UNIVERSITY OF BRITISH COLUMBIA

Department of Science

STAT 306 FINDING RELATIONSHIP IN DATA

TERM PROJECT

Group #18

Mengqi Lu

Catherine Cai

Ricky Xie

Josh Blas

December 2020

1. Introduction

In this project, the data collected includes *average working hours* (in hours), *geography* (provinces in Canada, no units), *sex* (no units) and *age group* (in years). The response variable is the *average working hours* in a week, which is derived by dividing the total actual hours worked by the number of workers who were at work during the reference week, where the reference week is usually the week containing the 15th day of the month. The explanatory variables (predictors) are *geography* (provinces in Canada), *sex* and *age group*. The data was collected from January 1st, 1976 to October 1st, 2020. A survey was targeted at the non-institutionalized population (15 years of age and over) living in all the ten provinces and territories in Canada.

The aim for this project is to explore how different factors of sex, age group and different provinces affect the average working hour in Canada, we will divide it to data visualization and data analysis to discuss.

2. Analysis

a. Data visualization

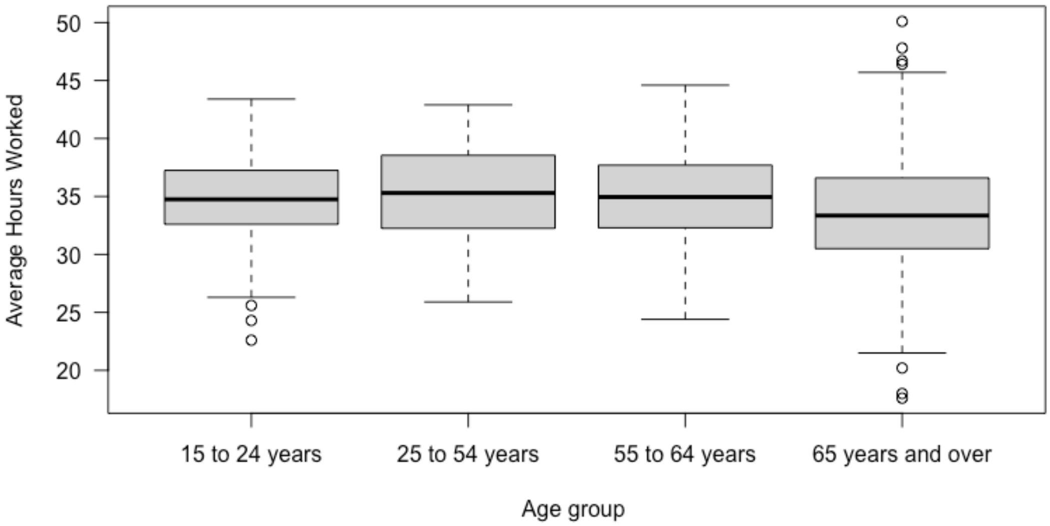
The entire dataset, as proposed in the early submission, was filtered to include only three explanatory variables. The response variable is average full-time working hours for all the months in the data set (January to October 2020). Only 800 observations were left and analyzed. The three explanatory variables are as follows:

Variable name	Level
Province	10 provinces in Canada.
Sex	Male and female
Age group	15-24 ,25-54, 55-64, 65 and over

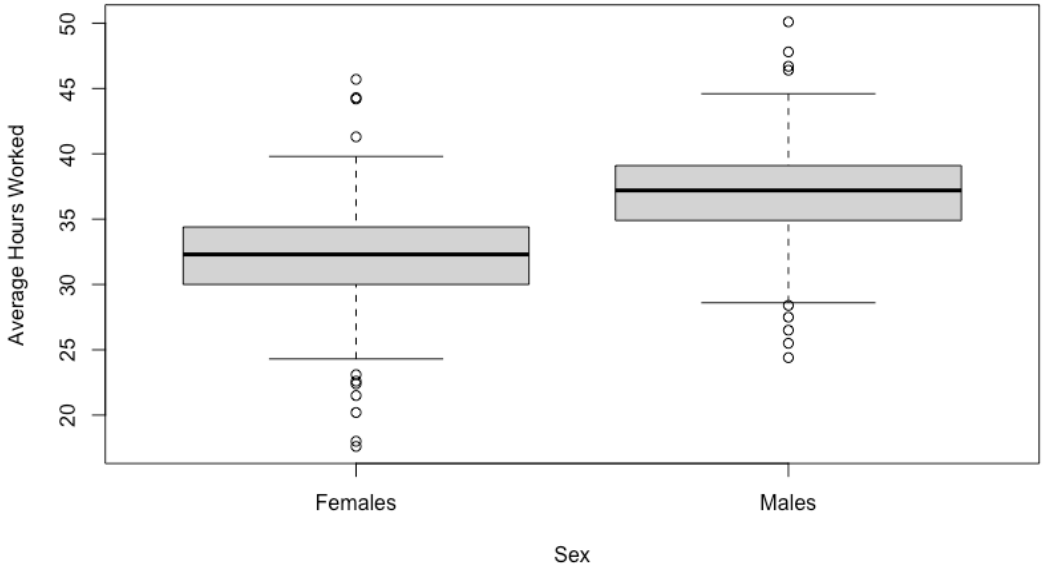
Note that the total 800 observations are equally distributed within each level of predictors. The dataset was then processed with R to visualize some key features.

