MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 0, spring 2023

Introduction and overview

https://canvas.education.lu.se/courses/23000

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

20/3-23

Formalia

Teachers

Lecturer: Anna Lindgren, MH:136, anna.lindgren@matstat.lu.se
TAs: Vasilii Goriachkin and Shokoufa Zeinali (PhD students)

Literature

- The lecture slides and accompanying R-code examples.

Suggested literature (available as e-books)

- Rawlings, Pantula, Dickey: Applied Regression Analysis - A Research Tool, 2ed 1998, Springer
- Agresti, A. An Introduction To Categorical Data Analysis, 2ed Wiley, 2007

Course registration by 31 March in Ladok

Course register yourself at www.student.lu.se or www.student.ladok.se.

Cannot be registered?

Required: a basic course in statistics or mathematical statistics.
Contact studierektor@matstat.lu.se

Early drop-out by 7 April in Ladok

- Not registered: decline your application
- Registered: make an early termination (tidigt avbrott).

Late application by 27 March

- FMSN30/FMSN40: see www.student.lth.se/kurs-och-programinformation/kursanmaelan-och-registrering/kursanmaelan/.
- MASM22: www.antagning.se

Mozquizto for Lab 1–3

Login at `quizms.maths.lth.se` using CAS-login with your student account (same as for Canvas) and make sure you see the tests for Lab 1–3 on Tuesday 21 March.

Pair up for Project 1

- Pair up (2 persons) for Project 1. Use the discussion forums to find someone to work with, then go to People and Groups and place yourselves in the same P1-group, by 2 April!
- Groups for Project 2 ("P2") and 3 ("P3" or "P3N40"). Default is clones of the Project 1 groups.

Handing in project reports

- Login to Canvas.
- Make sure you have paired yourself in a corresponding group (P1, P3N40, P2 or P3) with the correct co-authors **before** you submit! Only one person in the group should submit.
- Go to Assignments and the relevant assignment.
- Use the "Submit assignment" button in the top right corner.
- Only send the allowed file types, usually pdf and/or R files. Submit all files at the same time!

Course deadlines: part 1

- 23.59 Sun 2/4. Form groups for Project 1 (P1).
- 23.59 Wed 5/4. Pass Lab1.A+B+C in Mozquizto.
- 23.59 Wed 5/4. Pass Lab2.A+B+C in Mozquizto.
- Easter break
- 12.30 Wed 26/4. Preliminary report for Project 1.
- 13.00 Thu 27/4. Peer assessment comments for Project 1.
- 17.00 Thu 27/4. Preliminary plan for Project 3 (FMSN40).
- 17.00 Fri 28/4. Final report and R-code for Project 1.
- 23.59 Fri 28/4. Pass Lab3.A+B in Mozquizto.

Course deadlines: part 2

- 12.30 Wed 10/5. Preliminary report for Project 2.
- 13.00 Thu 11/5. Peer assessment comments for Project 2.
- 17.00 Fri 12/5. Final report and R-code for Project 2.

Course deadlines: part 3

- 17.00 Tue 16/5. Revised plan+data for Project 3 (FMSN40).
- Tue 23–Fri 26/5: Oral presentation of Project 3.
- Mon 29/5–Wed 21/6 and Mon 14–Fri 25/8: Oral exam.

Why Modelling?

This course introduces some ideas to deal with modelling of dependencies between several variables.

- "All models are wrong, but some are useful." [George Box]
- Everything in Nature and in Society varies. Most of such variability cannot be captured with deterministic mathematics and physics.
- Statistics is a powerful mathematical science to extract information, make predictions, find relations in large amounts of data and to model knowledge about uncertain phenomena. It's the Mathematics of Uncertainty!
- On TED-talk someone even proposed to teach Statistics before Calculus![1]

[1] http://tinyurl.com/nw8uyo

Why Modelling?

- New Statistical journals are created each year.
- Dozens of new articles are published every day.
- Larger and increasingly complex models are created in order to deal with the increasing amount of data we are all exposed to.
- In this course we start with basic but still very relevant statistical models.
- Understanding the main tools for linear models is fundamental as these are also used for nonlinear models, with some technical modifications.
- My hope for this course is to serve as a useful basis to enlarge your understanding of data dependencies, check the importance of statistical assumptions and motivate you in studying more ambitious methods beyond the scope of this course.

Which models?

We will consider modelling the linear relationship between some continuous variable Y depending on:

1. a single variable X (simple linear regression);
2. several variables $X_1, ..., X_p$ (multiple linear regression);

We will also consider modelling the nonlinear relationship between some binary variable Y

- depending on $X_1, ..., X_p$ (logistic regression).

Additional models will be considered for discrete Y.
All the above will be complemented with specific statistical inference tools.

Long term goals

During the next months we learn how to answer the following:

- are my modelling/probabilistic assumptions satisfied?
- how do I test whether there is an effect of a variable on another variable?
- or in other words: how do I assess the statistical significance of said effect?
- is my model "explaining" enough of the phenomenon variability?
- is my model satisfactory or is it too big/small to represent variability?

To consider the above we introduce concepts and constructs that will help you study further modelling tools beyond what is covered in this course.

Data variables

Typical variables:

- **Continuous**: e.g. blood pressure, height, result of a physical measurement,...
- **Count**: discrete, counting the number of events, e.g. number of accidents,...
- **Categorical**: discrete or qualitative. Such as *gender (M/F)*, *political preference (left/right)*, *eye color (green/blue/brown/...)*
  - nominal categorical have no intrinsic ordering: *gender*, *eye colour*...
  - ordinal categorical have some natural ordering/ranking: *studies degree (bachelor/master/PhD)*, *satisfaction (not at all/sometimes/often)*.

We start with simple linear regression.

We gradually evolve to:

- multiple linear regression
- logistic regression
- Poisson and negative binomial regression
- Generalized linear models: GLMs contain all the above as special cases
- quantile regression

Simple linear regression models are indeed very simple.

However, this way we can easily construct tools that will be trivially extended to more complex models.