Analyzing Census Data of Canadian Census Divisions and Using R to Ascertain the Most Impactful Confounders on Net Migration

Student: Francis Emmanuel Calingo (214451736)

Course: MATH 4330-Applied Categorical Data Analysis

Semester: Fall 2021

Instructor: Mr. Georges Monette

ABSTRACT

Despite Canada being a wealthy nation relative to much of the world, inequalities still exist here. Socioeconomic conditions are not uniform and can vary greatly from census division to census division. Which socioeconomic issue has had the most substantial effect on out-migration? Could there also be other explanations and factors that explain why out-migration is higher in some regions? This project collated different data of census divisions across Canada (2016 Census), and identified various demographic and socioeconomic variables of interest such as educational attainment and racial groups. After selecting specific variables from certain datasets, testing out different models via regression analysis, and measuring their statistical significance, average income and post-secondary education attainment rates were identified as the two most impactful variables on out-migration.This finding suggests that local governments should focus more on poverty reduction measures and increasing accessibility to post-secondary education if they seek to stabilize their population decline, if not completely reverse it.

INTRODUCTION

A simple survey of York University students (albeit, using an appropriate sample of the population) will most likely yield a highly diverse outcome. There are students who do not have to stress about student loans ever, while there are students who have to work multiple jobs in order to just barely scrape by. There are students who are only a 20-minute bus ride away from campus and there are students who would need to travel for two hours via GO

Transit (although that may be rendered moot next semester if the latest Omicron variant of COVID-19 continues to wreak havoc on reopening efforts across the globe). There are 6th-generation Canadians and students who have only landed in this country just a few short months ago. The socioeconomic diversity of this hypothetical survey would not only be endless, but it would be indicative of the socioeconomic diversity that can be seen across Canada’s 293 census division. Some divisions have much higher racial diversity than others (or alternatively, have a higher concentration of Indigenous people than others). Some divisions have higher educational attainment rates than others.

With some census divisions experiencing different levels of net migration (whether positive or negative), it leads one to wonder what socioeconomic variables would be most associated with levels of net migration. Would the findings and insights from modelling their causal relationships (or lack thereof) be sufficient to help guide public policy from the federal to the local level? Could there be unknown or known confounding variables that could make two or more correlated variables appear associated?

DATA COLLECTION

Datasets and Variables of Interest

Dataset 1:
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E

Dataset 1 Variables of Interest:

- Census Division
- Province
- 2016 Population
- Post secondary credentials attainment rate
- Average income
- Unemployment rate
- Visible Minority Rate
- Indigenous Rate
- Median Age

Dataset 2: **https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710014001**

Dataset 2 Variables of Interest:

- Mortality (2015-16, 2016-17, 2017-18, 2018-19)
- Net inter and intraprovincial migration (2015-16, 2016-17, 2017-18, 2018-19)

METHODOLOGY

Programming

The R language, a staple of statistical analysis, will be used via R Version: rstudio-2021.09.1
Libraries: car, spida2, effects, lattice, latticeExtra

Feature Engineering

- New variables (9):
    - 4-year average mortality
    - 4-year average mortality per million people
    - Net migration (sum of net interprovincial and intraprovincial migration for 2015-16, 2016-17, 2017-18, 2018-19)
    - 4-year average of calculated net migration
    - 4-year average of calculated net migration per million people
    - POC rate (sum of Visible Minority and Indigenous rate)
- Categorized Variables (7):
    - Mig, Mort, Post.Sec, Un, VM, Ind, POC
- Variables of Interest:
    - Mig - 4-year average of calculated net migration per million people
    - Mort - 4-year average mortality per million people
    - Post.Sec - Post secondary credentials attainment rate
    - Avg.Inc - Average income
    - Un - Unemployment rate
    - VM - Visible minority rate
    - Ind - Indigenous rate
    - POC - Combined visible minority and Indigenous rate
    - Med.Age - Median age

Reasonings Behind Variable Selection and Engineering Choices:

(1) Using yearly average for variables whose data is collected yearly (e.g., mortality) instead of one year will avoid inadvertently using a year where special circumstances occurred. 1 (2) Speaking of special circumstances, 2019/20 data was intentionally excluded to exclude the extraordinary impacts of the COVID-19 pandemic. (3) Some variables that appear related to socioeconomic push factors were selected (Mort and Med.Age could be related to health, Post.Sec could be related to educational opportunities, Avg.Inc and Un are related to each observation's economic conditions) (4) VM and Ind were included, not because they are direct push factors, but because due to Canada's ongoing legacy of systemic racism, both groups continue to be socioeconomically disadvantaged compared to the general population. Therefore, this project aims to explore their relation to the socioeconomic conditions of each observation.