1) Age group distribution

As indicated below, the mean average working hours for each age group is generally equal. Furthermore, the four age groups are mostly of equal variance. It could also be seen that the oldest group (65 years and over) have lower average working hours with more outliers. This is as expected since people work relatively short in terms of time after they retire.

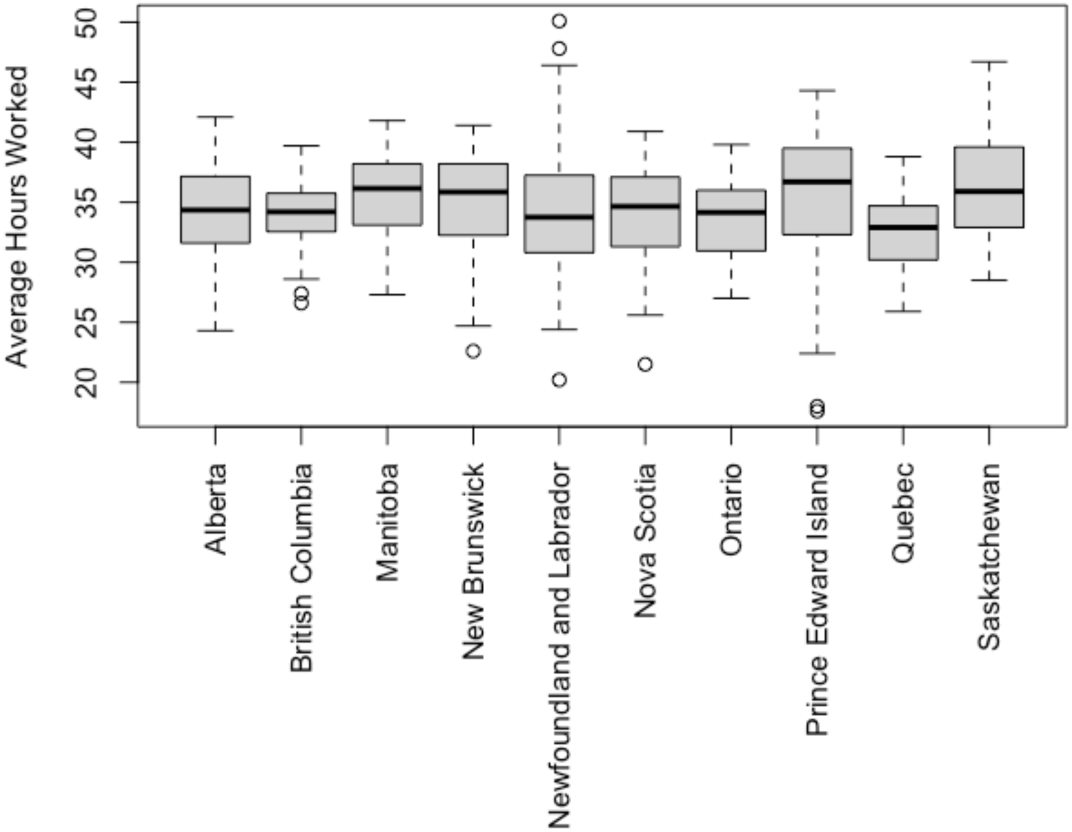


2) Sex group distribution



As illustrated in the above figure, males generally have longer working hours compared to females. However, they are of equal variance. This is also as expected. Also note that both sexes, outliers exist in two directions, which indicates that working hours in each sex group is widely distributed.

