STAT3503A – Final Project Report

Carleton University
School of Mathematics and Statistics

Table of Contents

Background and Motivation

According to the standard definition employed by Statistics Canada, the employed are persons having a job or business, whereas the unemployed are without work, are available for work, and are actively seeking work. Together the unemployed and the employed constitute the labor force, which reflects the country's economy growth and can be used to explain various social movements and issues.

Using Stat Canada's criteria, unemployed persons are those who, during reference week

- were on temporary layoff during the reference week with an expectation of recall and were available for work,
- were without work, had looked for work in the past four weeks
- were available for work or had a new job to start within four weeks from reference week and were available for work.

Statistics Canada believes the ratio of people in the labor force and people who is economically inactive can be used as a planning tool for decision making. Some noble uses of this number include developing government's policy to create balance of some programs such as public pension and changing the distribution of fund to social programs which focus on knowledge transfer, immigrant integration and employment equity.

The OpenGov has published and ongoing updated various datasets regarding to labor force survey, marital status distribution, and Gross Domestic Product (GDP) from past years. The first two datasets reflect how population is distributed across provinces and provide insight into the characteristics of population like gender and age distribution, educational levels, and employment statuses. On another hand, by definition, GDP is the final value of the goods and services produced within the geographic boundaries of a country within a period, which is normally a year. GDP growth rate is an important indicator of the economic performance of a country. The assumption we make for this project is the economic growth positively proportional to the number of job availability of the country.

The main goal of this project is to examine the number using various statistical analysis methods to answer question: "Can we estimate labor force efficiently using census information of population's characteristics and economy growth?". Currently, we obtain the labour force statistics from conducting surveys which can be expensive (from providing incentive, and expensive cost of labour and training) and have some limitations such as surveys can be bias, have risks of no-response and require an appropriate sampling method. At the same time, we are given lots of resources on census of population and GDP and strong frameworks to estimate that information. There are two main hypothesis we want to test here:

- Education level, marital status, and geography attributes are significant in our final Canada Labour Force estimator.
- The change in GDP can also be a helpful predictor.

Secondly, we build a model to estimate the Ontario Unemployed Rate using the available data. Knowing the factors that contribute to that rate can be extremely helpful to justify if the programs available from Employment Ontario are targeting the roots of unemployment. There are several questions that hope to be explained from the analysis.

- Is it effective to allocate fund to the program that helps unemployment people receive education and job training?
- Are martial statuses such are number of people getting married or divorced important in our model?

- Can we find any interaction between variables in our model?

We believe it will be helpful in extracting the attributes to predict Canada labor force and examining the relationship between attributes. There are many models can be built based on different attribute combination, and our goal is to find the best model that fit whatever we are exploring. In this project, we will define the best model to be the one with minimal AIC, also for safety check, we will also consider coefficient of determination (R^2) and Mean Square Error (MSE).

Dataset Description

Data Source

1. **Population.** This dataset covers the period 1990-2019 and includes number of persons in the labor force (employment and unemployment) and not in the labor force, unemployment rate, participation rate, and employment rate, by educational degree, sex, and age group, last 5 years. Each row contains the year, geography, labor force characteristics (Population, Employment either Full Time and Part Time, Unemployment…), educational degree (no degree/diploma, without/with high school graduation, post-secondary level and higher than post-secondary level), gender (male and female), age group (15-24, 25 and over, 25-54, 55-64), and the number within population corresponded to the declare statuses.
2. **Marital Status.** This dataset is customized from StatCan[1] website interface, and the downloadable file is available through my Dropbox link[2]. The dataset contains information of marital status information, which is grouped by Provincial, age group and by gender from 1990-2020.
3. **Gross Domestic Product (GDP).** This dataset covers basic prices from 1997 to 2020, by North American Industry Classification System (NAICS) aggregates, by Industry, volume measures, monthly, 5 most recent time periods.

Engineered Dataset for Model Building

1. For Labour Force Model: includes 23,457 rows and 29 columns.
   - **Year**(numerical): range from 1998-2020
   - **GEO**(categorical): Alberta, British Columbia, Manitoba, New Brunswick, Nova Scotia, Ontario, Prince Edward Island, Quebec, Saskatchewan.
   - **Sex**(categorical): Males and Females
   - **Educational.degree**(categorical): Above bachelor's degree, Bachelor's degree, High school graduate, High school graduate, some post-secondary, No degree, certificate or diploma, Postsecondary certificate or diploma, University degree, With high school graduation, Without high school graduation.
   - **Age.group**(categorical): 15 years and over,25 to 54 years,25 years and over,15 to 24 years, 55 to 64 years,55 years and over,65 years and over.
   - **Population** (in thousand): the population resides in the perspective province.

[1] StatCan Orignial Marital Status dataset can be retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=1710006001

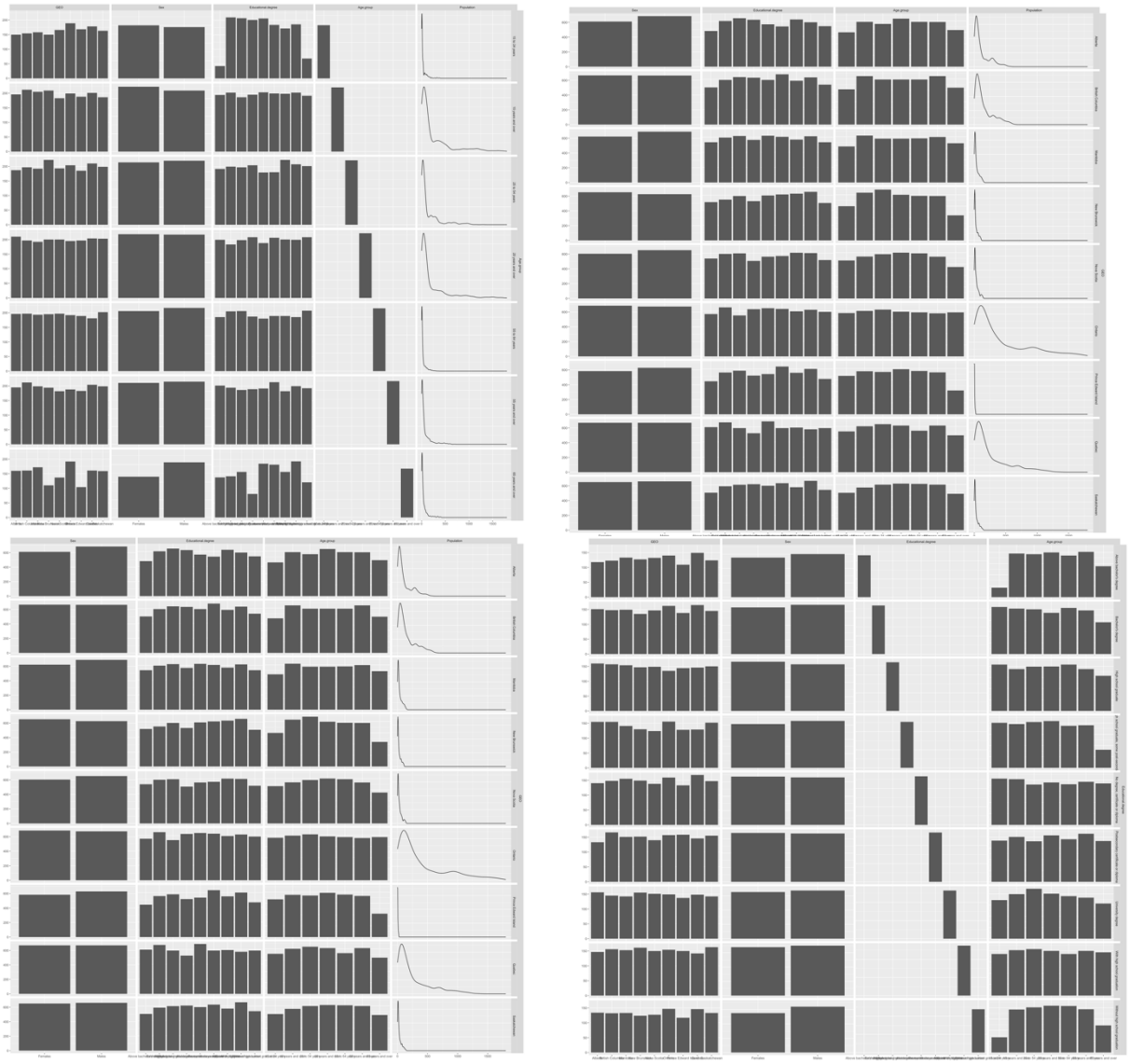
[2] Downloadable Martial Status dataset from https://www.dropbox.com/s/mocpwtrxdqn4zwy/martial_status.csv?dl=1

- Married.All.ages,Married.15.to.44.years, Married.45.to.64.years, and Married.65.years.and.over (numerical): number of people with married status in each age group.
- Divorced.All.ages, Divorced.15.to.44.years, Divorced.45.to.64.years, Divorced.65.years.and.over (numerical): number of people with divorced status in each age group.
- **LabourForce**(numerical): the number of people in labour force that matches categories above.
- GoodProduct_change, ServiceProduct_change, DurableManu_change, Arigiculutre_change, Mining_change, Utilities_change, Construction_change, Educational_change, PublicAdmin_change, Federal_change(numerical): change in industry's GDP by year.
- **CanadaGDP**: GDP change in Canada per year
- **GDP_whole**: total GDP of that year
- **Rate_GDP_change**: percentage of GDP increase in perspective of last year.

