SOC6707 Intermediate Data Analysis

Monica Alexander

Week 12: Summary and misc

Overview

- Summary of course
- What I hope you got out of it
- Extension: MRP
- Stuff you can do with these coding skills

Summary

Content

EDA:

- Data structures
- Descriptive statistics
- Data visualization

Regression

- Linear
- Logistic
- Multinomial
- Multilevel linear and logistic

Coding skills

R!!!!!

- Tidyverse
- ggplot
- R Markdown

What I hope you got out of it

Quantitative skills:

- Data analysis is not just regression as an end goal
- Need to get to know your data and understand what's going on there before you even start
- Focus on the whole pipeline of research, from getting data to writing a paper
- Hopefully this will help you to critically read quantitative methods papers
  - Do authors adequately describe the data?
  - Do the statistical results reconcile with what you expect based on the data?
  - Do they provide reproducibility materials? :)

What I hope you got out of it

R coding and data manipulation:

- This was hard, but you made it!
- Potentially useful even if you are not doing quant research
- How can you apply a reproducible workflow to your own research?
- If you get stuck, don't be afraid to look and ask for help

Multi-level regression and post-stratification (MRP)

Dealing with non-representative surveys

- We generally want to use responses from surveys to form estimates of groups of interest (e.g. national, state-level opinions)
- To do this we need to ensure that the characteristics of the people surveyed are similar to the group of interest
- But getting representative survey responses is expensive
- Even if you have a good sampling frame, not guaranteed you will get a representative set of responses (people don't have phones, or don't answer them)
- Often better to over-sample people of interest, and the re-weight (post-stratify) to get representative estimates

Me trying to stay relevant

We have 20 people in our class. Let's say 8 of you did undergrad at UofT (40%), and the other 12 did undergrad somewhere else.

Say I was interested in the the proportion of graduate students at UofT that use TikTok.

- I did a survey of our class, and out of 20 people, 10 people use TikTok and 10 do not.
- Of the people who did undergrad at UofT, 7 people use TikTok, and of those who didn't, 3 people use TikTok.

Based on our class survey, I could conclude that $10/20 = 50\%$ of graduate students use TikTok.

Post-stratification

- But say we knew that of all UofT grad students, 25% actually did undergrad at UofT.
- This is much lower than the proportion in our class
- A better estimate based on our survey, then, could be to post-stratify based on undergrad institution
- So our estimate of Pr(TikTok) $= 7/8 * 0.25 + 3/12 * 0.75 =$ 41%

Post-stratify based on more characteristics

- It might make even more sense to post-stratify on other characteristics, like gender, age, undergraduate degree
- These are characteristics we might expect to be associated with TikTok usage
- But as we choose more post-stratifying variables, the cell count (i.e. the number of people in each group) gets smaller
- So our estimates become more uncertain

A more robust approach

Instead of taking raw counts by group, we could model the probability of using TikTok (y_i) in a hierarchical (multi-level regression), with covariates such as age, gender, undergrad degree, e.g.

$$\text{logit}^{-1}(Pr(y_i = 1)) = \beta_0 + \beta_1 \text{gender}_i + \beta_2 \text{age}_i + \beta_3 \text{degree}_i + \beta_4 \text{institution}_i$$

with $\beta_2 \sim (0, \sigma^2_{\text{age}})$ and $\beta_3 \sim (0, \sigma^2_{\text{degree}})$ i.e. the effects of age and degree are modeled hierarchically (too few groups with gender and institution).

- Why the hierarchical/multi-level set-up? Remember from radon, etc, that estimates for groups with small counts get shrunk toward the global mean, effectively placing less weight on the outcomes for groups where we have less information

MRP

$$\text{logit}^{-1}(Pr(y_i = 1)) = \beta_0 + \beta_1 \text{gender}_i + \beta_2 \text{age}_i + \beta_3 \text{degree}_i + \beta_4 \text{institution}_i$$