Modelling

- Categorical Response Variable:
    - Mig:
        - 1 if average migration rate is below -5000
        - 2 if average migration rate is between -5000 and -1000
        - 3 if average migration rate is between -1000 and 1000
        - 4 if average migration rate is between 1000 and 5000
        - 5 if average migration rate is greater than 5000
- Categorical Predictor Variables:
    - Mort: 1 if average mortality rate is below 5000, 2 if average mortality rate is between 5000 and 7500 3 if average mortality rate is between 7500 and 10,000 4 if average mortalityrate is between 10,000 and 12,500 5 if average mortality rate is greater than 12,500
    - Post.Sec: 1 if post secondary credentials attainment rate is below 30%, 2 if post secondary credentials attainment rate is between 30-40%, 3 if post secondary credentials attainment rate is between 40-50%, 4 if post secondary credentials attainment rate is between 50-60%, 5 if post secondary credentials attainment rate is greater than 60%
    - Un: 1 if unemployment rate is below 5%, 2 if unemployment rate is between 5-10%, 3 if unemployment rate is between 10-15%, 4 if unemployment rate is between 15-20%, 5 if unemployment rate is greater than 20%
    - VM: 1 if Visible Minority rate is below 5%, 2 if Visible Minority rate is between 5-10%, 3 if Visible Minority rate is between 10-20%, 4 if Visible Minority rate is between 20-40%, 5 if Visible Minority rate is greater than 40%
    - Ind: 1 if Indigenous rate is below 5%, 2 if Indigenous rate is between 5-10%, 3 if Indigenous rate is between 10-20%, 4 if Indigenous rate is between 20-40%, 5 if Indigenous rate is greater than 40%
    - POC: 1 if POC rate is below 5%, 2 if POC rate is between 5-10%, 3 if POC rate is between 10-20%, 4 if POC rate is between 20-40%, 5 if POC rate is greater than 40%
- Continuous Predictor Variables:
    - Avg.Inc
    - Med.Age

Ind, VM, and POC will be used as interaction variables on each variable of interest.

RESULTS

Investigating VM, Ind, and POC as Interaction Variables:

Variable of interest: Mort

- Models: Mig ~ Mort*VM, Mig ~ Mort*Ind, Mig ~ Mort*POC,
- p-values: Consistently high for VM and POC, varied for Ind
- Standard errors: High for Ind
- Conclusion: Safe to only move forward with Ind for investigation. Due to high standard errors, Ind is not a good interaction variable for Mort.