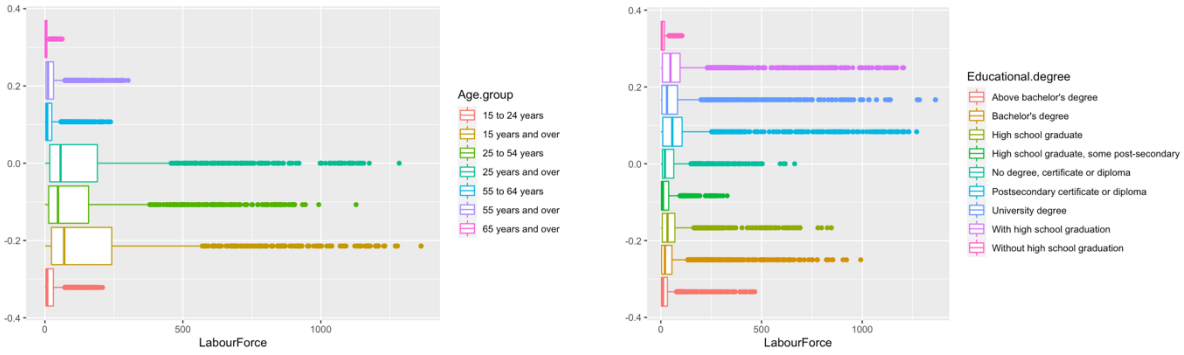
2. For Ontario's Unemployed Rate Model has 1,442 and 14 column.
   - **Year**(numerical): range from 1998-2020
   - **Sex**(categorical): Males and Females
   - **Educational.degree**: similar to the dataset above.
   - **Age.group**: 15 to 24 years, 25 to 54 years, and 25 years and ove.
   - **Married.All.ages,Married.15.to.44.years and Married.45.to.64.years**: similar to the above dataset.
   - **Divorced.All.ages,Divorced.15.to.44.years and Divorced.45.to.64.years**: similar to the above dataset.
   - **totalPopulation**: the total number of people resided in Ontario
   - **Ontario_change**: the change in population in Ontario each year
   - **Ontario_rate**: the change in percentage in population in Ontario each year
   - **UnempRate**: the rate of unemployed in perspective of group by Year,Sex, Educational.degree and Age.group.

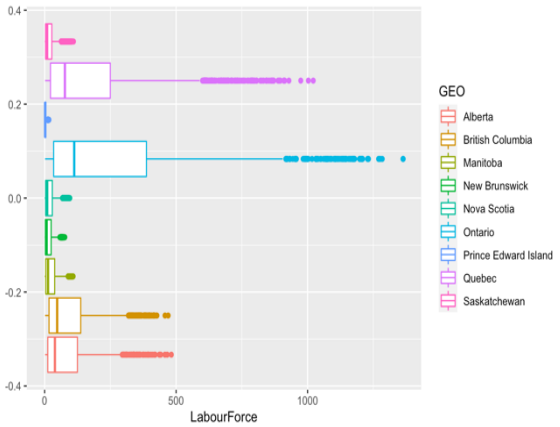
Explanatory Graph

1. Labour Force Dataset



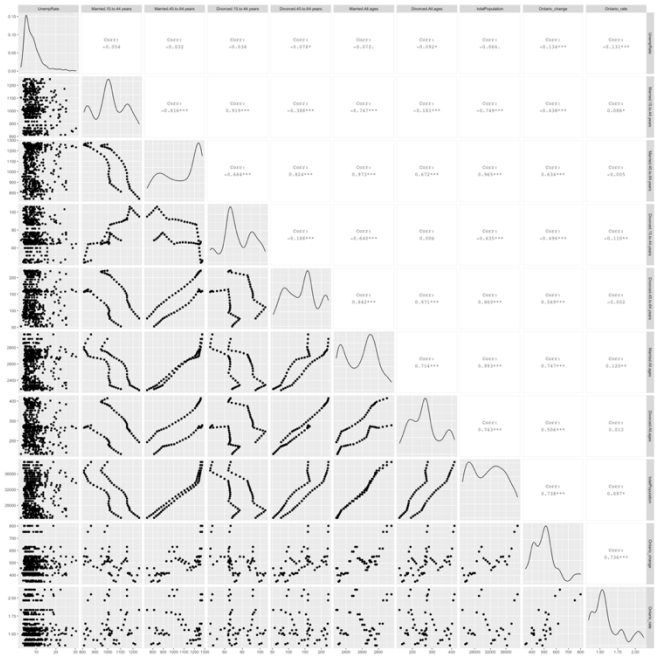
Comparing categorical predictor, we do not see many interactions regards to this. The ratio between categorical predictor does not show any clear interaction. We decide to not test any interaction in this model.





We notice there are some groups that are dominant in term of labour force. As we assume, the main working age is around 25 to 54, while 15-24 age group is quite low here because people tend to focus on education. Secondly, there is a noticeable difference in people with post-secondary, followed by high school graduation level in comparison with other type of degree holders. Lastly, the provinces with concentrated population such as Ontario, Quebec, Alberta, and Manitoba are predicted to have significant impact on our model.

2. Ontario Unemployment Rate Dataset



Married.all.age seems to have high correlation with other, so does population. Therefore, we believe that is worthy to test their interactions with other variables.

Regression Analysis

For both sections, we divide the datasets into train and test sections. Firstly, we fit all the model into train dataset and compare them to find the optimal. Then, I fit the best one to our test dataset, inspect it with diagnostic plot and apply transformation if needed.

Labour Force Model

At first, we test some variables which are believed to be important. The result is the population variable has the lowest AIC in term of estimating Labour Force Value.
Model selection based on AICc:

```
               K     AICc Delta_AICc AICcWt Cum.Wt        LL
pop.mod        3 121991.5       0.00      1      1 -60992.74
GEO.mod       10 148834.5   26842.99      0      1 -74407.23
married.mod    3 148880.5   26888.96      0      1 -74437.23
divorced.mod   3 149104.9   27113.40      0      1 -74549.44
age.mod        8 150457.2   28465.76      0      1 -75220.62
edu.mod       10 151589.5   29598.05      0      1 -75784.76
GDP.mod        3 152389.7   30398.22      0      1 -76191.85
```

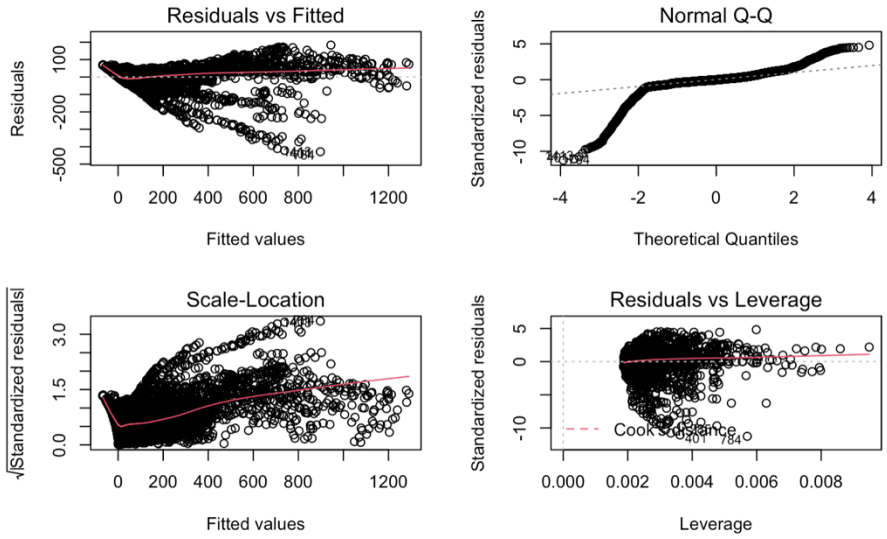
Cross-check model selection using ANOVA:

```
Analysis of Variance Table
Model 1: LabourForce ~ Age.group
Model 2: LabourForce ~ Educational.degree
Model 3: LabourForce ~ GEO
Model 4: LabourForce ~ Married.All.ages
Model 5: LabourForce ~ Divorced.All.ages
Model 6: LabourForce ~ Population
Model 7: LabourForce ~ CanadaGDP
  Res.Df       RSS Df  Sum of Sq      F    Pr(>F)
1  11645 276176408
2  11643 304256790  2  -28080382
3  11643 240189209  0   64067581
4  11650 241428857 -7   -1239648 6.7768 4.767e-08 ***
5  11650 246124276  0   -4695419
6  11650  24020454  0  222103822
7  11650 326277118  0 -302256664
```

Combining the results lead us to suspect population and the number of married people should be included in my final model. At the same time, we run the stepwise model selection on all the predictors we have. The model using population and total married has AIC = 122906 with 4 as df value while the model resulted from step model has 31 df and ends up with 118340. It is common to think more variables, the better. However, we aim to see if we can beat the stepwise mode with a model least variable and AIC.

We test random interactions and end has the Model1 that use GEO, total population, total of married and divorce people, Age.group, and education with AIC = 118431.5 with 2 as 27 degree of freedom. From AIC comparison, that model has the lower AIC than our resulted step model. However, ANOVA method disagrees with the results. However, I think having less variables in the model seems to be more optimal since we can save computational work in fact. Therefore, we discard the step model and select Model1 to be our final model.

We try to add GDP-related variables to Model1 and that results in the increase in AIC, so Model1 is chosen the way it is. Next, we fit Model1 variables to our test dataset and generate the diagnostic plot from it.
The basic statistics of model1 in our test dataset is:

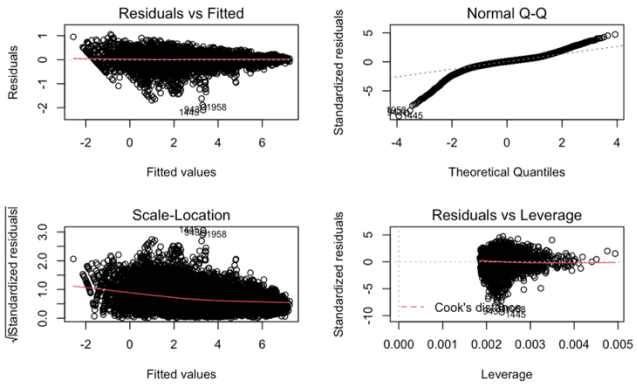



```
Residual standard error: 36.27 on 11735 degrees of freedom
Multiple R-squared: 0.954,Adjusted R-squared: 0.9539
F-statistic: 9733 on 25 and 11735 DF,  p-value: < 2.2e-16
```

There are some couple observations worth noticed:

- Residual vs Fitted has some problems in the exceedingly early values.
- Normal QQ shows the distribution is a bit of skewed.
- Scale-Location shows the variability (variances) of the residual points decreases at first then increases with the value of the fitted outcome variable, suggesting non-constant variances in the residual's errors (or *heteroscedasticity*).
- Residuals and Leverage has some extreme predictor value.