3) Province distribution



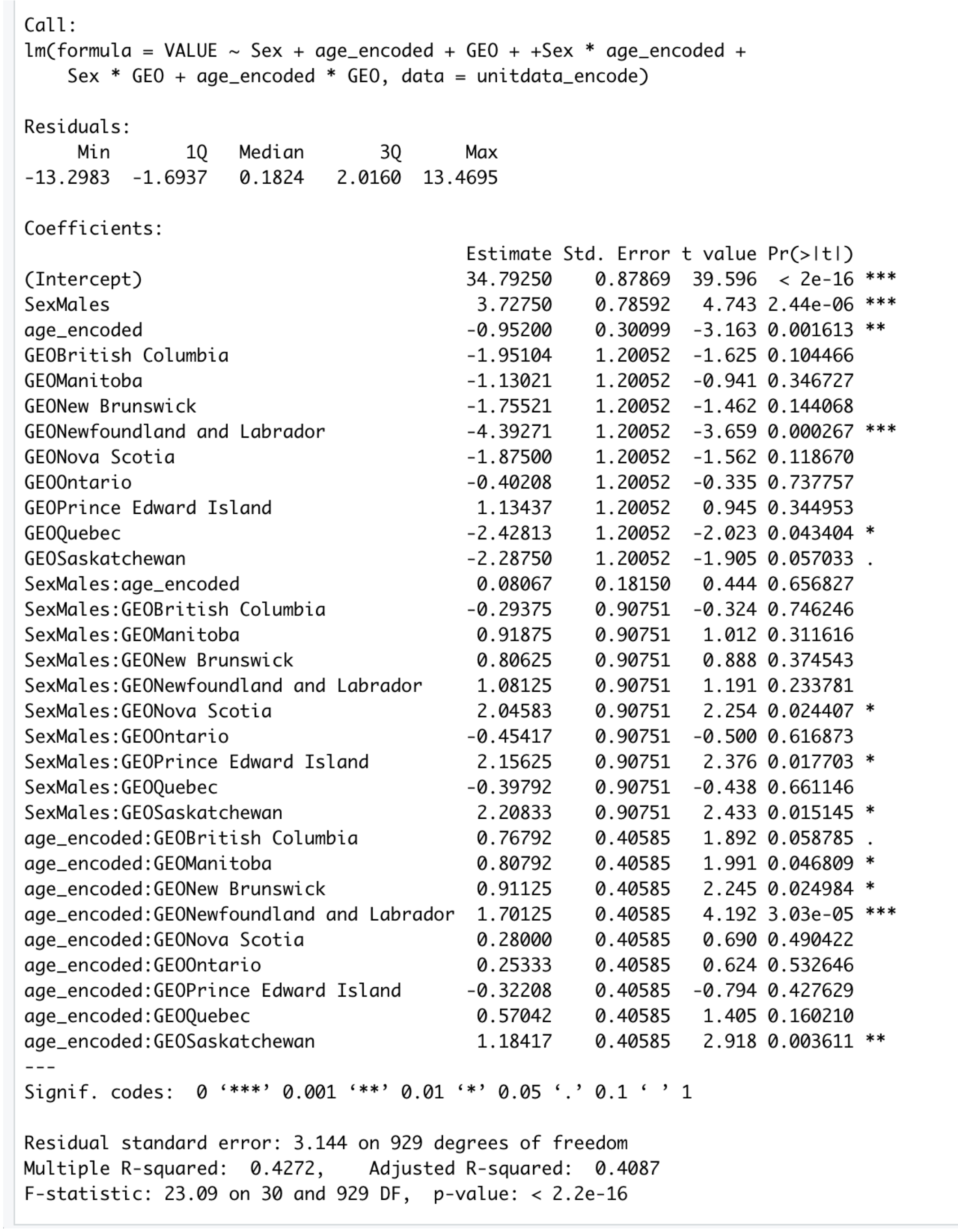
The average working hours for different provinces are generally equal with similar variance. It is also noticed that the distributions of working hours along different provinces are mostly symmetric. Half of the provinces have outliers.

b. Data Analysis

1) Full model

After the data visualization with boxplot, it is figured that the predicted response are significantly different only on sex group, but not in provinces or age groups. However, further analysis is still to be conducted to confirm the result, namely, are age groups and geographic conditions not useful at all when making predictions. In other words, how much of the model could be explained by the three explanatory variables.