Other variables of interest: Post.Sec, Avg.Income, Un, MedAge

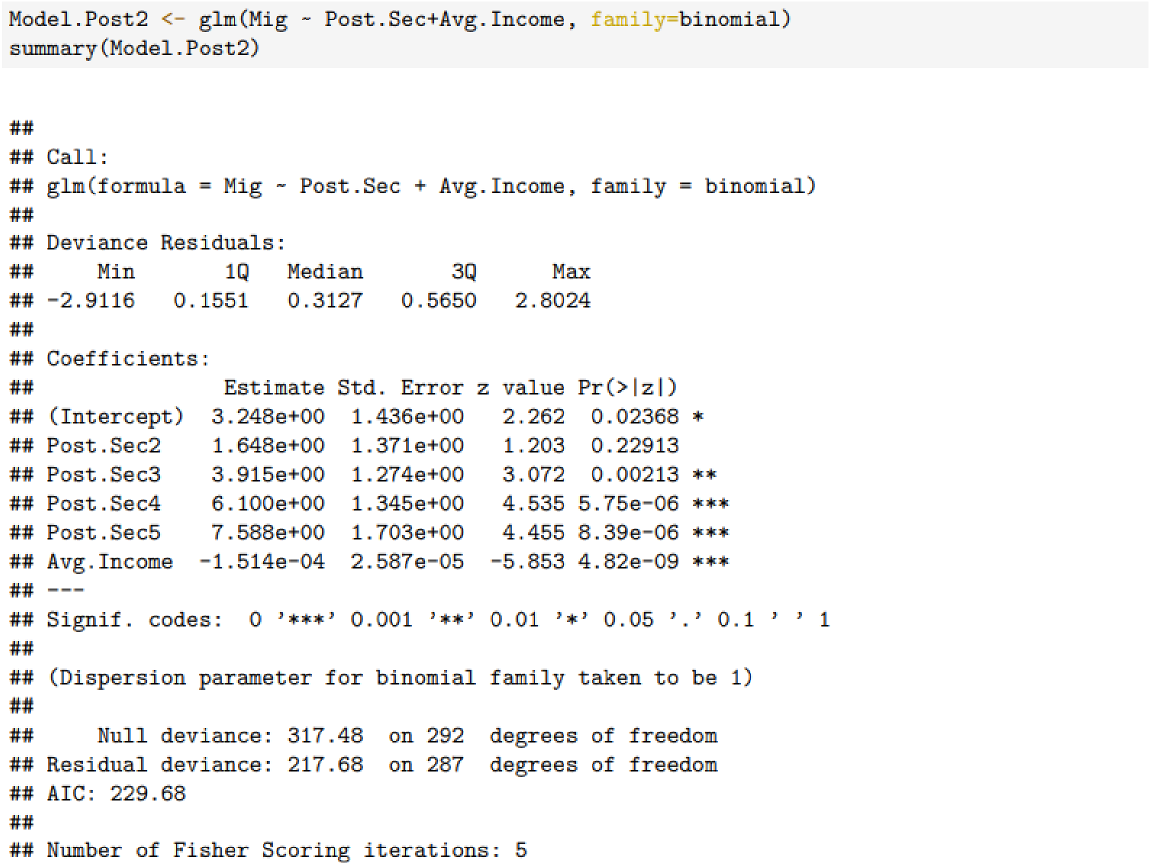
- ○ Models: Mig ~ Post.Sec*Ind, Mig ~ Avg.Inc*Ind, Mig ~ Un*Ind,Mig ~ MedAge*Ind
- ○ Results:
    - ■ Possible Hauck-Donner phenomenon for Post.Sec
    - ■ Lower p-values and more acceptable standard errors for Avg.Income, but still not good enough
    - ■ For third model, Only Ind2 has positive net mig.
    - ■ For fourth model, Standard errors are low but not as low as in Avg.Income. P-values are still too high.
- ○ Conclusion: Ind would probably not be the best interaction variable to use. Ind to be included as some other kind of predictor variable.

Testing Each Variable of Interest

Methodology for each variable of interest:

- STEP 1: Take main model: Mig ~ [variable of interest]
- STEP 2: Make 5 test models, adding other variable of interest: Mig ~ [variable of interest]+[second variable of interest]
- STEP 3: Comment and compare test models with main model to see effects of adding variables, e.g. SEs (standard errors) and est.

SAMPLE:



```
Model.Post2 <- glm(Mig ~ Post.Sec+Avg.Income, family=binomial)
summary(Model.Post2)
```

```
##
## Call:
## glm(formula = Mig ~ Post.Sec + Avg.Income, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9116   0.1551   0.3127   0.5650   2.8024
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.248e+00  1.436e+00   2.262  0.02368 *
## Post.Sec2    1.648e+00  1.371e+00   1.203  0.22913
## Post.Sec3    3.915e+00  1.274e+00   3.072  0.00213 **
## Post.Sec4    6.100e+00  1.345e+00   4.535 5.75e-06 ***
## Post.Sec5    7.588e+00  1.703e+00   4.455 8.39e-06 ***
## Avg.Income  -1.514e-04  2.587e-05  -5.853 4.82e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 317.48  on 292  degrees of freedom
## Residual deviance: 217.68  on 287  degrees of freedom
## AIC: 229.68
##
## Number of Fisher Scoring iterations: 5
```