To solve this, we apply log transformation on every predictor which is quantities to improve the diagnostic plot. Notice the Residual vs Fitted does not show any strange pattern anymore and QQ plot is better distributed. However, Scale Location is better but it is still concerning.




```
Residual standard error: 0.2247 on 11735 degrees of freedom
Multiple R-squared:  0.985,
      Adjusted R-squared: 0.985
F-statistic: 3.079e+04 on 25 and 11735 DF,  p-value: < 2.2e-16
```

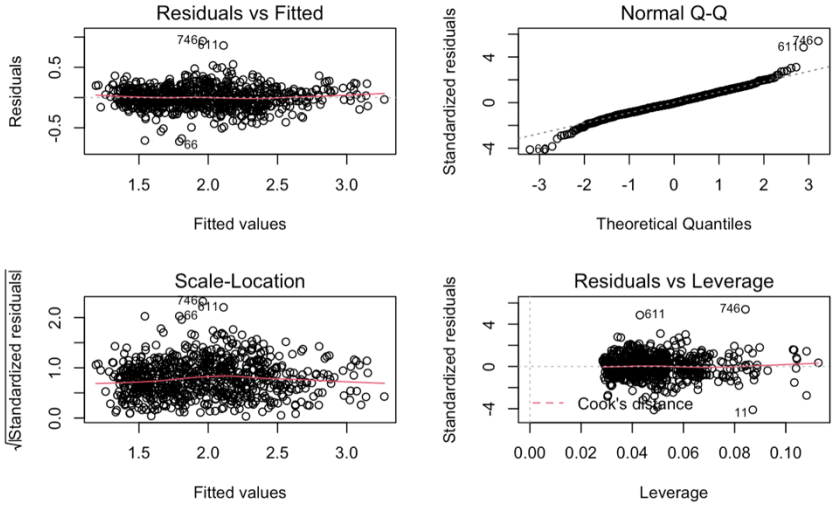
There is a big raised in R-squared, that means our transformation does show significant improvement.

Ontario Unemployment Rate Model

For this part, we focus on inspect the interaction between variables. Experiencing from another model, we decide to apply log transformation to all numerical predictor and run step model selection on all interactions. The R-squares for the best non-log transformation is at max 0.80 but when we apply transformation, so the selected model has a significant improve.

```
Residual standard error: 0.181 on 727 degrees of freedom
Multiple R-squared:  0.8416,Adjusted R-squared:  0.8338
F-statistic: 107.3 on 36 and 727 DF,  p-value: < 2.2e-16
```

Fit the model from stepwise selection to the testing dataset, now we inspect it using the diagnostic plot.



Model is not so bad to be honest, the pattern is not perfect, but it is not concerning.

Discussions

Labour Force Model

```
Call:
lm(formula = log(LabourForce) ~ GEO + Age.group + log(Population) +
    Educational.degree + log(Married.All.ages) + log(Divorced.All.ages),
    data = dataTestToUse)
Residual standard error: 0.2247 on 11735 degrees of freedom
Multiple R-squared:  0.985, Adjusted      R-squared:  0.985
F-statistic: 3.079e+04 on 25 and 11735 DF, p-value: < 2.2e-16
                                                          2.5 %      97.5 %
(Intercept)                                         -5.47026332 -4.53673644
GEOBritish Columbia                                 -0.26121883 -0.21596343
GEOManitoba                                          0.48911565  0.64449489
GEONew Brunswick                                     0.63488393  0.85216273
GEONova Scotia                                       0.57490492  0.76422340
GEOOntario                                          -1.05603600 -0.87289334
```

```
GEOPrince Edward Island                                          1.59657528  2.03693697
GEOQuebec                                                       -0.31238397 -0.23611240
GEOSaskatchewan                                                  0.60267485  0.77735553
Age.group15 years and over                                      -0.09693450 -0.04915361
Age.group25 to 54 years                                          0.13322999  0.17358615
Age.group25 years and over                                      -0.11178386 -0.06581270
Age.group55 to 64 years                                         -0.20900391 -0.17771749
Age.group55 years and over                                      -0.69601644 -0.66115802
Age.group65 years and over                                      -1.73617103 -1.70240635
log(Population)                                                  0.97763652  0.99402457
Educational.degreeBachelor's degree                             -0.08024095 -0.04282221
Educational.degreeHigh school graduate                          -0.22298791 -0.18155966
Educational.degreeHigh school graduate, some post-secondary     -0.20300412 -0.16676097
Educational.degreeNo degree, certificate or diploma             -0.68109158 -0.63731278
Educational.degreePostsecondary certificate or diploma          -0.16081903 -0.11621143
Educational.degreeUniversity degree                             -0.05527715 -0.01636546
Educational.degreeWith high school graduation                   -0.12085438 -0.07787244
Educational.degreeWithout high school graduation                -0.36892861 -0.33203574
log(Married.All.ages)                                            0.97292438  1.12422127
log(Divorced.All.ages)                                          -0.43878752 -0.39441270
```

All variables in the final model are all significant theoretically. However, I notice:

- "Age.group15 years and over" is quite redundant and that shows in the confident interval show very minimal effect by having their range remarkably close to 0.
- "Educational.degreeBachelor's degree" and "Educational.degreeUniversity degree" have truly negligible impact to the model. It can be explained by knowing people usually consider other postsecondary education options for the lower cost and years of schooling. Some papers suggest that there are different outcomes with respect to the educational levels and the mismatch between occupation and schooling. I think our result does back up the study.

In comparison to starting predictors, it looks like GDP-related variables have no significance in our labour force model. We can confirm by adding CanadaGDP to our estimator.

```
Model 1: log(LabourForce) ~ GEO + Age.group + log(Population) + Educational.degree +
    log(Married.All.ages) + log(Divorced.All.ages) + log(CanadaGDP)
Model 2: log(LabourForce) ~ GEO + Age.group + log(Population) + Educational.degree +
    log(Married.All.ages) + log(Divorced.All.ages)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  11734 537.84
2  11735 592.29 -1   -54.449 1187.9 < 2.2e-16 ***
```

We try to test our hypothesis if the economy growth from GDP can impact labour force and the anova test rejects that claim. However, we think GDP might not be the most accurate representation of economy growth, and the potential extension is to find the alternative representation. At this point, we can only make conclusion that annually GDP values has no impact to our model.

Ontario Unemployment Rate Model

We list the significant predictors along with their p-values.

```
Educational.degreeBachelor's degree   0.000120 ***
Educational.degreeHigh school graduate   2e-16 ***
Educational.degreeHigh school graduate, some post-secondary   < 2e-16 ***
Educational.degreeNo degree, certificate, or diploma   < 2e-16 ***
Educational.degreePostsecondary certificate or diploma .2.14e-08 ***
Educational.degreeWith high school graduation   2.97e-07 ***
Educational.degreeWithout high school graduation   < 2e-16 ***
Age.group25 to 54 years   < 2e-16 ***
Age.group25 years and over   < 2e-16 ***
log(Married.All.ages)   7.89e-05 ***
log(Married.15.to.44.years):log(Married.45.to.64.years)   1.63e-05 ***
log(Married.45.to.64.years):log(totalPopulation)   0.000112 ***
```

The educational degree attribute has significant impact on Ontarian ’s unemployability. Unlike the Canadian labour force model, higher education group has minimal impact in our unemployment rate model. The discovery somehow adds to the norm that people believe higher education can have better job security. Unemployment Ontario has a lot of support helping people get certificated and job training and Ontario government also accommodates citizens pursuing post-secondary education easier by giving out grant, bursary, and lower student loan interest.

The age group 25-44 which is major age group of working people in labour force is significant as expected. Any change in the group can impact the labour force of the country, and to the regional level like Ontario. The hypothesis we have at first is the number of young adult (15-24) would be significant to our model, but soon to be rejected.

It is surprised to discover that the number of married-related variables has more impact than the divorce's ones, and the model has some significant interactions of the married status within different age groups. Getting married can change a lot of things, from being eligible to different policies at workplace, tax benefits and incentive. Some studies do show that marital status and unemployment has two-way interactions.

Limitation

Data Validation:

There might some errors and misplacement when I try to merge three different datasets. The numbers are not aligning somewhere but since the final datasets are too big that we cannot visually inspect it. We try to remove N/A or 0 value off the dataset, but it might decrease the quality dataset along with the merge.

My intention at first is to predict the labour force of the next year using the past census data. However, my first attempted is to pivot all the categorical columns of GEO, Sex, Gender, Age. Group to make them numerical, hence easier to model in my knowledge. When we do that, more than 60,000 samples are minimized to 23 samples (correspond to each year in record) along with more than 60 predictors. We cannot proceed further since 23 samples are too little to test 60. To proceed with that we must research more in economic background to handpick the important variables and work around there. On another hand, as the professor Campbell

suggested, I could use Lasso Regression. Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This approach is a bit complex to my knowledge, therefore it is soon aborted.

Another point we want to mention is the use of log-transformation. This is a widely use technique to improve distribution and variability. According to some claims, log-transformation can be a bad trade off since it dismisses the presence of outliner. In general, this technique might change the data quality to look more promising, but we are confident after with the reliable level.

Project Extension

We think our project has a particularly good starting point for policy analyzing. There are many variables that we miss out in the model such as disability status, ethnicity, and government social benefits. Those can be especially important predictors to our labour force estimator according to some studies.

Rethinking the Project

There are many other things I could go with this project, like I can concatenate the other countries statistics to understand the labour force in nature and to test if the crucial factors can be varied based on difference of continents.
We look at different area instead of general labour force such as people working full-time and art-time or not-in-labour-force. Original population dataset gives a lots of information regard to those and may be align the current policy to any finding from those variables.

Conclusion

In this project, we built the labour force estimator model and Ontario Unemployment Rate model from different model selection methods, a mix of ANOVA, Stepwise selection and some decisions are based on R-Square. The model is selected by on the training data then validated by the testing data. Most of our hypothesis testing or interfering is based on optimal model fit using testing data. However, the current value fitting is prone to bad diagnostic plot nd having log-transformation. The significant observation from labour force model is GDP metric has no impact in our estimator. On another hand, we think GDP value is not the most accurate representation of economy growth. Therefore, we cannot make any statement regarding to the relationship between labour force and economy growth. Next, our Ontario unemployment model emphasizes the importance of marital status and educational degree, where the married data poses significant interaction with other variables. We also discussed our project has some flaws in data validation, and the auxiliary data from original datasets provide potential opportunities for this project extension.

