Introduction to MTHM506/COMM511

Module plan

- **Term 2, weeks 1-6**
- **“Drop-in” Sessions**
  - Tuesday 16:35-17.25
  - Wednesday 13:35-14:25
  - Wednesday 15:35-16:25
- During these sessions, I will review the notes and will highlight/expand on key material.
- I will not be covering the notes in detail like in the lecture videos. Instead I will use practical examples used to illustrate how we apply the knowledge learned from the recorded lectures.

Module plan

- The format of the drop-in sessions will be flexible and are an opportunity to ask questions. I will inform you of the material I intend to cover on the ELE page.
- If you have any questions beforehand or material you want covered specifically, please email me (m.l.thomas@exeter.ac.uk).
- You are strongly encouraged to read through the slides and watch the lecture videos beforehand to get the most out of these sessions.

Module plan

- **Term 2, weeks 1-6**
- **“Practical” Sessions**
  - Monday 13:35-14:25 (Group 1)
  - Friday 10.35-11.25 (Group 2)
- Due to limited space in computer labs, you have been split into two groups (Group 1 and 2) for the sessions that are labelled as "Practical" sessions.
- You are required to come to the sessions that you are timetable to attend, due to these limited numbers. I will be covering exactly the same material in each of the sessions.

Module plan

- During these sessions, you will have a series of practical exercises to work through and are able to ask questions on the material.
- These are designed to reinforce the material particularly the practical aspects.
- Solutions to problems will be made available a week after finishing the corresponding topic.
- Practical aspects of the course will use R.
- R scripts associated with each topic will be available on ELE, so you can go over these on your own time, even before the practical sessions.
- I do not intend to spend much time explicitly teaching you to use R from scratch. I will therefore not assess R programming skills.

Module plan

- I will set up synchronous zoom sessions to ensure those that have to self-isolate or are not in Exeter can participate in all the sessions.
- I will also record these so that they can be referred back to at a later date.
- If you are able to attend sessions in person though, you are strongly encouraged to do so to ensure you and I get the most out of the sessions.

Syllabus

- **Topic 0** - Foundation material (non-assessed) **(Weeks 1-6)**
- **Topic 1** - Introduction to statistical data modelling **(Weeks 1-2)**
- **Topic 2** - Generalized linear models (GLMs) **(Weeks 3-4)**
- **Topic 3** - Generalized additive models (GAMs) **(Weeks 5-6)**

Coursework

- The module assessment comprises of two pieces of coursework:
  - Individual exercises (worth 50%)
  - Group project (worth 50%)
- Individual exercises contain questions similar to the practical sheets and will cover Topics 1-2.
- The project will use the theory from Topic 3 as well as some independent learning (i.e. stuff I haven't taught you).
- The project will ask you as a group to do investigate models for a particular data set and write a report on your findings.

Coursework

- Individual exercises
  - Release date: Friday 10th February 2023
  - Deadline: Friday 3rd March 2023
- Group project
  - Release date: Friday 17th February 2023
  - Deadline: Friday 17th March 2023
- Deadlines to be confirmed.

Office hour

- I will not be running a formal office hour for the course but please email me (m.l.thomas@exeter.ac.uk) should you have any questions or issues with the course and the material.
- I will try my best to answer over email or but if there is a desire I will set up a short meeting to address questions individually.
- You can also post questions into the Module forum (which can be done anonymously) and I aim to answer those ASAP.
- I also encourage you all to help out your peers should you know the answer to the questions.
- Any announcements will be made on the ELE page so make sure you are subscribed.

Feedback

- I'm very open to receiving feedback during the course of the module, and will do my best to accommodate requests and fix issues.
- You can provide this either by email, or by using the course forum on ELE (noting you can post anonymously on that).
- Happy to take any questions now!

What is Statistical Data Modelling?

- Statistical data modelling (SDM) is a subset of Data Science.
- In general terms, we have some data $y_i, \ldots, y_n$ that we want to explain/model.
- Data modelling means we search for a model F that describes the data,

$$y_i = F(x_i)$$

  where x_i are some inputs/covariates/predictors.
- This invariably involves some mathematics.
- The hope is that F can be trained by looking at the data and the inputs.
- We can then use this to predict new data and/or understand relationships in existing data.
- The model F can be quite complicated.
  - Machine learning: deep neural networks, regression forests.

What is Statistical Data Modelling?

Modeling the distributions insted of data

- So where does SDM fall into all of this?
- SDM is fairly unique in this context as we a-priori assume that the data is random, they can not be deterministic.
- Instear we no longer have $y_i = F(x_i)$ but $Y_i \sim p(x_i)$, i.e. the data arises from some probability distribution p, it does not equal anything.
- A probability distribution is a mathematical function which we will determine the likelihood of our data.
- What we need to find is a p that generates the data and how this changes under different inputs.
- For example, the distribution of temperatures by season.
- This is what this course will do.