Effects of Other Variables on Mort:

- Main Model: Mig ~ Mort
- Test Models: Mig ~ Mort+Post.Sec, Mig ~ Mort + Avg.Income, Mig ~ Mort+Un, Mig ~ Mort+MedAge, Mig ~ Mort+Ind
- Results:
    - First test model: When Post.Sec added, Mort doesn't change SEs much, but changes estimates a lot (except for Mort2)
    - Second test model: After adding Avg.Income, SE changes minimally, Estimates greatly change.
    - Third test model: SE and Est minimal change.
    - Fourth test model: Estimates greatly change, SEs don't.
    - Fifth test model: Estimates greatly change, SEs don't.

Effects of Other Variables on Post.Sec:

- Main Model: Mig ~ Post.Sec
- Test Models: Mig ~ Post.Sec + Mort, Mig ~ Post.Sec + Avg.Income, Mig ~ Post.Sec + Un, Mig ~ Post.Sec+MedAge, Mig ~ Post.Sec+Ind
- Results:
    - First test model: Minimal change for both Est and SEs, but big change for Post.Sec.2
    - Second test model: Est greatly change, not so much for Standard Errors.
    - Third test model: Est change varies, but est mostly unchanged. SEs are also unchanged. assume independent.
    - Fourth test model: Est greatly changes, not so much for SEs.
    - Fifth test model: Est greatly changes, not so much for SEs.

Effects of Other Variables on Avg.Income:

- Main Model: Mig ~ Avg.Income
- Test Models: Mig ~ Avg.Income + Mort, Mig ~ Avg.Income + Avg.Income, Mig ~ Avg.Income + Un, Mig ~ Avg.Income+MedAge, Mig ~ Avg.Income+Ind
- Results:
    - First test model: This confirms that Average income and mort are ind.
    - Second test model: Est changes a lot.
    - Third test model: Est changes not as much as before. Assume independence.
    - Fourth test model: Est changes greatly.
    - Fifth test model: Minimal change in est and SEs. Independent.

Effects of Other Variables on Un:

- Main Model: Mig ~ Un
- Test Models: Mig ~ Un + Mort, Mig ~ Un + Post.Sec, Mig ~ Un + Avg.Income, Mig ~ Un + MedAge, Mig ~ Un + Ind
- Results:
    - First test model: Again all except Un2 has minimal est change.
    - Second test model: Massive change in Est not so much in SE.
    - Third test model: No Significant change in Est.

    - Fourth test model: All except Un2 no significant change.
    - Fifth test model: Massive change in Est not so much in SE.

Effects of Other Variables on MedAge:

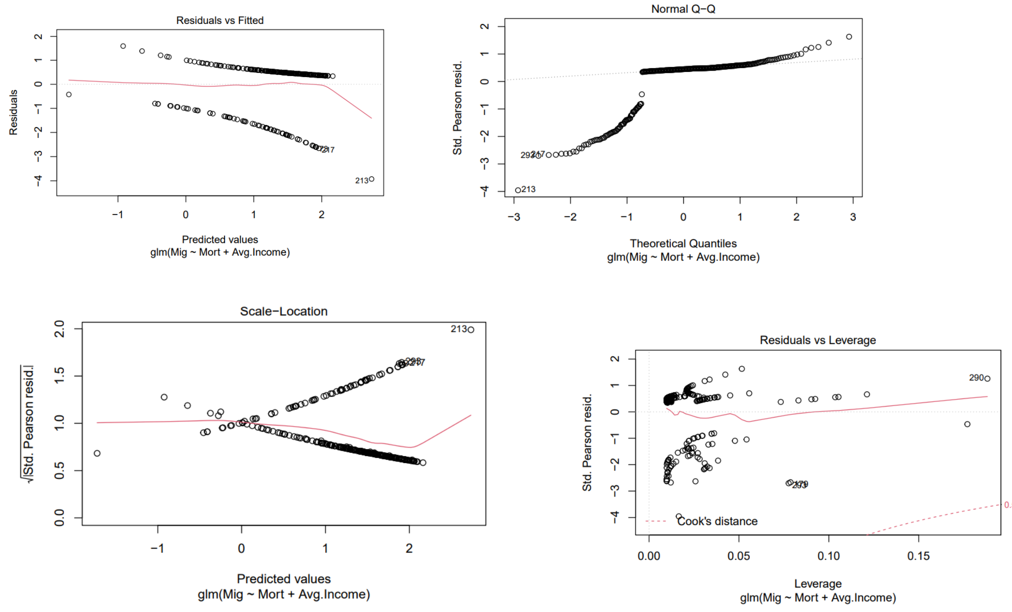
- Main Model: Mig ~ MedAge
- Test Models: Mig ~ MedAge + Mort, Mig ~ MedAge + Post.Sec, Mig ~ MedAge + Avg.Income, Mig ~ MedAge + Un, Mig ~ MedAge + Ind
- Results:
    - First test model: Sig in Est change not so much SE.
    - Second test model: Not as big of a change with est.
    - Third test model: Little changes.
    - Fourth test model: Little changes.
    - Fifth test model: Little changes.