Appendix I

Labour Force Dataset

Can be obtained through:
https://www.dropbox.com/s/sxlflueqf0dtol6/Veenguyen_STAT3503_labour_force.csv?dl=0

```
$ Year                      <dbl> 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1…
$ GEO                       <chr> "Alberta", "Alberta", "Alberta", "Alberta", "Alberta", "Alberta", "Alberta", "A…
$ Sex                       <chr> "Females", "Females", "Females", "Females", "Females", "Females", "Females", "F…
$ Educational.degree        <chr> "Above bachelor's degree", "Above bachelor's degree", "Above bachelor's degree"…
$ Age.group                 <chr> "15 years and over", "25 to 54 years", "25 years and over", "15 to 24 years", "…
$ Population                <dbl> 31.7, 27.3, 31.1, 8.2, 107.2, 87.4, 98.9, 6.7, 11.5, 49.0, 244.3, 148.9, 195.3,…
$ Total.All.ages            <dbl> 1436.211, 1436.211, 1436.211, 1436.211, 1436.211, 1436.211, 1436.211, 1436.211,…
$ Total.15.to.44.years      <dbl> 682.352, 682.352, 682.352, 682.352, 682.352, 682.352, 682.352, 682.352, 682.352…
$ Total.45.to.64.years      <dbl> 286.75, 286.75, 286.75, 286.75, 286.75, 286.75, 286.75, 286.75, 286.75, 286.75,…
$ Total.65.years.and.over   <dbl> 160.088, 160.088, 160.088, 160.088, 160.088, 160.088, 160.088, 160.088, 160.088…
$ Married.All.ages          <dbl> 588.083, 588.083, 588.083, 588.083, 588.083, 588.083, 588.083, 588.083, 588.083…
$ Married.15.to.44.years    <dbl> 315.1, 315.1, 315.1, 315.1, 315.1, 315.1, 315.1, 315.1, 315.1, 315.1, 315.1, 31…
$ Married.45.to.64.years    <dbl> 203.864, 203.864, 203.864, 203.864, 203.864, 203.864, 203.864, 203.864, 203.864…
$ Married.65.years.and.over <dbl> 69.119, 69.119, 69.119, 69.119, 69.119, 69.119, 69.119, 69.119, 69.119, 69.119,…
$ Divorced.All.ages         <dbl> 70.575, 70.575, 70.575, 70.575, 70.575, 70.575, 70.575, 70.575, 70.575, 70.575,…
$ Divorced.15.to.44.years   <dbl> 29.677, 29.677, 29.677, 29.677, 29.677, 29.677, 29.677, 29.677, 29.677, 29.677,…
$ Divorced.45.to.64.years   <dbl> 31.437, 31.437, 31.437, 31.437, 31.437, 31.437, 31.437, 31.437, 31.437, 31.437,…
$ Divorced.65.years.and.over <dbl> 9.461, 9.461, 9.461, 9.461, 9.461, 9.461, 9.461, 9.461, 9.461, 9.461, 9.…
$ LabourForce               <dbl> 26.3, 24.4, 25.7, 6.7, 85.4, 74.5, 78.7, 3.9, 4.2, 38.6, 173.1, 121.2, 134.5, 1…
$ GoodProduct_change        <dbl> 13708, 13708, 13708, 13708, 13708, 13708, 13708, 13708, 13708, 13708, 13708, 13…
$ ServiceProduct_change     <dbl> 24732, 24732, 24732, 24732, 24732, 24732, 24732, 24732, 24732, 24732, 24732, 24…
$ DurableManu_change        <dbl> 7845, 7845, 7845, 7845, 7845, 7845, 7845, 7845, 7845, 7845, 7845, 7845, 7845, 7…
$ Arigiculutre_change       <dbl> 2305, 2305, 2305, 2305, 2305, 2305, 2305, 2305, 2305, 2305, 2305, 2305, 2305, 2…
$ Mining_change             <dbl> 77270, 77270, 77270, 77270, 77270, 77270, 77270, 77270, 77270, 77270, 77270, 77…
$ Utilities_change          <dbl> 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37, 37,…
$ Construction_change       <dbl> -534, -534, -534, -534, -534, -534, -534, -534, -534, -534, -534, -534, -534, -…
$ Educational_change        <dbl> 245, 245, 245, 245, 245, 245, 245, 245, 245, 245, 245, 245, 245, 245, 245, 245,…
$ PublicAdmin_change        <dbl> 1869, 1869, 1869, 1869, 1869, 1869, 1869, 1869, 1869, 1869, 1869, 1869, 1869, 1…
$ Federal_change            <dbl> -83, -83, -83, -83, -83, -83, -83, -83, -83, -83, -83, -83, -83, -83, -83, -83,…
$ CanadaGDP                 <dbl> 1254166, 1254166, 1254166, 1254166, 1254166, 1254166, 1254166, 1254166, 1254166…
$ GDP_whole                 <dbl> 38503, 38503, 38503, 38503, 38503, 38503, 38503, 38503, 38503, 38503, 38503, 38…
$ Rate_GDP_change           <dbl> 3.070008, 3.070008, 3.070008, 3.070008, 3.070008, 3.070008, 3.070008, 3.070008,…
```

Ontario Unemployment Dataset

Can be obtained through:
https://www.dropbox.com/s/y4s1hfv216u8ol4/Veenguyen_STAT3503_ontario_unemployed.csv?dl=0

```
$ Year                      <dbl> 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991…
$ Sex                       <chr> "Females", "Females", "Females", "Females", "Females", "Females", "Females", "Fema…
$ Educational.degree        <chr> "University degree", "University degree", "High school graduate", "Without high sc…
$ Age.group                 <chr> "25 years and over", "25 to 54 years", "15 to 24 years", "25 years and over", "25 …
$ UnempRate                 <dbl> 5.0, 5.0, 12.3, 8.6, 4.2, 5.4, 4.2, 8.4, 6.7, 8.7, 5.9, 6.6, 5.4, 7.7, 8.0, 8.9, 9…
$ Married.15.to.44.years    <dbl> 1247.967, 1247.967, 1247.967, 1247.967, 1247.967, 1247.967, 1247.967, 1247.967, 12…
$ Married.45.to.64.years    <dbl> 744.443, 744.443, 744.443, 744.443, 744.443, 744.443, 744.443, 744.443, 744.443, 7…
$ Divorced.15.to.44.years   <dbl> 95.018, 95.018, 95.018, 95.018, 95.018, 95.018, 95.018, 95.018, 95.018, 95.018, 95…
$ Divorced.45.to.64.years   <dbl> 76.911, 76.911, 76.911, 76.911, 76.911, 76.911, 76.911, 76.911, 76.911, 76.911, 76…
$ Married.All.ages          <dbl> 2280.492, 2280.492, 2280.492, 2280.492, 2280.492, 2280.492, 2280.492, 2280.492, 22…
$ Divorced.All.ages         <dbl> 193.253, 193.253, 193.253, 193.253, 193.253, 193.253, 193.253, 193.253, 193.253, 1…
$ totalPopulation           <dbl> 24837.9, 24837.9, 24837.9, 24837.9, 24837.9, 24837.9, 24837.9, 24837.9, 24837.9, 2…
$ Ontario_change            <dbl> 455.7, 455.7, 455.7, 455.7, 455.7, 455.7, 455.7, 455.7, 455.7, 455.7, 455.7, 455.7…
$ Ontario_rate              <dbl> 1.834696, 1.834696, 1.834696, 1.834696, 1.834696, 1.834696, 1.834696, 1.834696, 1.…
```

Appendix II

---
title: "R Notebook"
output: html_notebook

```{r}
library(RColorBrewer)
library(tidyverse)
library(dplyr)

```



```{r}
temp <- tempfile()
download.file("https://www150.statcan.gc.ca/n1/tbl/csv/14100118-eng.zip",temp)
# download as a temporary file
(file_list <- as.character(unzip (temp, list = TRUE) $Name)) # unzip the file
population <- read_csv(unz(temp, "14100118.csv"))
unlink(temp) # Delete temporary file

population <- population %>%
  drop_na(VALUE) %>%
  rename_all(make.names)

```

Estimates of population as of July 1st, by marital status or legal marital status, age and sex 1, 2, 3
Frequency: Annual

Table: 17-10-0060-01 (formerly CANSIM 051-0042)

