STA2201H Methods of Applied Statistics II

Monica Alexander

Week 1: Introduction

Overview

- Introductions
- What is applied statistics
- Course outline and goals
- Motivating example
- Reproducible research
- Tools
- Lab: Intro to git, tidyverse, RMarkdown

Contact

- Email: monicaalexander@utoronto.ca.
  - I do not check/answer emails after 5pm or on weekends.
- Office: SS 6010.

Hello!

Me:

- statistics $\cap$ chemistry $\rightarrow$ social science $\cap$ statistics
- 50/50 Statistical Sciences and Sociology departments
- Not Canadian (Australia $\rightarrow$ USA $\rightarrow$ Canada)

What I work on: a mix of demography, applied stats, epidemiology and computational social science

What is demography?

Demography is the scientific study of population dynamics. We are interested the size, composition and distribution of populations over time, and study these changes with respect to the three main population processes:

- Births (Fertility)
- Deaths (Mortality)
- Migration

Statistical methods become important for estimation because:

- often have no/bad data
- often dealing with survey data/lots of measurement errors
- demographic events as stochastic processes

What is applied statistics? / how does it relate to data science?

What is applied statistics?

Using statistical methods to answer questions and draw reasonable conclusions from data that have uncertainty and randomness.

Emphasis is on **data**

- you need to understand your data in order to make decent inferences
- data generating process, measurement errors, correlations, dependence. . .
- as statisticians we often don't collect the data so easy to forget this
- EDA, data visualization
- By definition applied to some other area we may or may not be (probably not) trained in: need to be aware of substantive topic and issues

How does it relate to data science?

Data science is:

- "... a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data" (Wikipedia)
- Being able "to understand [data], to process it, to extract value from it, to visualize it, to communicate it." (Hal Varian)
- "... the ability to extract knowledge and insights from large and complex data sets." (DJ Patil)
- Statistics + data munging + data viz (Nathan Yau)

How does it relate to data science?

My take: data science is a mix of CS, stats and ML. An emphasis on the whole pipeline:

1. data collection/extraction
2. data storage/maintenance
3. data manipulation/processing
4. **data analysis** (applied statistics, ML)
5. communicate output (often predictions)

Notes:

- The applied stats part may only be a small part and may be automated (this is doable if focus is on prediction)
- Relatively easy to rebrand as a data scientist if that's your jam
- Reproducibility is important

Course overview and goals

Course overview

- Topics will include generalized linear models, Bayesian inference, generalized linear mixed models, generalized additive models involving non-parametric smoothing, model evaluation and selection. We will also cover some core statistical computing techniques.
- A large focus of the outcomes on this course will also be on reproducible research, identifying and dealing with data and modeling issues, and model interpretation and communication.
- Throughout the course we will be using R in all examples, labs and homework assignments. Exams will also require interpretation of R output.
- Please bring your laptop to class! Each week we will have a lecture then a lab.

Assessment

- Lab exercises, 1% per week
- Four assignments, 12.5% each
- Final exam, 40%

Goals

- Learn a useful suite of statistical techniques
- Be able to deal with real data
- Assess data and model issues
- Establish/streamline/improve project workflow
- Pass the comp!

Expectations

- Understand main ideas behind important techniques for applied statistics
- Coding in R
- R markdown
- Git
- Code readability
- Clear communication of methods, findings, limitations
  - Data exploration is part of this!
- Aim for reproducible research

Roadmap

Roadmap

Subject to change depending on time and priorities.

Planned lecture content:

- Generalized linear models
- Bayesian inference
- Multilevel models
- Temporal models
- Non-linear/ non-parametric models (splines)
- Sampling issues, measurement issues, data issues
- Model checking, selection

Roadmap

Planned lab content:

- Rmarkdown, git
- Tidyverse
- EDA, data viz
- RShiny
- `glm`
- Stan, `brms`
- Extracting data from API (e.g. Facebook)
- RStudio Cloud, AWS

Motivating example

Global estimation of the causes of maternal death

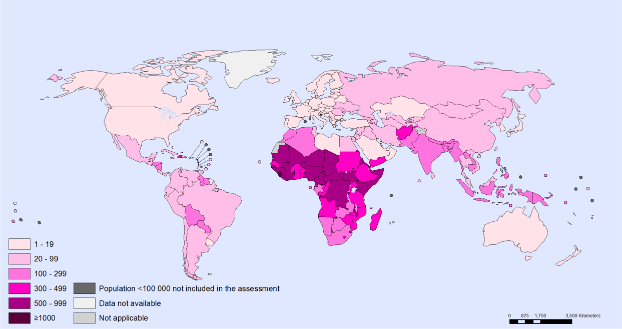
- **Maternal mortality**: the death of a woman while pregnant or within 42 days of termination of pregnancy, from any cause related to or aggravated by the pregnancy.
- Very important indicator of health and development of a country
- Part of the Sustainable Development Goals (3.1)