Effects of Other Variables on Ind:

- Main Model: Mig ~ MedAge
- Test Models: Mig ~ Ind + Mort, Mig ~ Ind + Post.Sec, Mig ~ Ind + Avg.Income, Mig ~ Ind + Un, Mig ~ Ind + MedAge
- Results:
    - First test model: All except Ind2 sees minimal change in Est.
    - Second test model: Minimal change in Est. Independent.
    - Third test model: Minimal change in Est. Independent.
    - Fourth test model: All except Ind2 see minimal change in Est.
    - Fifth test model: All except Ind2 see minimal change in Est.

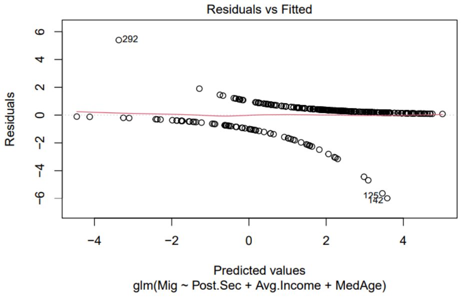
Residual Plots and ANOVA:

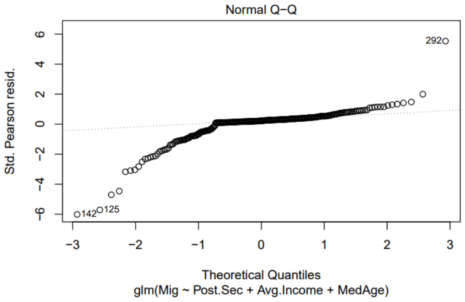
Mig ~ Mort+Avg.Income

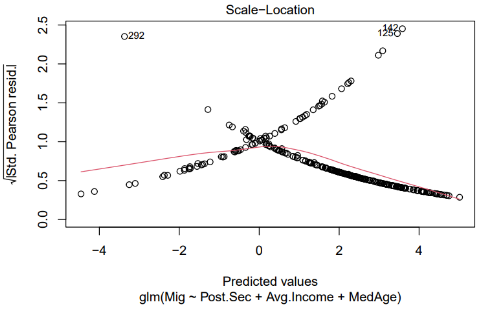


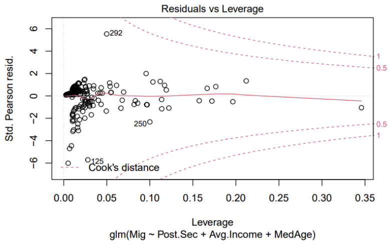
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Mig
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                        292     317.48
## Mort        4   11.247      288     306.24 0.023921 *
## Avg.Income  1   10.432      287     295.80 0.001238 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mig ~ Post.Sec+Avg.Income+MedAge