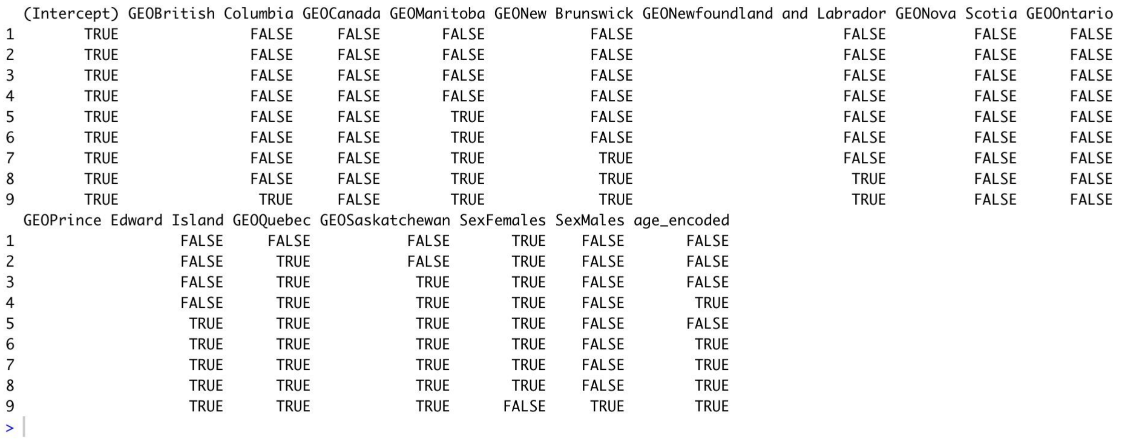
Firstly, the age groups are converted to factors. The full model using sex, age, interaction between sex and age, sex and geo, age and geo has been analyzed using the multi regression model in R. The output is summarized as follows:



```
Call:
lm(formula = VALUE ~ Sex + age_encoded + GEO + +Sex * age_encoded + 
    Sex * GEO + age_encoded * GEO, data = unitdata_encode)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.2983  -1.6937   0.1824   2.0160  13.4695 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              34.79250    0.87869  39.596  < 2e-16 ***
SexMales                                  3.72750    0.78592   4.743 2.44e-06 ***
age_encoded                              -0.95200    0.30099  -3.163 0.001613 ** 
GEOBritish Columbia                      -1.95104    1.20052  -1.625 0.104466    
GEOManitoba                              -1.13021    1.20052  -0.941 0.346727    
GEONew Brunswick                         -1.75521    1.20052  -1.462 0.144068    
GEONewfoundland and Labrador             -4.39271    1.20052  -3.659 0.000267 ***
GEONova Scotia                           -1.87500    1.20052  -1.562 0.118670    
GEOOntario                               -0.40208    1.20052  -0.335 0.737757    
GEOPrince Edward Island                   1.13437    1.20052   0.945 0.344953    
GEOQuebec                                -2.42813    1.20052  -2.023 0.043404 *  
GEOSaskatchewan                          -2.28750    1.20052  -1.905 0.057033 .  
SexMales:age_encoded                      0.08067    0.18150   0.444 0.656827    
SexMales:GEOBritish Columbia             -0.29375    0.90751  -0.324 0.746246    
SexMales:GEOManitoba                      0.91875    0.90751   1.012 0.311616    
SexMales:GEONew Brunswick                 0.80625    0.90751   0.888 0.374543    
SexMales:GEONewfoundland and Labrador     1.08125    0.90751   1.191 0.233781    
SexMales:GEONova Scotia                   2.04583    0.90751   2.254 0.024407 *  
SexMales:GEOOntario                      -0.45417    0.90751  -0.500 0.616873    
SexMales:GEOPrince Edward Island          2.15625    0.90751   2.376 0.017703 *  
SexMales:GEOQuebec                       -0.39792    0.90751  -0.438 0.661146    
SexMales:GEOSaskatchewan                  2.20833    0.90751   2.433 0.015145 *  
age_encoded:GEOBritish Columbia           0.76792    0.40585   1.892 0.058785 .  
age_encoded:GEOManitoba                   0.80792    0.40585   1.991 0.046809 *  
age_encoded:GEONew Brunswick              0.91125    0.40585   2.245 0.024984 *  
age_encoded:GEONewfoundland and Labrador  1.70125    0.40585   4.192 3.03e-05 ***
age_encoded:GEONova Scotia                0.28000    0.40585   0.690 0.490422    
age_encoded:GEOOntario                    0.25333    0.40585   0.624 0.532646    
age_encoded:GEOPrince Edward Island      -0.32208    0.40585  -0.794 0.427629    
age_encoded:GEOQuebec                     0.57042    0.40585   1.405 0.160210    
age_encoded:GEOSaskatchewan               1.18417    0.40585   2.918 0.003611 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.144 on 929 degrees of freedom
Multiple R-squared:  0.4272,	Adjusted R-squared:  0.4087 
F-statistic: 23.09 on 30 and 929 DF,  p-value: < 2.2e-16
```

According to the summary, we see the coefficients on sex, Newfoundland and Labrador as well as the interaction in between between are the most significant terms. Note that since it does not make sense to encode the 10 geographic conditions, the model seems to be redundant.