- Our model gives us **predicted** probabilities for TikTok usage by group.
- Once we have those, we can post-stratify as before to get a population-level (UofT-wide) estimate (and uncertainty)
- So the difference is we are using modeled proportions rather than raw proportions from the data
- Note that you must have info on post-stratification counts (cross-tabulated)! So in this example, need to know UofT counts by age, gender, undergrad degree, undergrad institution
- MRP often used to predict voting behavior, with post-stratification info coming from the census

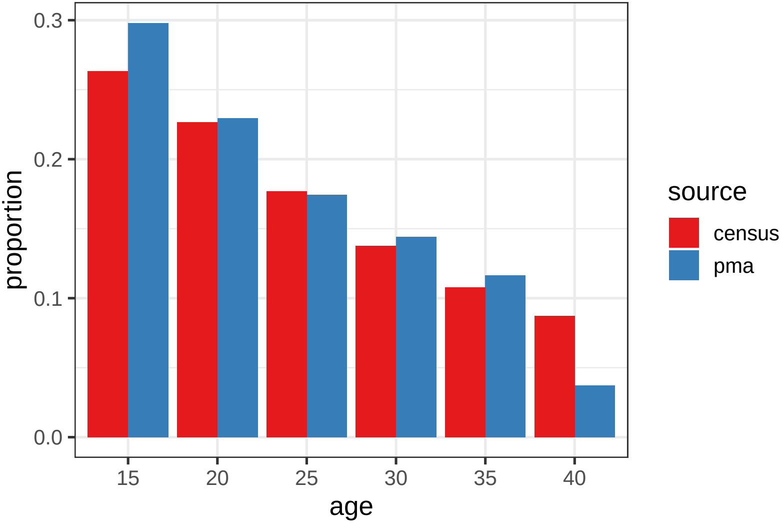
Example: abortion outcomes in Uganda

- Data from 2018 PMA survey (via IPUMS)
- Interested in factors associated with women ever having an abortion
- Outcome of interest: ‘ever had abortion (yes/no)’
- Notes: dropping don’t knows, including ‘unsuccessful abortions’ in ‘yes’.
- More notes: Self-reported abortion is very likely to be under-reported

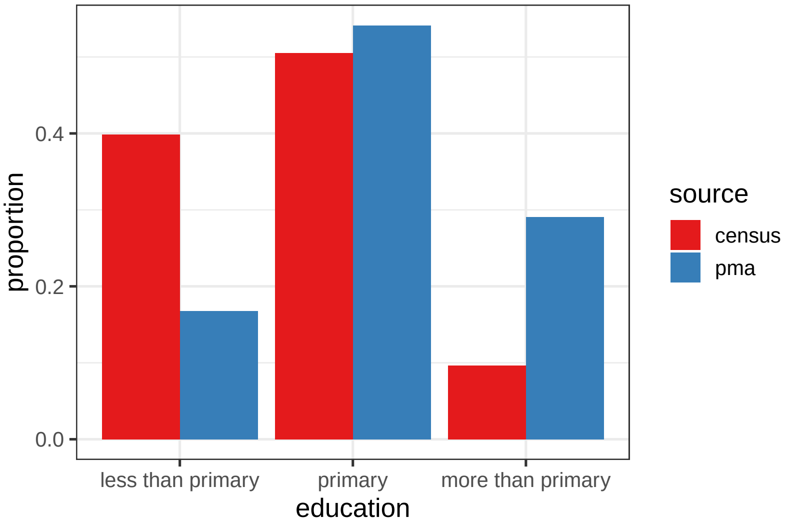
Sample versus census

How representative is the survey? Let's compare to 2014 census.