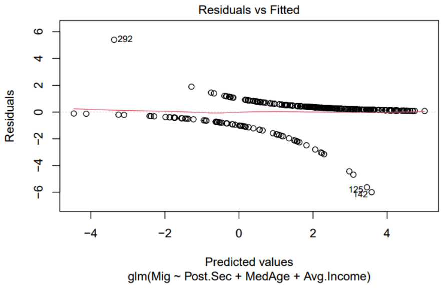


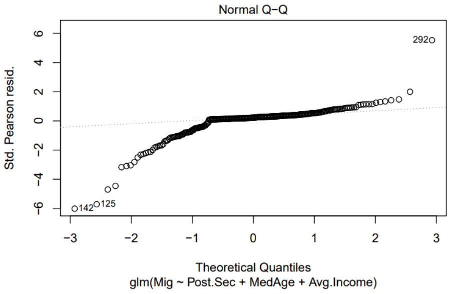


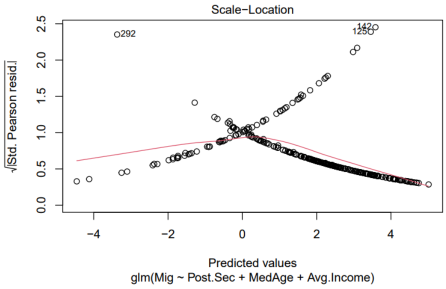


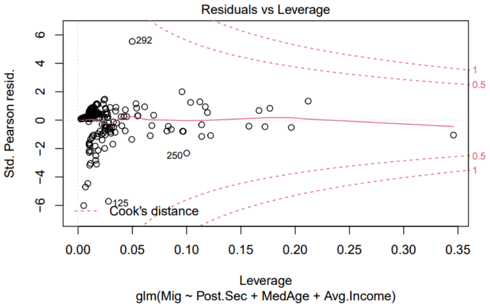
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Mig
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        292     317.48
## Post.Sec    4   52.026      288     265.46 1.362e-10 ***
## Avg.Income  1   47.778      287     217.68 4.773e-12 ***
## MedAge      1    6.777      286     210.90  0.009235 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mig ~ Mort+Avg.Income



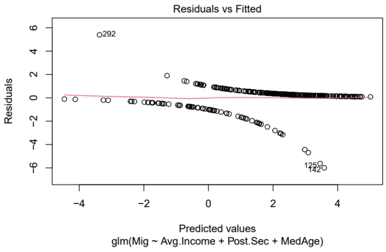


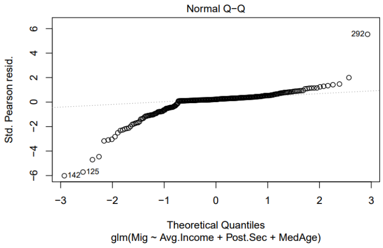


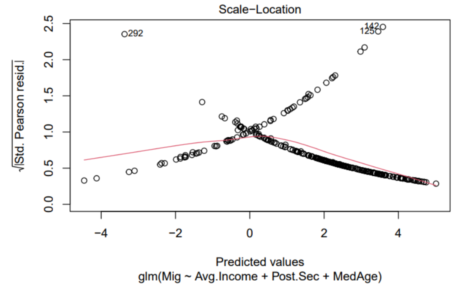


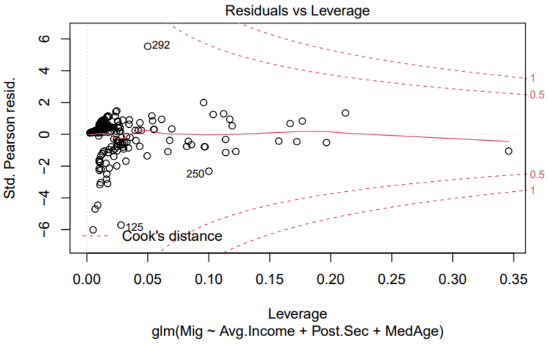
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Mig
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           292     317.48
## Post.Sec      4   52.026       288     265.46 1.362e-10 ***
## MedAge        1   39.919       287     225.54 2.647e-10 ***
## Avg.Income    1   14.636       286     210.90 0.0001304 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mig ~ Avg.Income+Post.Sec+MedAge



