CS7DS3 Applied Statistical Modeling: Overview

Arthur White

Preliminaries

CS7DS3 Applied Statistical Modeling

- Instructor: Arthur White
- Email: arwhite@tcd.ie
- Office: Room 144, Lloyd Building
- Office hours: 10-12am Fridays
- Email me to schedule a meeting, or I can also meet you remotely using Teams
- All material will appear on blackboard and class page: [scss.tcd.ie/~arwhite/Teaching/CS7DS3.html]

Basic Structure

- Lectures
  - Monday 1pm LB 1.07
  - Friday 9am LB 1.07
- Supporting videos will also be available on blackboard
- Case studies will accompany lecture material
  - Do these in your own time - i.e., no lab
  - Contact me and post on blackboard with any questions you have
  - These are highly relevant to assessment

Communication

- Email: arwhite@tcd.ie
  - Please use CS7DS3 in subject heading of all communication
- Discussion board
  - Ask any/all questions here
  - Create new threads or join existing ones
  - Anonymous questions are fine
  - *Please be respectful* to me and each other in all interactions
- Your interaction and feedback are *crucial*
- Input from class reps always useful

Assessment

- This module is assessed 100% by coursework, i.e., *no exam*
- 2 x small assignments: 15% each. These will be problem sets
- Main assignment: 70%. This will be a report describing a detailed analysis of a complex data set
- All assignments will be submitted through Turnitin
- These will be scheduled with goal to give you plenty of time to complete, especially main assignment. More details to follow.

Online materials

- Case studies will use R.
- You should download R [http://www.r-project.org/], and Rstudio [https://www.rstudio.com/home/]
- Both are open source and free to download
- You will also need Word, or similar (LaTeX and Markdown are also good options) for your main assignment.
- Python is an option, but will not be supported by me

Reading material

There is no compulsory textbook for this course, but the following cover different aspects of the material:

- P.D. Hoff, A first course in Bayesian statistical methods. Springer, 2009. Library e-link: http://stella.catalogue.tcd.ie/iii/encore/record/C__Rb17405199
- S.N. Wood, Core Statistics. Cambridge University Press, 2015. Library link: http://stella.catalogue.tcd.ie/iii/encore/record/C__Rb16031862 and free pdf online: https://people.maths.bris.ac.uk/~sw15190/core-statistics.pdf

Reading material continued

- C.M. Bishop, Pattern recognition and machine learning. Springer, 2006. Library link: http://stella.catalogue.tcd.ie/iii/encore/record/C__Rb16031862 Free pdf: https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/
- B. Efron & T. Hastie. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science Cambridge University Press, 2016. Free pdf: https://web.stanford.edu/~hastie/CASI_files/PDF/casi.pdf

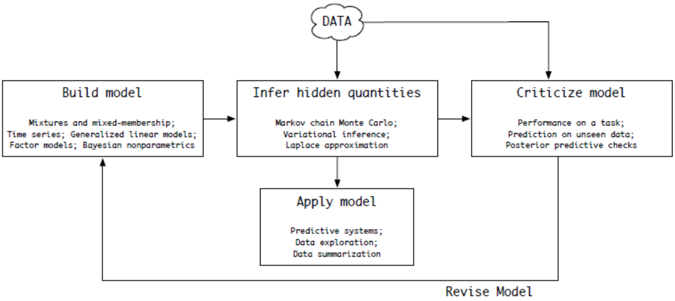
Questions?