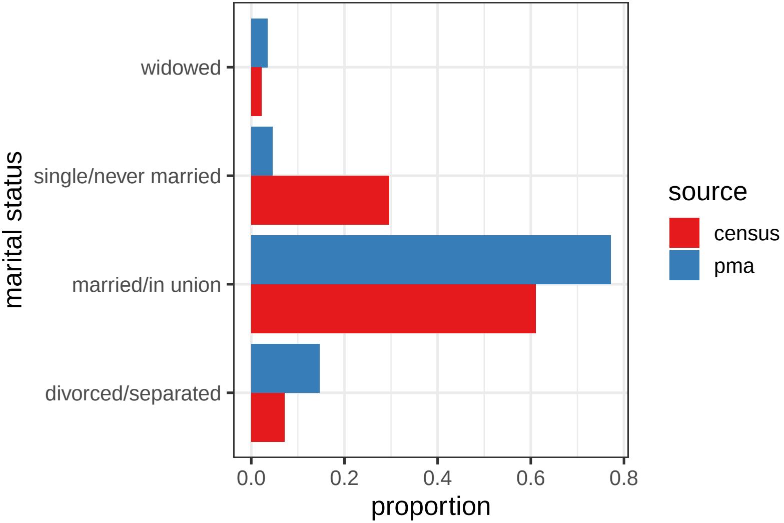
Proportion by age



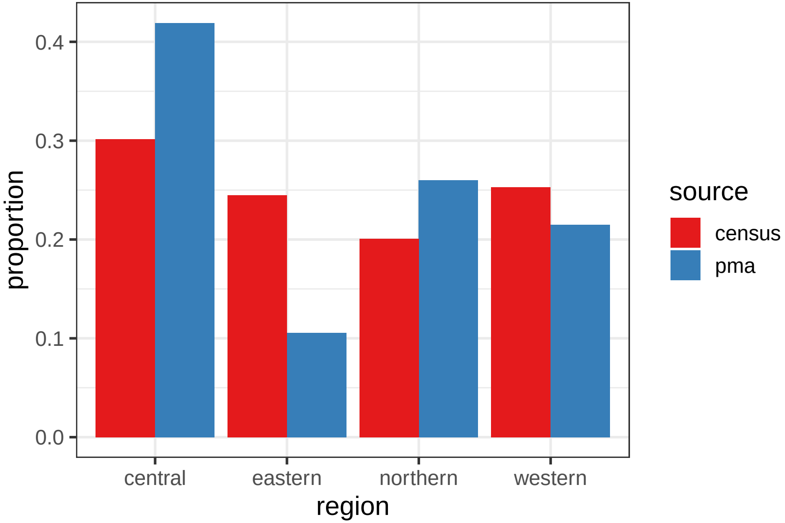
Proportion by education



Proportion by marital status



Proportion by region



Multilevel model

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit} p_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \alpha_{j[i]} + \gamma_{k[i]}$$

$$\alpha_j \sim N(0, \sigma_\alpha^2)$$

$$\gamma_k \sim N(0, \sigma_\gamma^2)$$

where

- Y_i refers to whether or not individual i has had an abortion
- The index j refers to region
- The index k refers to age group
- X_1 refers to marital status
- X_2 refers to education

Predicted probabilities

- Our model gives us a way of calculating the predicted probability of reporting abortion for a women in each age/educ/region/marital status group, call it $\hat{p}_g$ for groups $1, \dots G$ where $G = 6 \times 3 \times 4 \times 4$
- We can use these predicted probabilities to simulate individual outcomes based on the Bernoulli likelihood
- This gives us an estimate of the number of women in each group reporting abortion, with uncertainty

Post stratification

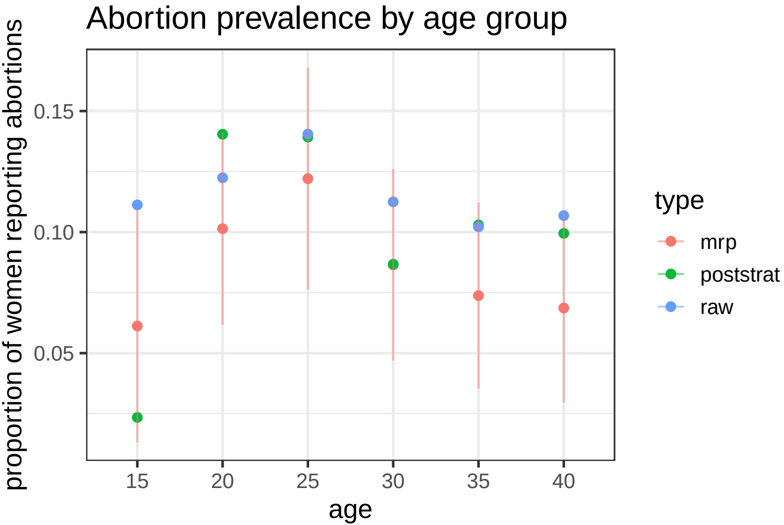
- Once we have an estimate of the probability of reporting abortion in each age/educ/region/marital status group, we can multiply these by the total number of women in each group based on the census, to get an estimate of the number of women reporting abortion in the population
- Can then use these estimates to calculate abortion incidence nationally and by different groupings
- This is called post-stratification
- E.g. national estimate of abortion incidence would be