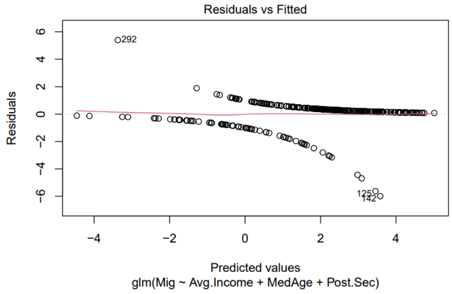


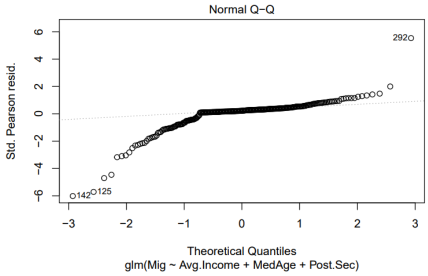


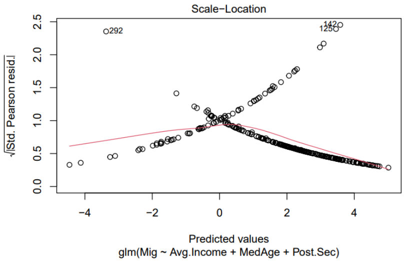


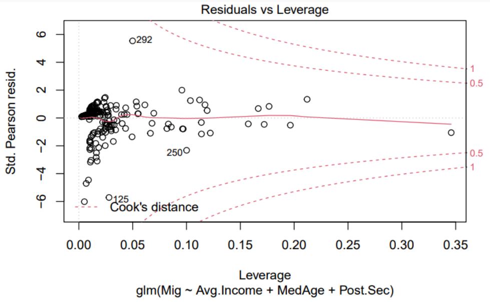
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Mig
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         292     317.48
## Avg.Income  1   18.014       291     299.47 2.193e-05 ***
## Post.Sec    4   81.791       287     217.68 < 2.2e-16 ***
## MedAge      1    6.777       286     210.90  0.009235 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Mig ~ Avg.Income+MedAge+Post.Sec



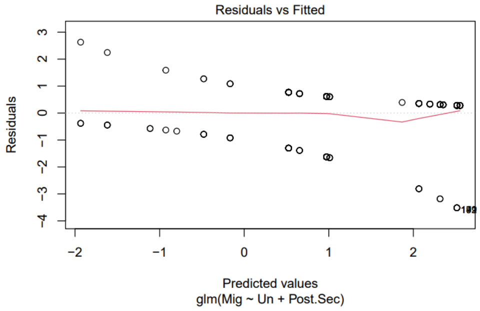


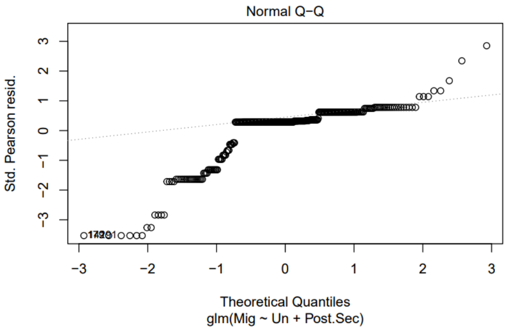


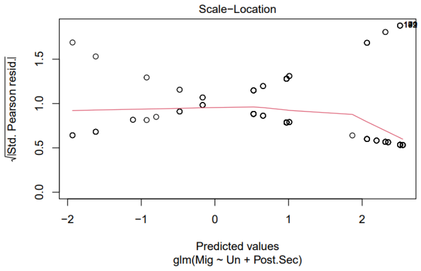


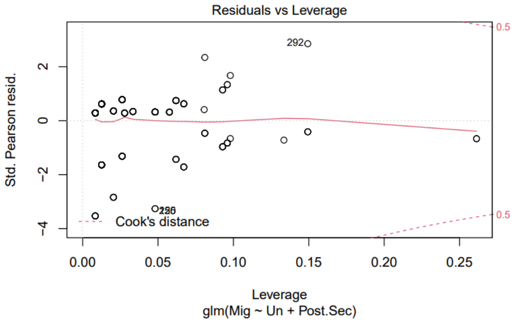
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Mig
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         292     317.48
## Avg.Income  1   18.014       291     299.47 2.193e-05 ***
## MedAge      1   35.686       290     263.78 2.319e-09 ***
## Post.Sec    4   52.882       286     210.90 9.021e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Mig ~ Un+Post.Sec