Module overview

Description

- This module will provide an overview of statistical models and how to apply them to analyse data.
- We will focus on theory and application:
  - how to build models
  - how to fit them to data
  - how to evaluate their performance
- Our models will be motivated by different research problems

Build, compute, critique, repeat



Description

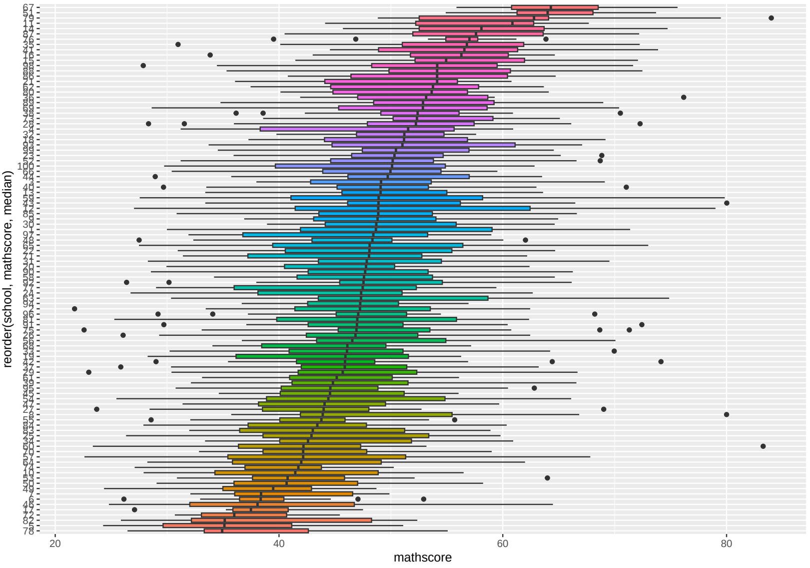
- how to build models:
  - simple and hierarchical models
  - regression models
  - latent variable models
- how to fit them to data:
  - frequentist and Bayesian frameworks for inference
  - optimisation and Monte Carlo computational methods
- how to evaluate their performance:
  - interpretation of parameters
  - model diagnostics
  - model comparison
  - visualisation and communication of results

Examples – schools data

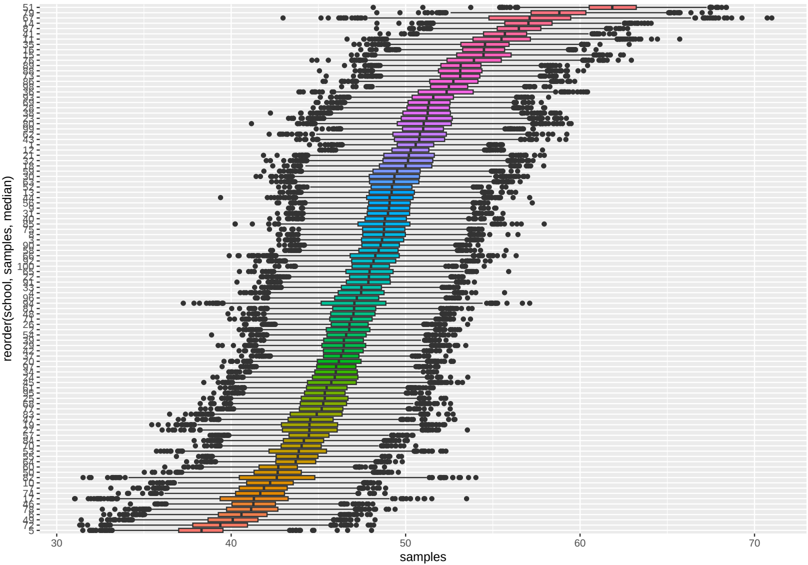
- Students from 100 different schools take a standardised test.
- Can we quantify which schools are best? By how much?

```
##    school mathscore
## 1       1     52.11
## 2       1     57.65
## 3       1     66.44
## 4       1     44.68
## 5       1     40.57
## 6       1     35.04
## 7       1     50.71
## 8       1     66.17
## 9       1     39.43
## 10      1     46.17
## 11      1     58.76
## 12      1     47.97
```