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS

Global estimation of the causes of maternal death

- Large variation in maternal mortality ratio (deaths per 100,000 births) across the world (highest: 1150; lowest: 2)
- In order to reduce number of deaths, need to know underlying causes
- But this is difficult information to obtain/estimate

Maternal mortality ratio (per 100 000 live births), 2015


1 - 19
20 - 99
100 - 299
300 - 499
500 - 999
≥1000
Population <100 000 not included in the assessment
Data not available
Not applicable
0 875 1,750 3,500 Kilometers

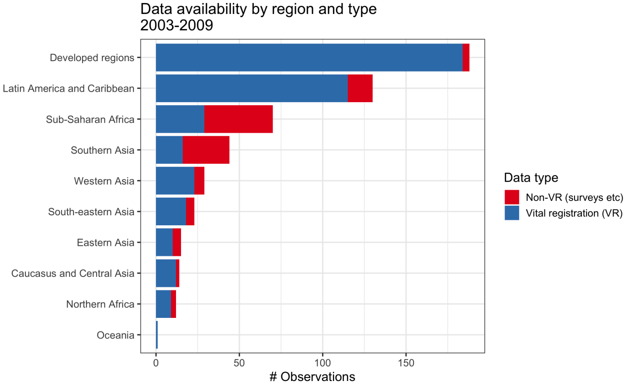
The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Data Source: World Health Organization
Map Production: Health Statistics and
Information Systems (HSI)
World Health Organization

World Health Organization

How do we get information on causes of (maternal) death?

- In high-income countries and some middle-income countries: civil registration systems
- In low-income countries: ???
  - surveys (why is this hard?)
  - facility-based administrative data
  - other specialized studies



How do we get information on causes of (maternal) death?

- If we had complete coverage of all deaths and a reliable way of classifying cause of death, then we could just count deaths and call it a day
- But in most countries (particularly high-burden countries) we have very little information, and what we do have is full of problems
- → Use statistical methods to obtain as reliable estimates as possible

Issues

To name a few:

- Years with no data
- Only some causes observed (even in high-income countries)
- Non-representative data (subnational, facility-based)
- Cause of death classification issues (death not witnessed, definition changes, differences across countries etc)
- Under/over-reporting (especially abortion)
- Not all civil registration systems are high quality
- Low death counts ($\sim$ 25 deaths in Australia)

Intro to statistical set-up

Notation:

- observations $i = 1, \ldots, n$
- d_i is total number of maternal deaths for the ith observation
- observed maternal deaths $\mathbf{y_i} = (y_{i,1}, \ldots, y_{i,7})$
- $y_{i,j}$ is the number of deaths due to cause j for the ith observation
- cause groups $j = 1, \ldots, 7$ corresponding to $\{$ABO, EMB, HEM, SEP, DIR, IND, HYP$\}$

Intro to statistical set-up

Think of deaths as a stochastic process:

- Given total number of maternal deaths d_i, the probability of a death is due to cause j is $p_{i,j}$ This is a Multinomial distribution, with 7 categories:

$$\mathbf{y_i} \sim \text{Multinomial}(d_i, \mathbf{p_i})$$
$$\mathbf{p_i} = (p_{i,1}, \dots, p_{i,7})$$

- We observe $y_{i,j}$ and d_i
- We are interested in estimating $\mathbf{p_i}$. These will help us get estimates for the 'true' proportions $\mathbf{p_c}$ for countries $c = 1, \dots, 193$ (UN member countries)

Intro to statistical set-up

$$\mathbf{y_i} \sim \text{Multinomial}(d_i, \mathbf{p_i})$$
$$\mathbf{p_i} = (p_{i,1}, \ldots, p_{i,7})$$

Put a model on $\mathbf{p_i}$:

- Transform to ensure probabilities sum to 1
- Model can include effects/adjustments for different things e.g. region, data quality, temporal changes, subnational adjustments. . .
- This is a (Bayesian) hierarchical model. We will learn about these!

Maternal mortality summary

- Real world problem, working with WHO and statisticians, epidemiologists, clinicians, public health officials
- So many data problems
- Data complexities lead to relatively complex models
- Substantive area knowledge helps to understand data issues
- Results have big impact (policy, $$$): need to be careful, transparent with assumptions, reproducible

Reproducible research

What is reproducibility?

