Canadian Election Forecasts and Computers

Eric Zhu, Brian Diep, Ashley (Jing Yuan) Zhang, Kristin (Xi Yu) Huang, Tanvir Hyder

All pictures taken by Eric Zhu

Roadmap

- Methodology
- Results
- Computational Talk
  - Bayesian problems
  - Lme4 issues
- Post-stratifying
- Next Steps

Methodology

Multilevel Regression with Poststratification (Mr. P)

- Mixed modelling formulation

$$y_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \ldots + U_{ij} + \epsilon$$
$$U_{ij} \sim N(0, \tau)$$

- Post-stratification

$$y_{post} = \sum_{i=1}^{D} w_i \cdot y_{pre}$$

- Fit using multinomial model - for classification task with more than 2 categories

Model Predictors

province - the province in which the individual resides

age

education - the highest education level attained by the individual

income - the household income of the individual separated into broad income brackets

home ownership - status of whether the individual is owner or tenant to their residence

language - official languages spoken by the individual, English and/or French (or neither)

immigration status - whether or not the individual themselves are naturalized citizens or born citizens

age - the individual’s age by decade bracket

Final model

- Log-linear multinomial model fit using **nnet**
  - Multinomial-GLM with a softmax link (activation)!
- Predicted outcomes using through the model:

  vote_choice ~ province + education + income_bracket + age_bracket + home_ownership + language + is_immigrant

- Then post-stratified over the various combinations of the predictors
  - ~7.7 million rows (categories) to postratify over
  - Create final predictions

Results

Accurately predicted a Liberal victory

Specific Seats:

- Election: Liberal 159, Conservative 119, Bloc Québécois 33, NDP 25, Green 2
- Prediction: Liberal 240, Conservative 86, NDP 1, Bloc Québécois 14

Reasons for difference:

- Data is not reflective of current demographics (2019)
- High bias
- Random effect
- Number of factors included in the model - lack granularity

Computational Considerations

Bayes and Brms

- Great package
  - Handles random effects
  - Many link functions - native support for multinomial regression
  - Default priors are weakly informative and provide regularization
  - Easy model diagnostics and integration with **bayesplot**
- Problems
  - **SLOW!** - 16+ hours to fit on a Ryzen 5800x (8 cores, 16 threads) @ ~4.8 GHz
  - “Too much” model complexity causes C++ compilation issues
    - Multiple random intercepts and multiple random slopes
  - ~4000 samples needed for decently independent sample

Lme4 and multinomial regression

- Also a great package
  - Objective function is penalized[1]
  - Models often fit using REML
    - REML per-iteration cost[2]: $\mathcal{O}(n^2 p)$

- Problems
  - Limited link functions - no support for softmax links
  - No good control over regularization

---

[1]: https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf
[2]: https://arxiv.org/pdf/1803.04431.pdf

Post-stratifying

Adding predictors greatly contribute to the space complexity of the post-stratification matrix

With our relatively limited set of predictors, our poststratification matrix already had over 7 million observations and it was difficult to load the matrix into memory

We did not have access to joint distribution of the population from the Canadian census data, we naïvely assumed that each of our predictors were independently distributed which may have contributed to our high bias

Next Steps

- Run model predicting using the CES 2021 survey answers
- Add omitted predictors such as sex, employment status, ethnicity, etc.
- Get some nicer computers (and packages) to accommodate the increased complexity
- Incorporate more predictors relevant to election topics (e.g. economy, COVID, housing crisis related predictors)
- Ideally we can use a different source of data for post stratification similar to the IPUMS dataset to get a more accurate joint distribution between predictors