Schools data



Schools data – hierarchical analysis

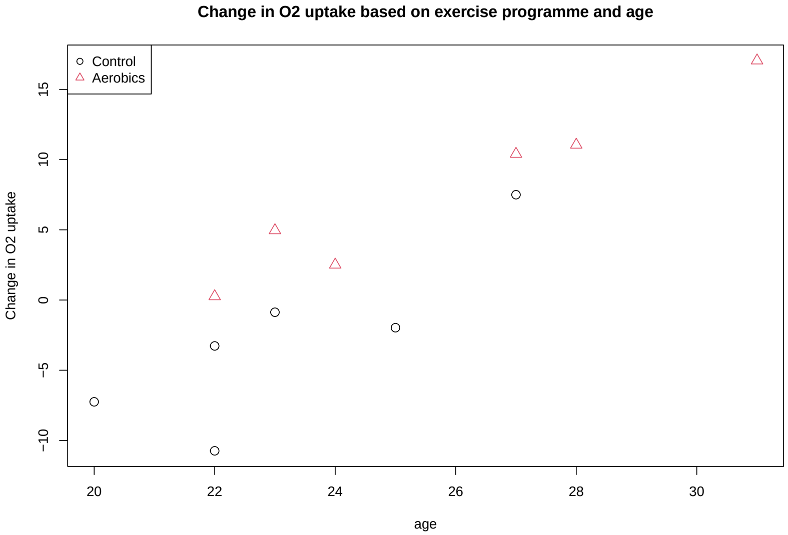


Examples – aerobics programme

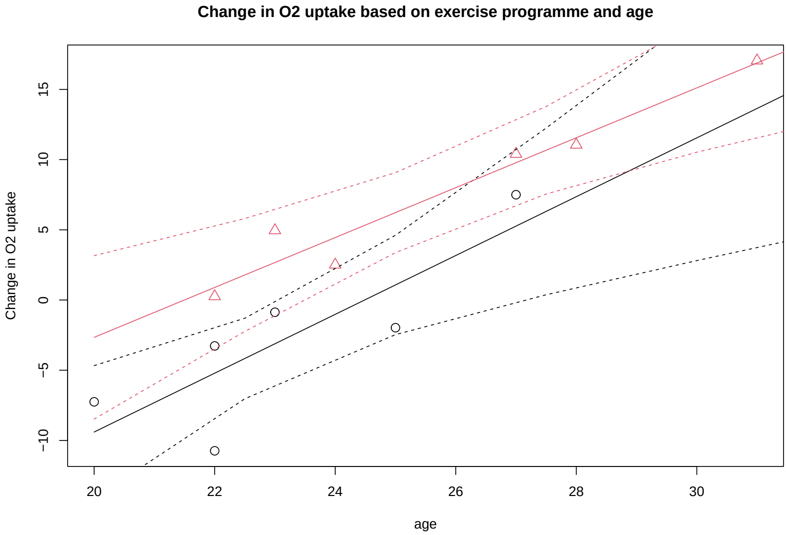
- Special aerobics vs standard running programme, $n = 12$
- Can we quantify the programme's effect on oxygen increase, **accounting for age**?

```
##    uptake aerobic age
## 1   -0.87       0  23
## 2  -10.74       0  22
## 3   -3.27       0  22
## 4   -1.97       0  25
## 5    7.50       0  27
## 6   -7.25       0  20
## 7   17.05       1  31
## 8    4.96       1  23
## 9   10.40       1  27
## 10  11.05       1  28
## 11   0.26       1  22
## 12   2.51       1  24
```