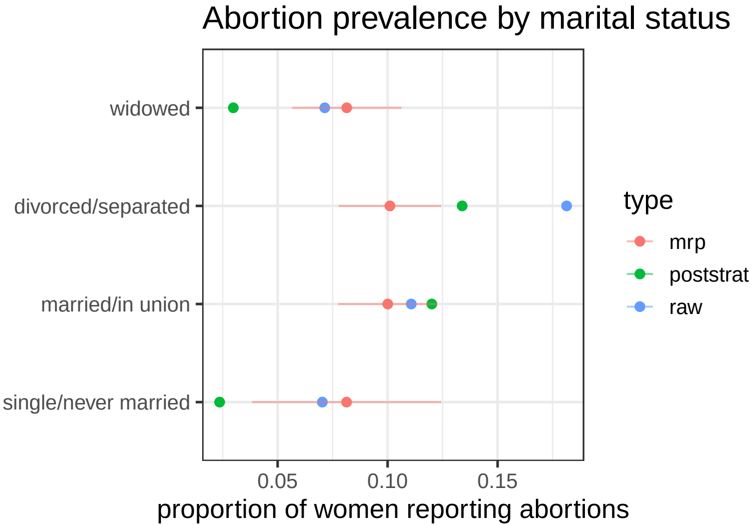
$$\frac{\sum_g \hat{p}_g \cdot N_g}{\sum_g N_g}$$