Geography: Canada, Province or territory
Martial status, link customized

https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=1710006001

```{r}
martial <- read.csv("https://www.dropbox.com/s/mocpwtrxdqn4zwy/martial_status.csv?dl=1")

martial <- martial %>%
  rename_all(make.names)
```

```
#glimpse(martial)

# martial[, 4:(ncol(martial)]/1000

# martial <- martial %>% select(Year,Divorced,Married)
martial[3:15] <- lapply(martial[3:15], as.numeric)
martial[4:15] <- martial[, 4:15]/1000
glimpse(martial)
# martial$Divorced <- martial$Divorced/1000
# martial$Married <- martial$Married/1000
```

```{r}
temp <- tempfile()
download.file("https://www150.statcan.gc.ca/n1/tbl/csv/36100434-eng.zip",temp)
# download as a temporary file
(file_list <- as.character(unzip(temp, list = TRUE)$Name)) # unzip the file
GDP <- read_csv(unz(temp, "36100434.csv"))
unlink(temp) # Delete temporary file
```

```{r}

GDP <- GDP %>%
  separate(REF_DATE, c("year", "month"), "-") %>%
  mutate(year = as.numeric(year))%>%
  drop_na(VALUE) %>%
  rename_all(make.names)

#names(GDP)[names(GDP) == "North.American.Industry.Classification.System..NAICS."] <-
'Industry.Name'

# mutate(year = as.numeric(year))%>%
#   drop_na(VALUE) %>%
#   rename_all(make.names)

names(GDP)[names(GDP) == "North.American.Industry.Classification.System..NAICS."] <-
'Industry.Name'

Essential_names <- c(
"Goods-producing industries [T002]",
"Service-producing industries [T003]",
"Industrial production [T010]",
```

```
"Non-durable manufacturing industries [T011]",
"Durable manufacturing industries [T012]",
"Agriculture, forestry, fishing and hunting [11]",
"Mining, quarrying, and oil and gas extraction [21]",
"Utilities [22]",
"Construction [23]",
"Public administration [91]",
"Federal government public administration [911]",
"Educational services [61]")


GDP_essential = GDP %>% subset(GEO == "Canada" & month == "12" & Industry.Name
%in% Essential_names & Prices == "2012 constant prices" & Seasonal.adjustment ==
"Seasonally adjusted at annual rates")

GDP_total = GDP %>% subset(GEO == "Canada" & month == "12" & Industry.Name == "All
industries [T001]" & Prices == "2012 constant prices" & Seasonal.adjustment == "Seasonally
adjusted at annual rates")

GDP_whole_growth = GDP_total %>% select(year,Industry.Name,VALUE) %>%
  mutate(GDP_whole = VALUE - lag(VALUE),
         Rate_GDP_change = (GDP_whole /VALUE)*100)

glimpse(GDP_whole_growth)

GDP_industry_rate = GDP_essential %>% select(year,Industry.Name,VALUE) %>%
  pivot_wider(names_from = Industry.Name, values_from = VALUE) %>%
rename_all(make.names)

colname <-
c("Year","GoodProduct","ServiceProduct","IndustrialProduct","NonDurableManu","DurableMa
nu","Arigiculutre","Mining","Utilities","Construction","Educational","PublicAdmin","Federal")

names(GDP_industry_rate) <- colname
glimpse(GDP_industry_rate)


GDP_industry_rate <- GDP_industry_rate %>%
  mutate(GoodProduct_change = GoodProduct - lag(GoodProduct),
         GoodProduct_rate = ((GoodProduct_change / GoodProduct)*100),
         ServiceProduct_change = ServiceProduct - lag(ServiceProduct),
         ServiceProduct_rate = ((ServiceProduct_change /ServiceProduct)*100),
         IndustrialProduct_change = IndustrialProduct - lag(IndustrialProduct),
         IndustrialProduct_rate = ((IndustrialProduct_change /IndustrialProduct)*100),
         NonDurableManu_change = NonDurableManu - lag(NonDurableManu),
         NonDurableManu_rate = ((NonDurableManu_change /NonDurableManu)*100),
```

```
        DurableManu_change = DurableManu - lag(DurableManu),
        DurableManu_rate = ((DurableManu_change /DurableManu)*100),
        Arigiculutre_change = Arigiculutre - lag(Arigiculutre),
        Arigiculutre_rate = ((Arigiculutre_change /Arigiculutre)*100),
        Mining_change = Mining - lag(Arigiculutre),
        Mining_rate = ((Mining_change /Mining)*100),
        Utilities_change = Utilities - lag(Utilities),
        Utilities_rate = ((Utilities_change /Utilities)*100),
        Construction_change = Construction - lag(Construction),
        Construction_rate = ((Construction_change /Construction)*100),
        Educational_change = Educational - lag(Educational),
        Educational_rate = ((Educational_change /Educational)*100),
        PublicAdmin_change = PublicAdmin - lag(PublicAdmin),
        PublicAdmin_rate = ((PublicAdmin_change /PublicAdmin)*100),
        Federal_change = Federal - lag(Federal),
        Federal_rate = ((Federal_change /Federal)*100))

glimpse(GDP_industry_rate)

```

````
```{r}
population <- population %>% rename(Year = REF_DATE)



martial <- martial %>% rename(Year = Reference.period,
                GEO = Geography)
```

```{r}
glimpse(population)
```

```{r}

martial <- martial %>% select(-Total.All.ages,-Total.15.to.44.years,-Total.45.to.64.years,-
Total.65.years.and.over)
glimpse(martial)
```

```{r}
# names(GDP_industry_rate)
````
````

```
essential_economy_growth <- GDP_industry_rate %>% select(Year,GoodProduct_change,
ServiceProduct_change,
DurableManu_change,
Arigiculutre_change,
Mining_change,
Utilities_change,
Construction_change,
Educational_change,
PublicAdmin_change,
Federal_change)
essential_economy_growth %>% glimpse
```

```{r}
mergeCols <- c("Year","GEO","Sex")

full_dataset <- merge(population,martial,
            by = mergeCols)
glimpse(full_dataset)
```

```{r}
population_dataset <- full_dataset %>% subset(GEO != "Canada" & Labour.force.characteristics
== "Population" & Sex != "Both sexes" &  Educational.degree != "Total, all education levels" )



total_population_dataset <- full_dataset %>% subset(Labour.force.characteristics ==
"Population" & Sex == "Both sexes" &  Educational.degree == "Total, all education levels" )


population_dataset <- population_dataset %>% rename(Population = VALUE)


population_dataset  <-population_dataset  %>%
select(Year,GEO,Sex,Educational.degree,Age.group,Population,Married.All.ages,Married.15.to.
44.years,
Married.45.to.64.years,Married.65.years.and.over,Divorced.All.ages,Divorced.15.to.44.years,Di
vorced.45.to.64.years,Divorced.65.years.and.over)

population_dataset  %>% glimpse
```

```{r}
labour_dataset <- full_dataset %>% subset(GEO != "Canada" & Labour.force.characteristics == "Labour force" & Sex != "Both sexes" &  Educational.degree != "Total, all education levels" )
#
labour_dataset <- labour_dataset %>%
select(Year,GEO,Sex,Educational.degree,Age.group,VALUE)
labour_dataset <- labour_dataset %>% rename(LabourForce = VALUE)
labour_dataset %>% glimpse

```


```{r}
population_dataset %>% glimpse
```



```{r}
in_model_dataset <- merge(population_dataset,labour_dataset,
          by = c("Year","GEO","Sex","Educational.degree","Age.group"))

in_model_dataset <- merge(in_model_dataset,essential_economy_growth,by="Year") %>%
drop_na()

#GDP_whole_growth <- GDP_whole_growth %>% rename(Year = year)
#GDP_whole_growth <- GDP_whole_growth %>% rename(CanadaGDP = VALUE)
in_model_dataset <- merge(in_model_dataset,GDP_whole_growth,by="Year") %>% select(-Industry.Name) %>% drop_na()

#in_model_dataset %>% drop_na()

in_model_dataset  %>% glimpse()

write_csv(in_model_dataset,"final_version_dataset_model.csv")
```


```{r}
GDP_whole_growth %>%glimpse

```



```{r}
#ggpairs(in_model_dataset %>% select(-Year,-GEO,-Sex,-Educational.degree,-Age.group))
```

Using DataSplit

```{r}
library(GGally)
#ggpairs(in_model_dataset)
N = in_model_dataset %>% nrow
datasplit = in_model_dataset %>% mutate(train = runif(N)<.5)
datatrain = datasplit %>% filter(train == TRUE)
datatest = datasplit %>% filter(train == FALSE)
```

```{r}
datatrain <- datatrain %>% select(-train)
datatest <- datatest %>% select(-train)
glimpse(datatrain)
```

```{r}
find_corr <- datatrain %>% select(6:28)

round(cor(find_corr),
  digits = 2 # rounded to 2 decimals
)

```

```{r}
#ggduo(datatrain, 1:28, 2, types = list(comboHorizontal = "facehist"))
```

```{r}
#ggcorr(datatrain)

#ggpairs(datatrain %>% select(5:28), diag=list(continuous="density"),axisLabels="show")
```

```{r}
glimpse(datatrain)
```

```{r}
#in_model_dataset <- select(in_model_dataset, -Year)

#dataToTrain <- select(-Year)

dataTrainToUse <- datatrain %>% select(-Year)

dataTestToUse <- datatest %>% select(-Year)



all_in_model <- lm(LabourForce~.,dataTrainToUse)
summary(all_in_model)

main_effect_model <- step(all_in_model)
summary(main_effect_model)

```


```{r}
dataTrainToUse <- datatrain %>% select(-Married.65.years.and.over,-
Divorced.65.years.and.over)

dataTestToUse <- datatest %>% select(-Married.65.years.and.over,-Divorced.65.years.and.over)

all_in_model <- lm(LabourForce~.,dataTrainToUse)
summary(all_in_model)

main_effect_model <- step(all_in_model)
summary(main_effect_model)
```



```{r}
age.mod <- lm(LabourForce~Age.group,dataTrainToUse)
edu.mod <- lm(LabourForce~Educational.degree,dataTrainToUse)
GEO.mod <- lm(LabourForce~GEO,dataTrainToUse)
married.mod <- lm(LabourForce~Married.All.ages,dataTrainToUse)
divorced.mod <- lm(LabourForce~Divorced.All.ages,dataTrainToUse)
pop.mod <- lm(LabourForce~Population,dataTrainToUse)
GDP.mod <- lm(LabourForce~CanadaGDP,dataTrainToUse)
GDP.growth.mod <- lm(log(LabourForce)~Rate_GDP_change,dataTrainToUse)
```

```{r}
```

```
install.packages("AICcmodavg")
library(AICcmodavg)
```

```{r}
models <- list(age.mod, edu.mod, GEO.mod, married.mod, divorced.mod, pop.mod,GDP.mod)

model.names <- c("age.mod", "edu.mod", "GEO.mod", "married.mod", "divorced.mod",
"pop.mod","GDP.mod")
```

```{r}
aictab(cand.set = models, modnames = model.names)
```

```{r}
anova(age.mod, edu.mod, GEO.mod, married.mod, divorced.mod, pop.mod,GDP.mod)
```

```{r}
pop_married.mod <- lm(LabourForce~Population+Married.All.ages,dataTrainToUse)
anova(married.mod,pop.mod,pop_married.mod,all_in_model,main_effect_model)
AIC(married.mod,pop.mod,pop_married.mod,all_in_model,main_effect_model)
```

```{r}
pop.married.mod <- lm(LabourForce~Population+Married.All.ages,dataTrainToUse)
pop.married.divorced.mod <-
lm(LabourForce~Population+Married.All.ages+Divorced.All.ages,dataTrainToUse)

GEO.pop.married.age.divorced.mod <-
lm(LabourForce~GEO+Age.group+Population+Married.All.ages+Divorced.All.ages,dataTrainT
oUse)

GEO.pop.married.age.edu.divorced.mod <-
lm(LabourForce~GEO+Age.group+Population+Educational.degree+Married.All.ages+Divorced.
All.ages,dataTrainToUse)

AIC(pop.married.mod,pop.mod,pop.married.divorced.mod,main_effect_model,GEO.pop.marrie
d.age.divorced.mod,GEO.pop.married.age.edu.divorced.mod)
```

```{r}
```

```
anova(GEO.pop.married.age.edu.divorced.mod,main_effect_model)
AIC(GEO.pop.married.age.edu.divorced.mod,main_effect_model)
```

Is it significant to add GDP to the eqution

```{r}

GDP_add <- lm(LabourForce ~ GEO + Age.group + Population + Educational.degree +
   Married.All.ages + Divorced.All.ages + CanadaGDP,dataTrainToUse)

anova(GEO.pop.married.age.edu.divorced.mod,main_effect_model,GDP_add)
AIC(GEO.pop.married.age.edu.divorced.mod,main_effect_model,GDP_add)
```

```{r}

par(mfrow=c(2,2)) # Change the panel layout to 2 x 2
plot(GEO.pop.married.age.edu.divorced.mod)
par(mfrow=c(1,1)) # Change back to 1 x 1