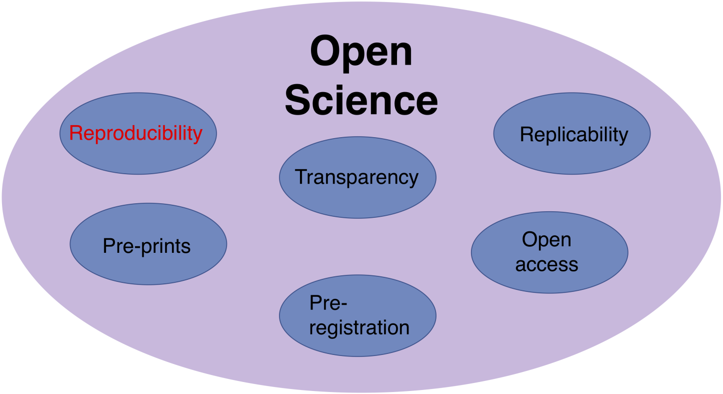
- Research is **reproducible** if it can be reproduced exactly, given all the materials used in the study
  - Note that materials need to be provided!
  - For us, ‘materials’ usually means data, code and software
  - Reproduce the data, methods and results (including figures, tables)
- Another person should be able to take the exact same data, run the exact same analysis, and produce the exact same results
- Different to **replicability**
  - carrying out a new study based on the description of the data and method provided in the original publication, and obtaining results that are similar enough.

Increased awareness recently

‘Replication crisis’, starting in Psychology, now extended to other fields

- Famous results called into question (e.g. Marshmallow test, power poses, ego depletion)
- Issues range from weaker evidence than originally thought, to fabrication of data

Just one part of doing open science



Tools

Tools

- R
- Tidyverse
- RMarkdown
- git

R

We will be using R in this course. Pros:

- Free
  - reproducibility
  - portability
- Open
  - large community
  - lots of pacakages
  - lots of help

RStudio:

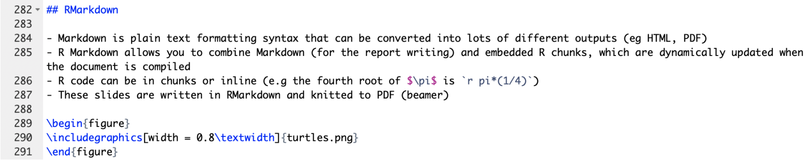
- IDE for R that makes using R a lot nicer and easier
- If you haven't already got it, download the free version here: https://rstudio.com/products/rstudio/download/
- In the labs I will be using RStudio Cloud. You should be able to access all the code there.

Tidyverse

- R Packages contain R functions, the documentation that describes how to use them, and sample data.
- The ‘tidyverse’ is “an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.” https://www.tidyverse.org/
- `ggplot` probably the most well known
- Style of coding funadmentally different to base R.
- A lot of other packages now produce output objects in the ‘tidy’ form

RMarkdown

- Markdown is plain text formatting syntax that can be converted into lots of different outputs (eg HTML, PDF)
- R Markdown allows you to combine Markdown (for the report writing) and embedded R chunks, which are dynamically updated when the document is compiled
- R code can be in chunks or inline (e.g the fourth root of π is 0.7853982)
- These slides are written in RMarkdown and knitted to PDF (beamer)



```
## RMarkdown

- Markdown is plain text formatting syntax that can be converted into lots of different outputs (eg HTML, PDF)
- R Markdown allows you to combine Markdown (for the report writing) and embedded R chunks, which are dynamically updated when the document is compiled
- R code can be in chunks or inline (e.g the fourth root of $\pi$ is `r pi*(1/4)`)
- These slides are written in RMarkdown and knitted to PDF (beamer)

\begin{figure}
\includegraphics[width = 0.8\textwidth]{turtles.png}
\end{figure}
```

RMarkdown

- Good reproducibility tool
- Can do most things you can do in LaTeX (writing math is the same)
- You are expected to write up assignments in RMarkdown

git

- git is a version control system (think a more complicated Dropbox)
- Designed for software engineers, but useful for all sorts of code
- Useful for both collaborative and solo projects
- GitHub is useful place to host open source projects

Summary

Summary

- Applied statistics has a focus on **data**
    - Understanding where the data come from, generating process, issues, etc
- The implication is that we need to think carefully about model assumptions and whether they make sense
- Reproducibility is important, especially if we want our research to be useful

Lab

To-dos

- Install R/RStudio if you haven't already
- Get GitHub account if you haven't already
- Get RStudio Cloud account if you haven't already