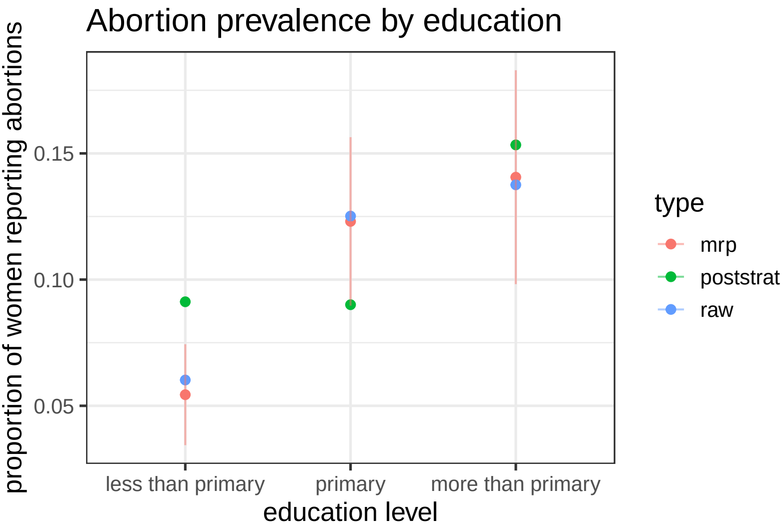
MRP versus other options

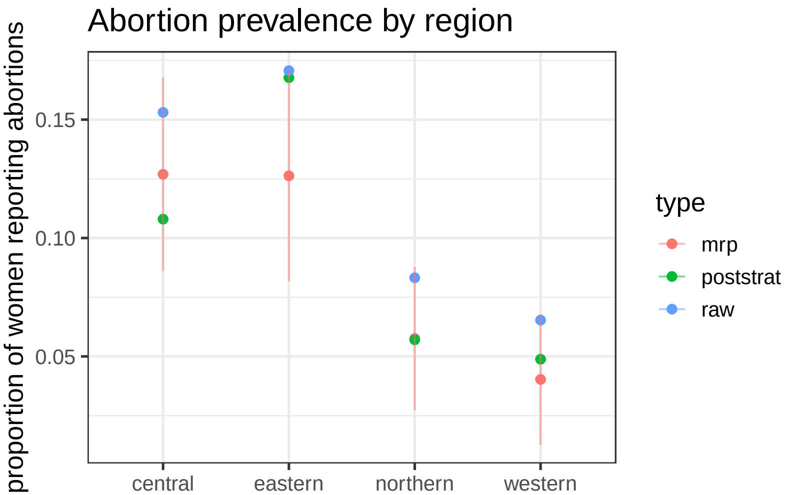
- This is called multilevel regression and post-stratification because we do just that – run a ML regression then post-stratify based on a representative data source
- Other options:
    1. just take raw incidence proportions from survey
    2. calculate raw incidence proportions from survey then post-stratify these
- Compared to 1, 2 and MRP are more representative
- Compared to 2, MRP downweights raw proportion estimates from groups that have small sample sizes (but potentially at the expense of increased variance in estimates)

Results


Abortion prevalence by age group
proportion of women reporting abortions
0.15
0.10
0.05
15
20
25
30
35
40
age
type
mrp
poststrat
raw


Abortion prevalence by marital status
widowed
divorced/separated
married/in union
single/never married
0.05
0.10
0.15
proportion of women reporting abortions
type
mrp
poststrat
raw


Abortion prevalence by education
proportion of women reporting abortions
0.15
0.10
0.05
less than primary
primary
more than primary
education level
type
mrp
poststrat
raw


Abortion prevalence by region
proportion of women reporting abortions
0.15
0.10
0.05
central
eastern
northern
western
type
mrp
poststrat
raw

MRP: further reading

- A famous paper, using surveys of Xbox users: Wang et al., “Forecasting elections with non-representative polls”
- Here’s a worked example, where I used data from a survey about changing name after marriage: https://www.monicaalexander.com/posts/2019-08-07-mrp/

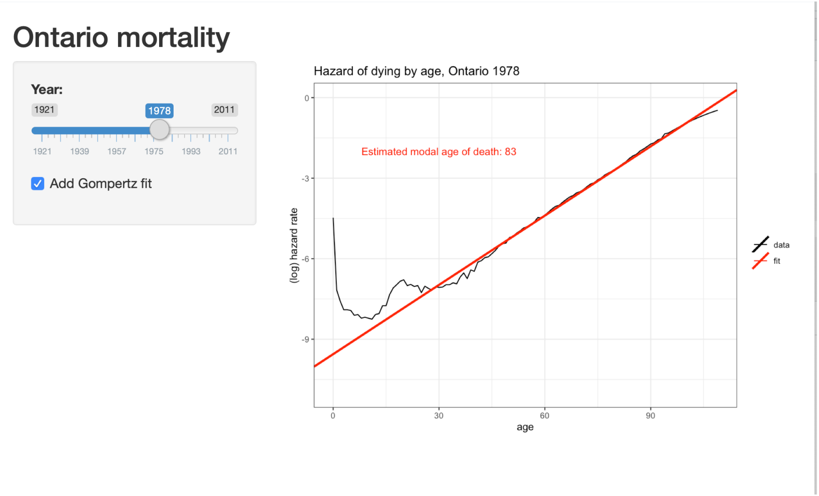
More things you can do with these coding skills

R Shiny

- An R package to create interactive web-based applications to visualize results
- Essentially inserts your R code (ggplot or otherwise) into functions that create an interactive interface so the user can change inputs based on a widget (slider, dropdown, etc)
- Can then host on the web for free with `shinyapps.io` or your own server