```
```{r}
summary(GEO.pop.married.age.edu.divorced.mod)
```

```{r}
summary(lm(formula = LabourForce ~ GEO + Age.group + Population + Educational.degree +
   Married.All.ages + Divorced.All.ages, data = dataTrainToUse))

```

```{r}
log_best <- lm(formula = log(LabourForce) ~ GEO + Age.group + log(Population) +
Educational.degree +
   log(Married.All.ages) + log(Divorced.All.ages), data = dataTrainToUse)

summary(log_best)

```

```{r}
par(mfrow=c(2,2)) # Change the panel layout to 2 x 2
plot(log_best)
```

```
par(mfrow=c(1,1)) # Change back to 1 x 1
```

```{r}
log.pop.married.divorced.mod <-
lm(log(LabourForce)~log(Population)+log(Married.All.ages)+log(Divorced.All.ages),dataTrain
ToUse)

log.GEO.pop.married.divorced.mod <-
lm(log(LabourForce)~GEO+log(Population)+log(Married.All.ages)+log(Divorced.All.ages),data
TrainToUse)

log.GEO.pop.married.age.divorced.mod <-
lm(log(LabourForce)~GEO+Age.group+log(Population)+log(Married.All.ages)+log(Divorced.A
ll.ages),dataTrainToUse)


log.GEO.pop.married.age.edu.divorced.mod <-
lm(log(LabourForce)~GEO+Age.group+log(Population)+log(Married.All.ages)+log(Divorced.A
ll.ages)+Educational.degree,dataTrainToUse)

log.GEO.pop.married.age.edu.divorced.GDP.mod <-
lm(log(LabourForce)~GEO+Age.group+log(Population)+log(Married.All.ages)+log(Divorced.A
ll.ages)+log(CanadaGDP)+Educational.degree,dataTrainToUse)

AIC(log.pop.married.divorced.mod,log.GEO.pop.married.divorced.mod,log.GEO.pop.married.a
ge.divorced.mod,log.GEO.pop.married.age.edu.divorced.mod,log.GEO.pop.married.age.edu.div
orced.GDP.mod)

```

```{r}
AIC(log.GEO.pop.married.age.edu.divorced.GDP.mod,log_best)
anova(log.GEO.pop.married.age.edu.divorced.GDP.mod,log_best)
```

```{r}
summary(log_best)
```

```{r}
final_model_labour <- lm(formula = log(LabourForce) ~ GEO + Age.group + log(Population) +
    Educational.degree + log(Married.All.ages) + log(Divorced.All.ages),
    data = dataTestToUse)
```

```

```{r}
par(mfrow=c(2,2)) # Change the panel layout to 2 x 2
plot(lm(formula = LabourForce ~ GEO + Age.group + Population +
    Educational.degree + Married.All.ages + Divorced.All.ages,
    data = dataTestToUse))
par(mfrow=c(1,1)) # Change back to 1 x 1


```

```{r}
summary(lm(formula = LabourForce ~ GEO + Age.group + Population +
    Educational.degree + Married.All.ages + Divorced.All.ages,
    data = dataTestToUse))
```


```{r}
final_test_model_labour <- lm(formula = log(LabourForce) ~ GEO + Age.group + Population +
    Educational.degree + Married.All.ages + Divorced.All.ages,
    data = dataTestToUse)

par(mfrow=c(2,2)) # Change the panel layout to 2 x 2
plot(final_test_model_labour)
par(mfrow=c(1,1)) # Change back to 1 x 1
```
```{r}
summary(final_test_model_labour)
```

```{r}
summary(final_model_labour)
```

```{r}
anova(final_model_labour)
```

```{r}
par(mfrow=c(2,2)) # Change the panel layout to 2 x 2
plot(final_model_labour)
par(mfrow=c(1,1)) # Change back to 1 x 1
```
```

```{r}
confint(final_model_labour)
```

```{r}
summary(final_model_labour)
```

```{r}
AIC(log.GEO.pop.married.age.divorced.mod,log.GEO.pop.married.age.edu.divorced.mod,log.GEO.pop.married.age.edu.divorced.GDP.mod)

# data_mut <- datatrain
#
# data_mut = data_mut %>% mutate(pred1 = predict(log.GEO.pop.married.age.divorced.mod),
#                                pred2 = predict(log.GEO.pop.married.age.edu.divorced.mod),
#                                logLabour = log(LabourForce))
# data_mut %>% ggplot(aes(x=Year))+
#   geom_line(aes(y=pred1),colour= "red",lwd=1.25)+
#   geom_line(aes(y=pred2),colour= "blue",lwd=1)
  #geom_point(aes(y=logLabour))

dataTestOkModel1 <-
lm(lm(log(LabourForce)~GEO+Age.group+log(Population)+log(Married.All.ages)+log(Divorced.All.ages)+Educational.degree,dataTrainToUse),dataTestToUse)

dataTestOkModel2 <- lm(formula = log(LabourForce) ~ GEO + Age.group + log(Population) +
    log(Married.All.ages) + log(Divorced.All.ages) + log(CanadaGDP) +
    Educational.degree, data = dataTestToUse)
summary(dataTestOkModel1)
summary(dataTestOkModel2)
```

```{r}
library(dplyr)
library(broom)

glance(dataTestOkModel1) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)

glance(dataTestOkModel2) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```{r}
sigma(dataTestOkModel1)/mean(dataTrainToUse$LabourForce)


sigma(dataTestOkModel2)/mean(dataTrainToUse$LabourForce)
```

Note that, the RMSE and the RSE are measured in the same scale as the outcome variable. Dividing the RSE by the average value of the outcome variable will give you the prediction error rate, which should be as small as possible:

http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/

```{r}
anova(dataTestOkModel1,dataTestOkModel2)
```

```{r}
plot(dataTestOkModel2)
```

```{r}
confint(dataTestOkModel1)
confint(dataTestOkModel2)
```

```{r}
AIC(dataTestOkModel)
```

```{r}
summary(dataTestOkModel)
```

```{r}

# datatrain %>%
#   ggplot(aes(LabourForce,colour = Age.group)) +
#   geom_boxplot()


datatrain %>%
```

```
  ggplot(aes(LabourForce,colour = Educational.degree)) +
  geom_boxplot()


# datatrain %>%
#   ggplot(aes(y=LabourForce,colour = Age.group, x = Year)) +
#   geom_point()
```

```{r}
datatrain %>%
  ggplot(aes(LabourForce,colour = Educational.degree)) +
  geom_boxplot()
```

```{r}
datatrain %>%
  ggplot(aes(LabourForce,colour = GEO)) +
  geom_boxplot()
```

```{r}
ggplot(data = dataTrainToUse, aes(Population)) + geom_histogram()
ggplot(data = dataTrainToUse, aes(Population,Married.All.ages)) + geom_point()
```

```{r}
ggplot(data = dataTrainToUse, aes(x = Population, y = Married.All.ages)) +
  geom_point() +
  scale_x_log10() + scale_y_log10()
```

```{r}
full_dataset%>% subset(GEO != "Canada" & Labour.force.characteristics == "Labour force" &
Educational.degree == "Total, all education levels" & Sex == "Both sexes") %>%
        group_by(Year, GEO) %>%
        summarize(total = sum(VALUE)) %>%
        ungroup() %>%
```

```
ggplot(aes(x = Year, y = total, group = GEO)) +
  labs(y = "Population (thounsand)") +
  geom_line(aes(colour = GEO)) +
  scale_color_brewer(palette = "Paired")+
  ggtitle("Total Labour Force, by year, seperated by Province")
```
```{r}
ggduo(datatrain, 1:10, 4, types = list(discrete = "ratio",comboVertical = "facethist"))
```

```{r}
datatrain %>% glimpse
```

```{r,fig.width=10, fig.height=10, fig.fullwidth=TRUE}
ggpairs(datatrain[,6:32])
```

Testing Categorical value

Age Group
Sex
Educational.degree
GEO

```{r,fig.width=10, fig.height=10, fig.fullwidth=TRUE}
ggduo(datatrain, 3:6, 2, types = list(discrete = "facetbar",comboHorizontal = "facetdensity"))
```

```{r,fig.width=10, fig.height=2, fig.fullwidth=TRUE}
ggduo(datatrain, 2:6, 3, types = list(discrete = "facetbar",comboHorizontal = "facetdensity"))
```
```{r,fig.width=10, fig.height=10, fig.fullwidth=TRUE}
ggduo(datatrain, 2:6, 4, types = list(discrete = "facetbar",comboHorizontal = "facetdensity"))
```
```{r,fig.width=10, fig.height=10, fig.fullwidth=TRUE}
ggduo(datatrain, 2:6, 5, types = list(discrete = "facetbar",comboHorizontal = "facetdensity"))
```

INVESTIAGE ON ONTARIO EMPLOYBILITY PERCENTAGE

```{r}
`%notin%` <- Negate(`%in%`)
Ontario_dataset <- full_dataset %>% subset(GEO == "Ontario" & Labour.force.characteristics %in% c("Participation rate","Employment rate","Unemployment rate") & Sex != "Both sexes" & Age.group %notin% c("55 years and over","15 years and over","55 to 64 years","65 years and over") & Educational.degree != "Total, all education levels" )

#glimpse(Ontario_dataset)

Ontario_dataset <- Ontario_dataset %>% select(Year,Labour.force.characteristics,GEO,Sex,Educational.degree,Age.group,VALUE,Married.15.to.44.years,Married.45.to.64.years,Divorced.15.to.44.years,Divorced.45.to.64.years,Married.All.ages,Divorced.All.ages) %>% rename(Rate = VALUE)

glimpse(Ontario_dataset)


