STA2201H Methods of Applied Statistics II

Monica Alexander

Week 1: Introduction

Overview

- Introductions
- Course outline and goals
- Tools
- GLMs review
- Lab: Intro to git, tidyverse, RMarkdown

Introductions

Instructor

- Monica Alexander
- Email: monicaalexander@utoronto.ca.
- Office hours time TBC.

TA

- Michael Chong
- Email: myc.chong@mail.utoronto.ca.

We do not check/answer emails after 5pm or on weekends!

Course outline and goals

Course outline

- 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.
- The focus in terms of methods is advanced regression techniques, fit using Bayesian inference. The focus in terms of coding/computation is becoming more comfortable and adept at efficient, reproducible coding and workflows (data, analysis, reporting and communicating results)

Course outline

- Throughout the course we will be using R in all examples, labs and homework assignments.
- Each week will be a lecture ($\sim$1-1.5hrs) then a lab
- The first three lectures are definitely online, after that, ?????

We're all out here doing our best

- The current situation makes both learning and teaching challenging
- Try to be understanding of everyone's sub-optimal situation
- Communication is key
- There may be guest appearances from my children

Assessment

- Lab exercises, 8 in total, 2.5% each
  - Due 9am the following Monday
  - Hand in via git
  - Practice of concepts covered in the lecture
- Three assignments, 15% each
  - Mostly data analysis, very R heavy
  - Hand in via Quercus
- Research project 35%
  - Pick a dataset, research question and statistical approach (that is covered in class)
  - Research proposal (7.5%)
  - Research paper (20 %)
  - Presentation last week of class (7.5%)

Expectations

We will be doing applied statistics in the truest sense of the term

- Understand main ideas behind important techniques for applied statistics
- Coding in R (and in particular, the tidyverse, ggplot)
- Dealing with real data!
- R markdown
- Git (terminal or desktop, not direct file upload)
- Code readability
- Clear communication of methods, findings, limitations
  - Data exploration is part of this!
- Aim for reproducible research

Research project

- The goal is to write a short applied statistics paper in the academic style
- Intro/background (some reference to previous literature/work!), data, methods, results, discussion, limitation
- Increased length does not equal increased quality
- (Increased number of graphs does not equal increased quality of EDA)
- Should be written in RMarkdown (ideally self-contained, but if data/code too big/slow, rmd should call scripts in a reproducible way)

Research project tips

- Start thinking about it now?!
- We will regression techniques to deal with
  - a range of outcomes (continuous, binary, categorical, counts)
  - nested groups (e.g. individuals within schools within districts within provinces)
  - non-representative surveys
  - time series (can have missing data!)
- Think of a question –> find a dataset –> if you can't find a dataset then maybe change your question :)
- Must be different to AS1 projects

Course roadmap

Subject to change depending on time and priorities.

Planned lecture content:

- Generalized linear models recap
- Bayesian inference
- Visualizing the Bayesian workflow and model checks
- Multilevel models
- Non-linear/ non-parametric models (splines)
- Temporal models / dealing with correlation
- Time/interest permitting: text analysis?

Roadmap

Planned lab content:

- Rmarkdown, git
- Tidyverse
- EDA, data viz
- RShiny
- Stan, `brms`
- Probably: web scraping
- Maybe: Extracting data from API (e.g. Facebook or Twitter), AWS

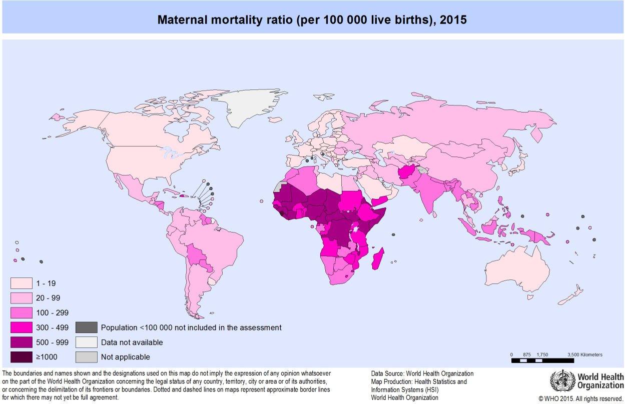
Motivating example

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)