A simple example

https://monica-alexander.shinyapps.io/example_shiny/



Ontario mortality

Year: 1978

☑ Add Gompertz fit

more examples: https://shiny.rstudio.com/

A simple example

```r
library(tidyverse)
d <- read_rds("data/ON_mortality.RDS")

# Define a server for the Shiny app
function(input, output) {

  # Fill in the spot we created for a plot
  output$hazardPlot <- renderPlot({

    p <- d %>%
      mutate(age = as.numeric(age)) %>%
      filter(year==input$year) %>%
      ggplot(aes(age, log(hx))) +
      geom_line(aes(color = "data")) +
      #scale_y_log10() +
      theme_bw() +
      ylab("(log) hazard rate") +
      ggtitle(paste0("Hazard of dying by age, Ontario ", input$year )) +
      scale_color_manual(name = "", values = c("data" = "black", "fit" = "red")) +
      ylim(c(-11,0))

      if(input$addGompertz==FALSE){
        p
      }

    ...

  })
}
```

A simple example

```r
library(tidyverse)
d <- read_rds("data/ON_mortality.RDS")

# Use a fluid Bootstrap layout
fluidPage(

  # Give the page a title
  titlePanel("Ontario mortality"),

  sidebarLayout(

    sidebarPanel(
      sliderInput("year",
                  "Year:",
                  value = 1960,
                  min = min(d$year),
                  max = max(d$year), sep = ""),
      checkboxInput("addGompertz", "Add Gompertz fit", FALSE)
      ),

    # Create a spot for the plot
    mainPanel(
      plotOutput("hazardPlot")
    )

  )
)
```

Other examples

- Foster care: https://monica-alexander.shinyapps.io/foster_care/
- Baby names: https://monica-alexander.shinyapps.io/babynames_app/
- Built-in explorer in R package: https://github.com/jessieyeung/rcbayes

Blogdown

Unsolicited advice

- Many things that are useful after grad school are not taught in grad school
- Dissertation is only one bit
- Broad versus specific knowledge
- “Admin” stuff: Grant writing, teaching, mentoring
- Build your ‘brand’
    - What’s your niche?
    - What’s your narrative?
    - 1/5/15min spiels of research
    - Make a website!

Websites with blogdown

- Consider making a website, if you don't have one already!
- If you are on the job market (academic or otherwise) people will Google you. It's a useful way to partially control what they see.
- Even before you're on the market, good to have, to build up a profile
- Lots of good tools, but you can also make websites in R :)

Blogdown

- Blogdown is an R package that let's you create websites in RMarkdown
- Built by people at RStudio so nicely integrated
- Builds on website templates from Hugo (https://gohugo.io/)

Example wesbites built with blogdown:

- Mine: https://www.monicaalexander.com/
- Julia Silge: https://juliasilge.com/
- Sharla Gelfand: https://sharla.party/

High-level steps

Follow Alison Hill's blog:
https://alison.rbind.io/post/new-year-new-blogdown/

High-level steps

1. Create a new folder with an RStudio project. Best to make it a git repo also (e.g. my_website) because it will be easier to get online later
2. Choose a Hugo theme, hugo-academic is common one to start with. Then in Rstudio type

```
blogdown::new_site(theme = "wowchemy/starter-academic")
```

This will download a bunch of files into your folder and begin "serving" your site locally (i.e. within RStudio)

3. Add your own basic content. Some of this will be editing the `config.toml` file that got downloaded. You can also add a headshot (in the `static/img` folder).

High-level steps

4. Add more detailed content.

- If you look at the the `content` folder, for hugo-academic there are some markdown (.md, similar to .Rmd) files called things like `about.md`, `publications.md` etc. Can edit as neccessary.
- If you want to add blog posts written in RMarkdown, you can add them in the `content/post` folder.
- This step will invole a lot of playing around and editing to get things how you want. There's lots of help online, and I've included some good blog posts and resources below.
- To come back your website once you've closed R Studio, open the RStudio project, then type 'blogdown:::serve_site()' into the console to serve your site and then continue editing.

5. Make your website public

- Commit and push to GitHub. Then two options: deploy using Netlify or GitHub Pages

Blogdown: further resources

Lots of good resources out there, here's a selection

- https://bookdown.org/yihui/blogdown/
- https://alison.rbind.io/post/2017-06-12-up-and-running-with-blogdown/
- https://masalmon.eu/2020/02/29/hugo-maintenance/
- https://djnavarro.net/post/starting-blogdown/

But Alison's blog is really the best to get started:
https://alison.rbind.io/post/new-year-new-blogdown/

Thank you :)