```

```{r}
# a <- Ontario_dataset %>%
#   group_by(Year,Labour.force.characteristics,GEO,Sex,Educational.degree,Age.group) %>%
#   summarise(total_young_marr = sum(Married.15.to.44.years),
#             total_young_divorced = sum(Divorced.15.to.44.years),
#   mutate(young_marriage_change = Married.15.to.44.years - lag(Married.15.to.44.years),
#                                  young_marriage_rate = (young_marriage_change/Married.15.to.44.years)*100,
#                                  old_marriage_change = Married.45.to.64.years - lag(Married.45.to.64.years),
#                                  old_marriage_rate = (old_marriage_change/Married.45.to.64.years)*100,
#                                  young_divorced_change = Divorced.15.to.44.years - lag(Divorced.15.to.44.years),
#                                  young_divorced_rate = (young_divorced_change/Divorced.15.to.44.years)*100,
#                                  old_divorced_change = Divorced.45.to.64.years - lag(Divorced.45.to.64.years),
#                                  old_divorced_rate = (old_divorced_change/Divorced.45.to.64.years)*100) %>% drop_na()
#
# a %>% glimpse


```

```{r}
Ontario_population_dataset  <- total_population_dataset %>% subset(GEO=="Ontario") %>%
group_by(Year) %>% summarise(totalPopulation=sum(VALUE)) %>% mutate(Ontario_change
= totalPopulation - lag(totalPopulation), Ontario_rate = (Ontario_change/totalPopulation)*100)
%>% drop_na()


glimpse(Ontario_population_dataset)
```

```{r}
dataOntario <- merge(Ontario_dataset,Ontario_population_dataset,by="Year")
glimpse(dataOntario)
```

```{r}
unemployed <- dataOntario %>% subset(Labour.force.characteristics == "Unemployment rate")
%>% select(-Labour.force.characteristics) %>% rename(UnempRate = Rate)

Ontario.unemp.row <- unemployed %>% nrow
datasplitOntario = unemployed %>% mutate(train = runif(Ontario.unemp.row)<.5)
datatrainOntario = datasplitOntario %>% filter(train == TRUE) %>% select(-train,-GEO) %>%
drop_na()
datatestOntario = datasplitOntario%>% filter(train == FALSE) %>% select(-train,-GEO)%>%
drop_na()




```

```{r}
unemployed1 <- dataOntario %>% subset(Labour.force.characteristics == "Unemployment
rate") %>% select(-Labour.force.characteristics,-GEO) %>% rename(UnempRate = Rate)

write_csv(unemployed1,"ontario_unemployed.csv")
```


```{r}
glimpse(datatrainOntario)
```

```{r}
all_Ontario <- lm(UnempRate~. -Year,datatrainOntario)
summary(all_Ontario)
```

```{r}
stepmodel = step(all_Ontario)
summary(stepmodel)
```

```{r}
log_mod = lm(formula = UnempRate ~ Sex + Educational.degree + Age.group +
    log(Married.15.to.44.years) + log(Married.45.to.64.years) + log(Divorced.15.to.44.years) +
    log(Divorced.45.to.64.years) + log(Married.All.ages) + log(totalPopulation) +
    Ontario_change + Ontario_rate, data = datatrainOntario)


log_mod = lm(formula = UnempRate ~ Sex + Educational.degree + Age.group +
    log(Married.15.to.44.years) + log(Married.45.to.64.years) + log(Divorced.15.to.44.years) +
    log(Divorced.45.to.64.years) + log(Married.All.ages) + log(totalPopulation) +
    Ontario_change + Ontario_rate, data = datatrainOntario)

AIC(log_mod,stepmodel,all_Ontario)
```
Apply transformation does help
```{r}
glance(log_mod) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)

glance(stepmodel) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```{r,fig.width=10, fig.height=10, fig.fullwidth=TRUE}
ggpairs(datatrainOntario[,5:14])
```
```{r}
anova(log_mod,lm(UnempRate~.+log(Married.All.ages):log(totalPopulation), datatrainOntario))
```

Significant improvement

```{r}
all_interaction <- (lm(formula =  UnempRate ~ Sex + Educational.degree + Age.group +
(log(Married.15.to.44.years) +
    log(Married.45.to.64.years) + log(Divorced.15.to.44.years) +
```

```
   log(Divorced.45.to.64.years) + log(Married.All.ages) + log(totalPopulation) +
   Ontario_change + Ontario_rate)^2,data = datatrainOntario))
```

```{r}
step_interaction = (step(lm(formula =  UnempRate ~ Sex + Educational.degree + Age.group +
(log(Married.15.to.44.years) +
   log(Married.45.to.64.years) + log(Divorced.15.to.44.years) +
   log(Divorced.45.to.64.years) + log(Married.All.ages) + log(totalPopulation) +
   Ontario_change + Ontario_rate)^2,data = datatrainOntario)))
```

UnempRate ~ Year + Sex + Educational.degree + Age.group + Married.15.to.44.years +
   Married.45.to.64.years + Divorced.15.to.44.years + Divorced.45.to.64.years +
   Married.All.ages + Divorced.All.ages + totalPopulation +
   Ontario_change + Ontario_rate + log(Married.All.ages):log(totalPopulation)
```{r}
log_mod_marriage_year <- lm(UnempRate ~ Year + Sex + Educational.degree + Age.group +
Married.15.to.44.years +
   Married.45.to.64.years + Divorced.15.to.44.years + Divorced.45.to.64.years +
   Married.All.ages + Divorced.All.ages + totalPopulation +
   Ontario_change + Ontario_rate + log(Married.All.ages):log(totalPopulation),datatrainOntario)
```

```{r}
AIC(all_interaction,step_interaction,stepmodel,log_mod,log_mod_marriage_year)
```

```{r}
glance(step_interaction) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)

glance(log_mod_marriage_year) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```{r}
anova(log_mod_marriage_year,step_interaction)
```

```{r}
summary(step_interaction)
```

```{r}
test1 <- lm(formula = UnempRate ~ Sex + Educational.degree + Age.group +
```

```
    log(Married.15.to.44.years) + log(Married.45.to.64.years) +
    log(Divorced.15.to.44.years) + log(Divorced.45.to.64.years) +
    log(Married.All.ages) + log(totalPopulation)
    + log(Married.15.to.44.years):log(Married.45.to.64.years) +
    log(Married.15.to.44.years):log(totalPopulation) +
log(Married.15.to.44.years):Ontario_change +
    log(Married.15.to.44.years):Ontario_rate + log(Married.45.to.64.years):log(totalPopulation) +
    log(Married.45.to.64.years):Ontario_change + log(Married.45.to.64.years):Ontario_rate +
    log(Divorced.15.to.44.years):Ontario_change + log(Divorced.15.to.44.years):Ontario_rate +
    log(Divorced.45.to.64.years):log(Married.All.ages) +
log(Divorced.45.to.64.years):log(totalPopulation) +
    log(Married.All.ages):log(totalPopulation) + log(Married.All.ages):Ontario_change +
    log(totalPopulation):Ontario_rate, data = datatrainOntario)

test2 <- lm(formula = UnempRate ~ Sex + Educational.degree + Age.group +
log(Married.45.to.64.years) +
    log(Divorced.15.to.44.years) + log(Divorced.45.to.64.years) +
    log(Married.All.ages)
    + log(Married.15.to.44.years):log(Married.45.to.64.years) + log(totalPopulation) +
    log(Married.15.to.44.years):log(totalPopulation) +
log(Married.15.to.44.years):Ontario_change +
    log(Married.15.to.44.years):Ontario_rate + log(Married.45.to.64.years):log(totalPopulation) +
    log(Married.45.to.64.years):Ontario_change + log(Married.45.to.64.years):Ontario_rate +
    log(Divorced.15.to.44.years):Ontario_change + log(Divorced.15.to.44.years):Ontario_rate +
    log(Divorced.45.to.64.years):log(Married.All.ages) +
log(Divorced.45.to.64.years):log(totalPopulation) +
    log(Married.All.ages):log(totalPopulation) + log(Married.All.ages):Ontario_change +
    log(totalPopulation):Ontario_rate, data = datatrainOntario)

AIC(test1,test2,step_interaction)
```

```{r}
summary(step_interaction)
```

```{r}
final_model <- lm(UnempRate ~ Sex + Educational.degree + Age.group +
log(Married.15.to.44.years) +
    log(Married.45.to.64.years) + log(Divorced.15.to.44.years) +
    log(Divorced.45.to.64.years) + log(Married.All.ages) + log(totalPopulation) +
    Ontario_change + Ontario_rate + log(Married.15.to.44.years):log(Married.45.to.64.years) +
```

```
   log(Married.15.to.44.years):log(totalPopulation) +
log(Married.15.to.44.years):Ontario_change +
   log(Married.15.to.44.years):Ontario_rate + log(Married.45.to.64.years):log(totalPopulation) +
   log(Married.45.to.64.years):Ontario_change + log(Married.45.to.64.years):Ontario_rate +
   log(Divorced.15.to.44.years):Ontario_change + log(Divorced.15.to.44.years):Ontario_rate +
   log(Divorced.45.to.64.years):log(Married.All.ages) +
log(Divorced.45.to.64.years):log(totalPopulation) +
   log(Married.All.ages):log(totalPopulation) + log(Married.All.ages):Ontario_change +
   log(totalPopulation):Ontario_rate,datatestOntario)
```

```{r}
summary(final_model)
```

```{r}
final_model2 <- lm(formula = UnempRate ~ Sex + Educational.degree + Age.group +
log(Married.45.to.64.years) +
   log(Divorced.15.to.44.years) + log(Divorced.45.to.64.years) +
   log(Married.All.ages)
   + log(Married.15.to.44.years):log(Married.45.to.64.years) + log(totalPopulation) +
   log(Married.15.to.44.years):log(totalPopulation) +
log(Married.15.to.44.years):Ontario_change +
   log(Married.15.to.44.years):Ontario_rate + log(Married.45.to.64.years):log(totalPopulation) +
   log(Married.45.to.64.years):Ontario_change + log(Married.45.to.64.years):Ontario_rate +
   log(Divorced.15.to.44.years):Ontario_change + log(Divorced.15.to.44.years):Ontario_rate +
   log(Divorced.45.to.64.years):log(Married.All.ages) +
log(Divorced.45.to.64.years):log(totalPopulation) +
   log(Married.All.ages):log(totalPopulation) + log(Married.All.ages):Ontario_change +
   log(totalPopulation):Ontario_rate, data = datatrainOntario)

summary(final_model2)
```

```{r}
confint(final_model2)
```

```{r}
glance(final_model) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)

glance(final_model2) %>%
  dplyr::select(adj.r.squared, sigma, AIC, BIC, p.value)
```

```{r}
plot(final_modelf)
```

```{r}
plot(final_model)
```

```{r}
(l <- sapply(datatrainOntario, function(x) is.factor(x)))
lapply(unemployed[colnames(unemployed)], unique)
```

```{r}
logfinal1 <-lm(formula = log(UnempRate) ~ Sex + Educational.degree + Age.group +
    log(Married.45.to.64.years) + log(Divorced.15.to.44.years) +
    log(Divorced.45.to.64.years) + log(Married.All.ages) +
log(Married.15.to.44.years):log(Married.45.to.64.years) +
    log(totalPopulation) + log(Married.15.to.44.years):log(totalPopulation) +
    log(Married.15.to.44.years):Ontario_change + log(Married.15.to.44.years):Ontario_rate +
    log(Married.45.to.64.years):log(totalPopulation) +
log(Married.45.to.64.years):Ontario_change +
    log(Married.45.to.64.years):Ontario_rate + log(Divorced.15.to.44.years):Ontario_change +
    log(Divorced.15.to.44.years):Ontario_rate +
log(Divorced.45.to.64.years):log(Married.All.ages) +
    log(Divorced.45.to.64.years):log(totalPopulation) +
log(Married.All.ages):log(totalPopulation) +
    log(Married.All.ages):Ontario_change + log(totalPopulation):Ontario_rate,
    data = datatrainOntario)

summary(logfinal1)
```

```{r}
plot(logfinal1)
```

```{r}
log_final_model2 <- lm(formula = log(UnempRate) ~ Sex + Educational.degree + Age.group +
log(Married.45.to.64.years) +
    log(Divorced.15.to.44.years) + log(Divorced.45.to.64.years) +
```

```
    log(Married.All.ages)
    + log(Married.15.to.44.years):log(Married.45.to.64.years) + log(totalPopulation) +
    log(Married.15.to.44.years):log(totalPopulation) +
log(Married.15.to.44.years):Ontario_change +
    log(Married.15.to.44.years):Ontario_rate + log(Married.45.to.64.years):log(totalPopulation) +
    log(Married.45.to.64.years):Ontario_change + log(Married.45.to.64.years):Ontario_rate +
    log(Divorced.15.to.44.years):Ontario_change + log(Divorced.15.to.44.years):Ontario_rate +
    log(Divorced.45.to.64.years):log(Married.All.ages) +
log(Divorced.45.to.64.years):log(totalPopulation) +
    log(Married.All.ages):log(totalPopulation) + log(Married.All.ages):Ontario_change +
    log(totalPopulation):Ontario_rate, data = datatrainOntario)

summary(log_final_model2)
```

```{r}
plot(log_final_model2)
```
```{r}
 summary(final_model2)$r.squared

 summary(log_final_model2)$r.squared
```

```{r}
par(mfrow=c(2,2)) # Change the panel layout to 2 x 2
plot(final_model)
par(mfrow=c(1,1)) # Change back to 1 x 1