Aerobics programme data



Aerobics programme – regression output

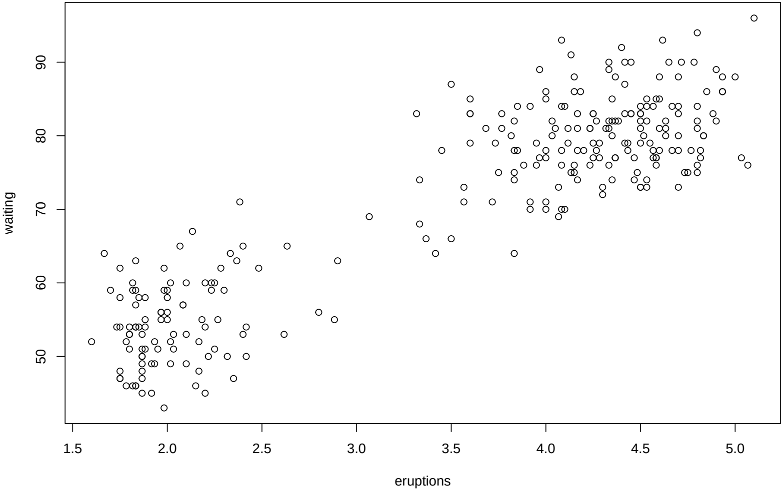


Examples – Old Faithful data

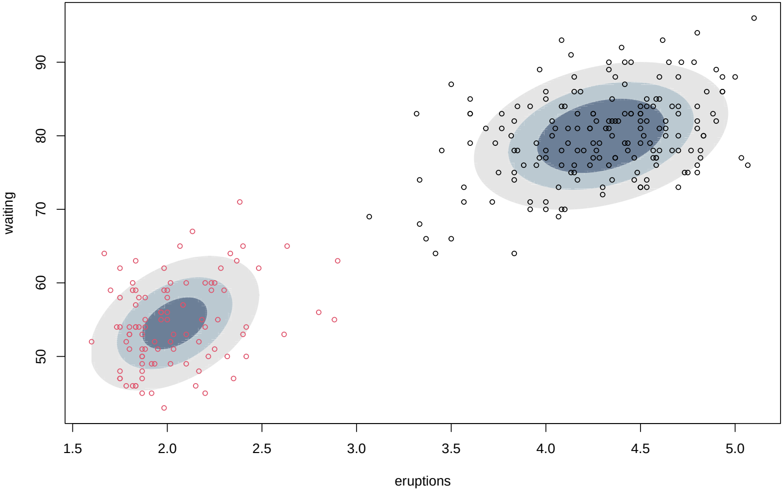
- Duration of eruption and time between eruptions of Old Faithful geyser, $n = 272$
- Can we identify groups of similar eruptions times?

```
##    eruptions waiting
## 1      3.600      79
## 2      1.800      54
## 3      3.333      74
## 4      2.283      62
## 5      4.533      85
## 6      2.883      55
## 7      4.700      88
## 8      3.600      85
## 9      1.950      51
## 10     4.350      85
## 11     1.833      54
## 12     3.917      84
```

Old Faithful data



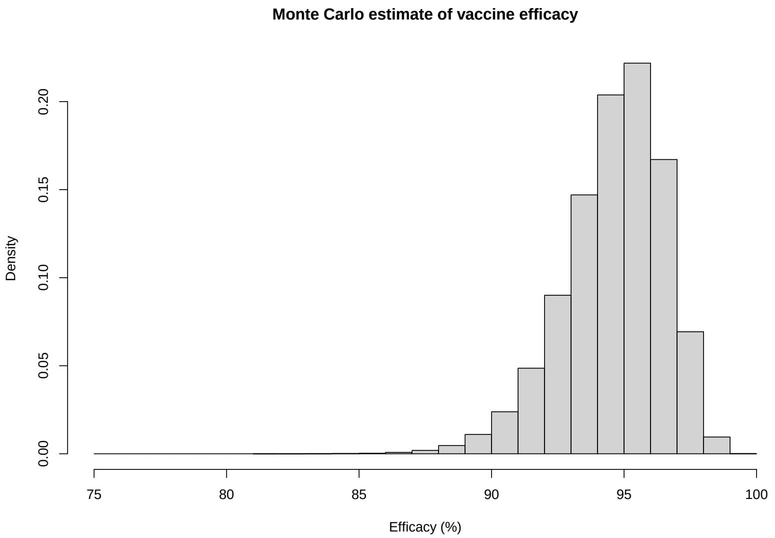
Old Faithful data – cluster analysis



Example: COVID-19 Vaccine

- "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine," Polack et al (2020)
- Sample of 43,548 participants randomized to receive mRNA Covid-19 Vaccine or placebo (i.e., control group).
- Authors report that 8 cases of Covid-19 recorded from vaccinated patients while 162 cases recorded from the control arm.
- Vaccine efficacy estimated by $VE = 100 \times (1 - RR)$, where RR is the estimated ratio of confirmed cases of Covid-19 in vaccine vs. placebo groups.
- How effective is the vaccine?

