concepts and assumptions.

**Preparation**

Review the material in the previous lectures.

## Week 7. Regression analysis in practice: Open datasets for SDG monitoring.

**Objectives**

Students will:

* Become familiar with all major open datasets for social science research.
* Use said datasets to collect information for their research projects.
* Understand the tools used to monitor the achievement of Sustainable Development Growth.

**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 final 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 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

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 with basic concepts and assumptions through weeks.

## Week 8.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, 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.
* 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 thus gaining exposure to presenting data driven reports.
* Publish result 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%)**

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 (<=3) 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 Nestor.

**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 Nestor.

**Course Participation (10%)**

Students are expected to actively participate during the course lectures and tutorials. This will also make up 10% of their final grade.