```

```{r}

final_model <- lm(formula = log(UnempRate ~ Sex + Educational.degree + Age.group +
    log(Married.15.to.44.years) + log(Married.45.to.64.years) +
    log(Divorced.15.to.44.years) + log(Divorced.45.to.64.years) +
    log(Married.All.ages) + log(totalPopulation) + Ontario_change +
    Ontario_rate + log(Married.15.to.44.years):log(Married.45.to.64.years) +
    log(Married.15.to.44.years):log(totalPopulation) +
log(Married.15.to.44.years):Ontario_change +
    log(Married.15.to.44.years):Ontario_rate + log(Married.45.to.64.years):log(totalPopulation) +
    log(Married.45.to.64.years):Ontario_change + log(Married.45.to.64.years):Ontario_rate +
    log(Divorced.15.to.44.years):Ontario_change + log(Divorced.15.to.44.years):Ontario_rate +
    log(Divorced.45.to.64.years):log(Married.All.ages) +
log(Divorced.45.to.64.years):log(totalPopulation) +
    log(Married.All.ages):log(totalPopulation) + log(Married.All.ages):Ontario_change +
```

```
   log(totalPopulation):Ontario_rate, data = datatestOntario))

summary(final_model)
```
```{r}
confint(final_model)
```

```{r}
confint(final_model_labour)
```

```{r}
step_final <- (lm(formula = log(UnempRate) ~ Sex + Educational.degree + Age.group +
   log(Married.15.to.44.years) + log(Married.45.to.64.years) +
   log(Divorced.15.to.44.years) + log(Divorced.45.to.64.years) +
   log(Married.All.ages) + log(totalPopulation) + Ontario_change +
   Ontario_rate + log(Married.15.to.44.years):log(Married.45.to.64.years) +
   log(Married.15.to.44.years):log(totalPopulation) +
log(Married.15.to.44.years):Ontario_change +
   log(Married.15.to.44.years):Ontario_rate + log(Married.45.to.64.years):log(totalPopulation) +
   log(Married.45.to.64.years):Ontario_change + log(Married.45.to.64.years):Ontario_rate +
   log(Divorced.15.to.44.years):Ontario_change + log(Divorced.15.to.44.years):Ontario_rate +
   log(Divorced.45.to.64.years):log(Married.All.ages) +
log(Divorced.45.to.64.years):log(totalPopulation) +
   log(Married.All.ages):log(totalPopulation) + log(Married.All.ages):Ontario_change +
   log(totalPopulation):Ontario_rate, data = datatrainOntario))

test1 <- lm(formula = log(UnempRate) ~ Sex + Educational.degree + Age.group +
   log(Married.45.to.64.years) + log(Divorced.15.to.44.years) +
   log(Divorced.45.to.64.years) + log(Married.All.ages) +
log(Married.15.to.44.years):log(Married.45.to.64.years) +
   log(totalPopulation) + log(Married.15.to.44.years):log(totalPopulation) +
   log(Married.15.to.44.years):Ontario_change + log(Married.15.to.44.years):Ontario_rate +
   log(Married.45.to.64.years):log(totalPopulation) +
log(Married.45.to.64.years):Ontario_change +
   log(Married.45.to.64.years):Ontario_rate + log(Divorced.15.to.44.years):Ontario_change +
   log(Divorced.15.to.44.years):Ontario_rate +
log(Divorced.45.to.64.years):log(Married.All.ages) +
   log(Divorced.45.to.64.years):log(totalPopulation) +
log(Married.All.ages):log(totalPopulation) +
   log(Married.All.ages):Ontario_change + log(totalPopulation):Ontario_rate,
   data = datatrainOntario)

AIC(step_final,test1)
anova(step_final,test1)
```

```

```{r}
all_interaction_log <- (lm(formula =  log(UnempRate) ~ Sex + Educational.degree + Age.group
+ (log(Married.15.to.44.years) +
    log(Married.45.to.64.years) + log(Divorced.15.to.44.years) +
    log(Divorced.45.to.64.years) + log(Married.All.ages) + log(totalPopulation) +
    Ontario_change + Ontario_rate)^2,data = datatrainOntario))

summary(all_interaction_log)
```

```{r}
step_interaction_log <- step(all_interaction_log)
summary(step_interaction_log)
```

```{r}
step_model_test <- lm(formula = log(UnempRate) ~ Sex + Educational.degree + Age.group +
    log(Married.15.to.44.years) + log(Married.45.to.64.years) +
    log(Divorced.15.to.44.years) + log(Divorced.45.to.64.years) +
    log(Married.All.ages) + log(totalPopulation) + Ontario_change +
    Ontario_rate + log(Married.15.to.44.years):log(Married.45.to.64.years) +
    log(Married.15.to.44.years):log(Divorced.15.to.44.years) +
    log(Married.15.to.44.years):log(Divorced.45.to.64.years) +
    log(Married.15.to.44.years):log(totalPopulation) +
log(Married.15.to.44.years):Ontario_change +
    log(Married.15.to.44.years):Ontario_rate +
log(Married.45.to.64.years):log(Divorced.15.to.44.years) +
    log(Married.45.to.64.years):log(Divorced.45.to.64.years) +
    log(Married.45.to.64.years):log(Married.All.ages) +
log(Married.45.to.64.years):log(totalPopulation) +
    log(Married.45.to.64.years):Ontario_change + log(Married.45.to.64.years):Ontario_rate +
    log(Divorced.45.to.64.years):log(Married.All.ages) +
log(Divorced.45.to.64.years):log(totalPopulation) +
    log(Married.All.ages):log(totalPopulation) + log(Married.All.ages):Ontario_change +
    log(Married.All.ages):Ontario_rate, data = datatestOntario)
summary(step_model_test)

par(mfrow=c(2,2)) # Change the panel layout to 2 x 2
plot(step_model_test)
par(mfrow=c(1,1)) # Change back to 1 x 1
```

```{r}

```
test <- lm(formula = log(LabourForce) ~ GEO + Age.group + log(Population) +
    Educational.degree + log(Married.All.ages) + log(Divorced.All.ages)+log(CanadaGDP),
    data = dataTestToUse)
anova(test,final_model_labour)
```

```{r}
confint(level=0.90,final_model_labour)
```

```{r}
confint(step_model_test)
```
```{r}
summary(step_model_test)
```

```{r}
test1 <- lm(formula = log(UnempRate) ~  Educational.degree + Age.group +
    log(Married.15.to.44.years) + log(Married.45.to.64.years) +
    log(Divorced.15.to.44.years) + log(Divorced.45.to.64.years) +
    log(Married.All.ages) + log(totalPopulation) + Ontario_change +
    Ontario_rate + log(Married.15.to.44.years):log(Married.45.to.64.years) +
    log(Married.15.to.44.years):log(Divorced.15.to.44.years) +
    log(Married.15.to.44.years):log(Divorced.45.to.64.years) +
    log(Married.15.to.44.years):log(totalPopulation) +
log(Married.15.to.44.years):Ontario_change +
    log(Married.15.to.44.years):Ontario_rate +
log(Married.45.to.64.years):log(Divorced.15.to.44.years) +
    log(Married.45.to.64.years):log(Divorced.45.to.64.years) +
    log(Married.45.to.64.years):log(Married.All.ages) +
log(Married.45.to.64.years):log(totalPopulation) +
    log(Married.45.to.64.years):Ontario_change + log(Married.45.to.64.years):Ontario_rate +
    log(Divorced.45.to.64.years):log(Married.All.ages) +
log(Divorced.45.to.64.years):log(totalPopulation) +
    log(Married.All.ages):log(totalPopulation) + log(Married.All.ages):Ontario_change +
    log(Married.All.ages):Ontario_rate, data = datatestOntario)

anova(test1,step_model_test)
```