COVID-19 Vaccine – Efficacy estimate



Summary of examples

- These examples have some aspects in common
  - Using data (specific individuals) to describe a population (model parameters)
  - Limited/finite sample sizes (at least e.g., per group)
  - Presence of variability in the data – **must** be accounted for
- The models that were used, and the research question involved had different aims
  - Schools data: a hierarchical model is used to compare the performance between schools
  - Aerobics data: a regression model compares groups of participants, while accounting for their age
  - Faithful data: a cluster model identifies hidden structure in the data
  - Covid vaccine: we quantify the efficacy of the vaccine, including the uncertainty in our estimate
- In general, the statistical models that we fit to the data will depend on the nature of the data in question **and** the point of view of the research stakeholders

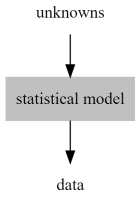
Questions of interest

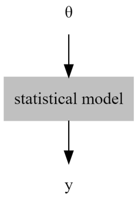
- Generically, a statistical model will have parameter(s) θ.
- A typical statistical analysis will be interested in answering some or all of of these questions:
  - **Point estimation:** What value(s) of θ, is(are) most consistent with the observed data y?
  - **Hypothesis testing** Are the estimated value(s) of θ consistent with some pre-specified value(s) θ_0?
  - **Interval estimation:** What range(s) of value(s) of θ are most plausibly consistent with y?
- Often we will have different, related models for consideration, in which case we will have to decide:
  - **Model choice:** Which model is the most appropriate to use for the data? Which aspects of the data (i.e., variables) are most important to our research objective?
- We will examine how to answer these questions using both frequentist and Bayesian methods

Statistical models

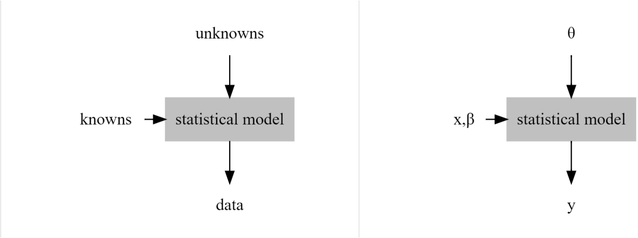
- The statistical models we will investigate will involve some of the below:
  - y: a random vector of the observed data
  - θ: a vector of parameters of unknown value
  - x: covariates/explanatory variables/predictor variable
  - β : the associated coefficients to x
  - Z : a latent variable that identifies hidden structure in the data

A simple/hierarchical model - schematic

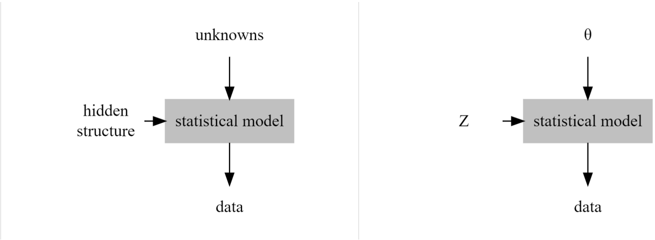




A regression model - schematic



A latent variable model - schematic



Summary

- We are going to study different models:
  - simple and hierarchical models
  - regression models
  - latent variable models
- We will use frequentist and Bayesian inference frameworks to estimate model parameters
- We will use optimisation and Monte Carlo computational methods
- We will communicate our findings in terms of the original context of the research question