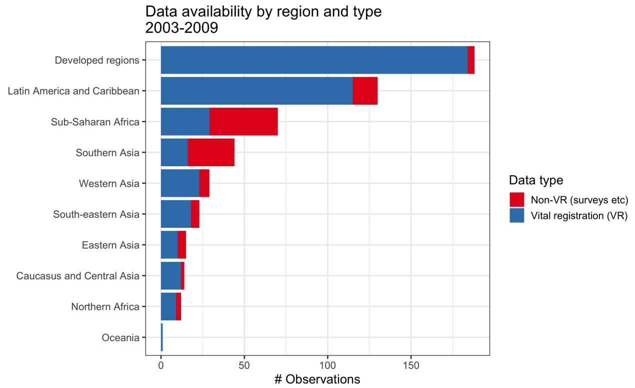
Global estimation of the causes of maternal death

- 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



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
- $\rightarrow$ 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:

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

- 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 $c = 1, \ldots, 193$ (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!

More info, see paper: https://arxiv.org/abs/2101.05240

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

Tools

Tools

- R
- Tidyverse
- RMarkdown
- git

R

We will be using R in this course. Pros:

- Free
  - reproducibility
  - portability
- Open
  - large community
  - lots of packages
  - 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/

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 fundamentally 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

GLMs recap

Motivating examples

Outcomes we may be interested in investigating (in relation to other explanatory variables):

- Police stop and frisks in NYC
- Infant deaths in the US
- Who voted for the Liberal party v other party
- Who voted Liberal, Conservatives, LDP
- Concentration of drug at particular times after ingestion

The take-away: none of these are Normal.

General linear models

Let's start with a recap of general linear models. We observe $y_1, y_2, \ldots y_n$ which are realizations of the random variables $Y_1, Y_2, \ldots, Y_n$

In linear models the y_i's have two pieces:

1. A **systematic part**, with the form

$$E(\mathbf{Y}|\mathbf{X}) = \mu = \mathbf{X}\beta$$

2. A **random part**, where errors are assumed to be i.i.d such that $E[\epsilon] = 0$ and $var[\epsilon] = \sigma^2$. We usually further assume that errors are Normal with constant variance σ^2.

Multiple linear regression

One of the most common examples of a general linear model.

Goal: we are trying to measure the association between response/outcome/dependent variable Y_i and one or more explanatory variables/covariates $X_{i,1}, X_{i,2}, \ldots, X_{i,k}$

- The conditional expectation function (CEF) $E(Y_i|X_{i,1}, X_{i,2}, \ldots, X_{i,k})$ describes the expected value (population mean) of Y_i given values of the variables $X_{i,1}, X_{i,2}, \ldots, X_{i,k}$.

Multiple linear regression

MLR is a model for the CEF:

$$\begin{aligned} Y_i &= E\left(Y_i \mid X_{i,1}, X_{i,2}, \ldots, X_{i,k}\right) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \end{aligned}$$

Specifically, the most basic MLR model is a simple linear function of the X's and associated parameters β.

Estimation

Minimizing the sum of squared residuals

$$S(\beta) = \sum_{i=1}^{n} \left(y_i - x_i^{\mathrm{T}} \beta \right)^2 = (y - X\beta)^{\mathrm{T}} (y - X\beta)$$

Leads to the MLR-OLS estimator

$$\hat{\beta} = \left(X^{\mathrm{T}} X \right)^{-1} X^{\mathrm{T}} y$$

Sampling distribution of the MLR-OLS estimator

- Under the Gauss Markov and normality assumptions, the OLS estimator, $\hat{\beta}_k$ is normally distributed with a mean equal to

$$E\left(\hat{\beta}_k\right) = \beta_k$$

and variance

$$\text{Var}\left(\hat{\beta}_k\right) = \frac{\sigma^2}{\sum_i \left(X_{ik} - \bar{X}_{ik}\right)^2 (1 - R_k^2)}$$

We can use this property for inference: The sampling distribution of standard error standardized estimator follows a t-distribution with $n - (k+1)$ degrees of freedom.