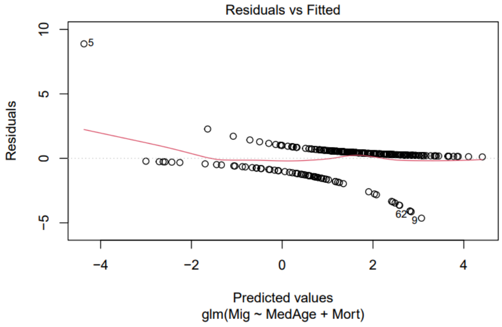


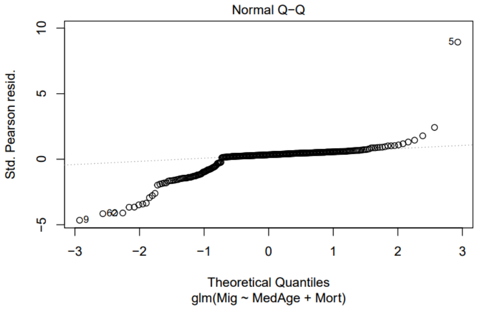


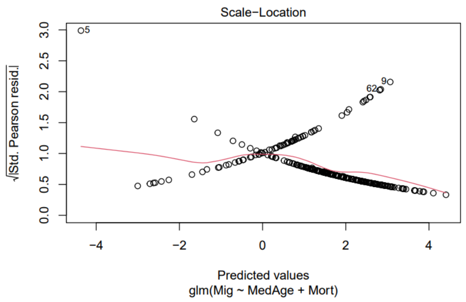


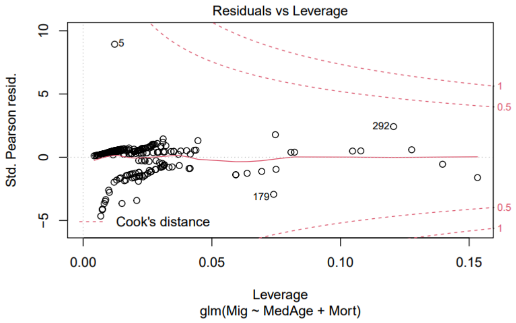
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Mig
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       292     317.48
## Un        4   24.195       288     293.29 7.299e-05 ***
## Post.Sec  4   31.798       284     261.49 2.104e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mig ~ MedAge+Mort









```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Mig
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     292     317.48
## MedAge  1   52.198       291     265.29 5.019e-13 ***
## Mort    4   13.157       287     252.13   0.01053 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CONCLUSION

It appears that, based on the p-values and the coefficient estimates, that average income and post-secondary opportunities have the biggest influence on out-migration. That makes sense, since average income is tied to economic opportunities, and post-secondary opportunities are tied to job prospects. Poverty reduction measures can be a key policy in helping increase economic opportunities. For example, launching a basic income pilot project can help low-income earners. By increasing their financial fluidity, they have an increased capability of easily providing for themselves such as groceries and healthcare, reducing the stress and allowing them to shift their focus towards other matters that could augment their financial health such as career advancement and educational attainment. Speaking of education, not every community will have the capacity to host a university nor any post-secondary institution, which could explain education's confounding effect on out-migration. Regardless, local governments could still incentivize educated community members to return home with a more diversified economy.

Of course, this is not to say that this is perfect modelling. This is simply just one interpretation out of many. It is possible that different factorizations of categorical variables could yield dramatically different results. It is also just as possible that there are socioeconomic indicators not captured by Statistics Canada that would both create better models and also better answer my questions. The main takeaway from this project is that official national data gathering institutions still do not fully capture the socioeconomic picture of Canada's diverse geopolitical divisions. There is a lack of specific data on things such as how much systemic racism plays into such as healthcare access, mental health indicators, job prospects, and infrastructure development for communities with higher rates of Visible Minorities and Indigenous communities.