General linear models

$$
\begin{aligned}
Y_i &\sim N(\mu_i, \sigma^2) \\
\mu_i &= \mathbf{X}_i^T \beta
\end{aligned}
$$

General linear models are not appropriate when

- The range of Y is restricted
- The variance of Y depends on the mean

Generalized Linear Models extend the classical set-up to allow for a wider range of distributions. Introduced by Nelder and Wedderburn (1972) [Later, GAMs in 1990].

Generalized linear models

Generalized linear models

GLMs have an additional piece on top of the classical linear models:

1. **random component**: $Y_i \sim$ some distribution with $E[Y_i|\mathbf{X}_i] = \mu_i$
2. **systematic component**: $\mathbf{X}_i^T\beta$
3. The **link function** that links the random and systematic components $g(u_i) = \mathbf{X}_i^T\beta$

- Set-up is almost the same, particularly in terms of specifying a good linear predictor $\mathbf{X}_i^T\beta$
- Just need to think about the link and the distribution of the outcome

GLMs

$$
\begin{aligned}
Y_i &\sim G(\mu_i, \phi) \\
E[Y_i|\mathbf{X}_i] &= \mu_i \\
g(\mu_i) &= \mathbf{X}_i^T \beta
\end{aligned}
$$

- ϕ is the scale parameter.

What can Y be distributed as? In principle, anything. In practice (and original formulation), distributions come from the **exponential family**.

Exponential Family

Exponential Family

The random variable Y belongs to the exponential family of distributions if its support does not depend upon any unknown parameters and its density or probability mass function takes the form

$$p(y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

- $\theta = h(\mu)$ depends on the expected value of y and is the **canonical parameter**
- ϕ is the scale parameter (if known: one-parameter family)
- b and c are arbitrary functions

Example: Poisson distribution

$$p(y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

Poisson:

$$p(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}$$

Write as

$$p(y|\mu) = \exp\{y \log \mu - \mu - \log y!\}$$

- $\theta = \log \mu$
- $b(\theta) = e^\theta$
- $c(y, \phi) = -\log y!$
- Note that the scale parameter $\phi = 1$ so the variance is entirely determined by the mean

Example: Normal distribution

$$p(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

Normal:

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

Write as

$$p(y|\mu,\sigma^2) = \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right\}$$

- $\theta = \mu$
- $b(\theta) = \frac{1}{2}\theta^2$
- $\phi = \sigma^2$
- $c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]$

Other examples

Other common examples:

- Binomial
- Gamma
- Negative binomial
- Inverse Gaussian

Propertites of exponential families

Mean and variance for exponential families

$$p(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

It can be shown that

$$E(Y|\theta,\phi) = b'(\theta) = \mu$$

and

$$Var(Y|\theta,\phi) = \phi b''(\theta) = \phi V(\mu)$$

Note the variance of Y depends not only on the scale parameter but also on a function of the mean.

Examples:

$$E(Y|\theta,\phi) = b'(\theta)$$

and

$$Var(Y|\theta,\phi) = \phi b''(\theta)$$

- Poisson: $E(Y|\theta,\phi) = e^{\theta} = \mu$, $Var(Y|\theta,\phi) = 1 \times e^{\theta} = \mu$
- Normal: $E(Y|\theta,\phi) = \theta = \mu$, $Var(Y|\theta,\phi) = \sigma^2 \times 1 = \sigma^2$

The canonical link

The link function $\eta_i = g(\mu)$ could in theory be any function linking the linear predictor to the distribution of the outcome variable, which is also is **monotonic** and **smooth**.

Recall $\theta = h(\mu)$. If we choose $g = h$, then

$$\theta_i = h(\mu_i) = h(h^{-1}(\eta_i)) = \eta_i = \mathbf{x}_i^T \beta$$

In other words, it ensures that the systematic component of our model is modeling the parameter of interest.

Canonical links

- Normal: identity $\theta = h(\mu) = \mu$
- Poisson: $\theta = h(\mu) = \log \mu$
- Binomial: $\theta = h(\mu) = \log(\frac{\mu}{1-\mu})$
- Exponential/Gamma: $\theta = h(\mu) = -\mu^{-1}$
- Inverse Gaussian: $\theta = h(\mu) = \mu^{-2}$

Likelihood-based estimation and inference

Estimation

- Inference is based on MLE, but cannot derive closed form solutions for regression coefficients

The log-likelihood function is:

$$\ell(\theta) = \sum_i \ell(\theta_i) = \sum_i \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi)$$

- Differentiate with respect to β to get the score function $\mathbf{S}(\beta)$ and then set this equal to 0
- Use Newton-Raphson to approximate the score function

Estimation

Estimator can be written in the form:

$$\widehat{\beta}^{(\mathbf{t+1})} = (\mathbf{x}^{\mathbf{T}}\mathbf{W}\mathbf{x})^{-\mathbf{1}}\mathbf{x}^{\mathbf{T}}\mathbf{W}\mathbf{z}$$

where $\mathbf{W}$ is diagonal with $w_i = (\frac{\partial \mu_i}{\partial \eta_i})^2 / \phi b''(\theta_i)$ and

$$z_i = x_i\beta + (Y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$$

- $\mathbf{W}$ and $\mathbf{z}$ change depending on $\hat{\beta}$ and vice versa
- Use iteratively weighted least squares (IWLS)
  1. Choose initial value $\hat{\beta}^{(0)}$
  2. Calculate $\mathbf{W}$ and $\mathbf{z}$
  3. Repeat until convergence

Inference

Inference

We know that for the MLE, the limiting distribution is

$$\hat{\beta} \sim N(\beta, (\mathbf{x}^{\mathbf{T}}\mathbf{W}\mathbf{x})^{-1})$$

$$\hat{\beta} \sim N(\beta, I(\hat{\beta})^{-1})$$

Standard errors are the square roots of the inverse of the information matrix.

- Use this for the classic Wald Tests e.g. $\sqrt{W} = \frac{\hat{\beta}-\beta_0}{se(\hat{\beta})}$ follows z distribution.

Likelihood ratio test

Testing nested models, ω_1 and ω_2, $\omega_1 \in \omega_2$ and number of parameters $p_2 > p1$

$$2[\log \ell(\widehat{\beta}_{\mathbf{1}}|\mathbf{y}) - \log \ell(\widehat{\beta}_{\mathbf{2}}|\mathbf{y})] \sim \chi_{p_1 - p_2}$$

- Comparing fit of two models
- Model with more predictors will almost always fit better, but is the difference significant?

GLM in R

- `glm()`
- same set up as `lm()`; additional `family` argument with a link
- e.g. `glm(y~x, family = binomial(link = 'logit')`

Lab

Lab this week

- Setting up git, with a small exercise
- Intro to tidyverse and ggplot

Git

- Git is a version control system
- The system tracks changes you make to git repositories (‘repos’)
- Think of repos as folders
- In order for file versions to be tracked, they need to be **committed** to the git repo
- Think of committing as like saving, but with slightly more steps

GitHub

https://github.com/

- A hosting service for git repos
- You can sign up for free, and host an unlimited number of public or private repos
- You will be submitting lab exercises via GitHub, so you need to set up an account!

The simplest Git/GitHub workflow

- New repo on GitHub
- Clone onto local computer
- Do work on local computer
- Save
- Add and commit to git repo
- Push to GitHub (this means your new work will appear on the GitHub website)

If you are working on your own, on one computer, this is it!

Git/GitHub

- If you are working on a couple of different computers / servers, you may also need to **pull** from GitHub to update any new work done elsewhere
- Git is designed for collaborative work. More complicated workflows have branches, pull requests, merges (more later)

Steps on GitHub

Monica to now demonstrate

- creating a new repository
- Adding Monica and Michael as collaborators

Then using the terminal (and GitHub Desktop?):

- cloning to your computer
- doing some work
- git status
- add, commit, push

Disclaimer: I use the terminal. You are welcome to use the GitHub Desktop: https://desktop.github.com/ But do not use direct file upload.

This week's lab assessment (git part)

- Make a repo and add me as a collaborator
- Add a text file with your name, program (e.g. statistics Masters), whether or not you're currently in Toronto, and your favorite type of food
- Push changes to GitHub