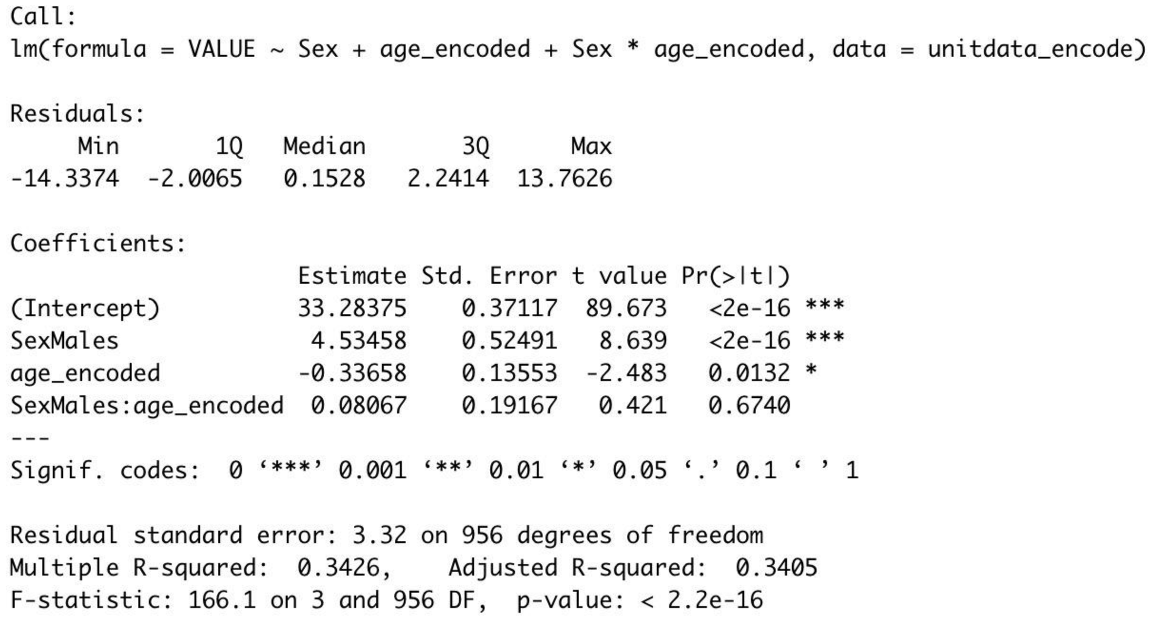
Then the best subset selection algorithm was run for all the ten provinces and sex groups as binary variable along with age group to be encoded.



```
  (Intercept) GEOBritish Columbia GEOCanada GEOManitoba GEONew Brunswick GEONewfoundland and Labrador GEONova Scotia GEOOntario
1        TRUE               FALSE     FALSE       FALSE            FALSE                        FALSE          FALSE      FALSE
2        TRUE               FALSE     FALSE       FALSE            FALSE                        FALSE          FALSE      FALSE
3        TRUE               FALSE     FALSE       FALSE            FALSE                        FALSE          FALSE      FALSE
4        TRUE               FALSE     FALSE       FALSE            FALSE                        FALSE          FALSE      FALSE
5        TRUE               FALSE     FALSE        TRUE            FALSE                        FALSE          FALSE      FALSE
6        TRUE               FALSE     FALSE        TRUE            FALSE                        FALSE          FALSE      FALSE
7        TRUE               FALSE     FALSE        TRUE             TRUE                        FALSE          FALSE      FALSE
8        TRUE               FALSE     FALSE        TRUE             TRUE                         TRUE          FALSE      FALSE
9        TRUE                TRUE     FALSE        TRUE             TRUE                         TRUE          FALSE      FALSE
  GEOPrince Edward Island GEOQuebec GEOSaskatchewan SexFemales SexMales age_encoded
1                   FALSE     FALSE           FALSE       TRUE    FALSE       FALSE
2                   FALSE      TRUE           FALSE       TRUE    FALSE       FALSE
3                   FALSE      TRUE            TRUE       TRUE    FALSE       FALSE
4                   FALSE      TRUE            TRUE       TRUE    FALSE        TRUE
5                    TRUE      TRUE            TRUE       TRUE    FALSE       FALSE
6                    TRUE      TRUE            TRUE       TRUE    FALSE        TRUE
7                    TRUE      TRUE            TRUE       TRUE    FALSE        TRUE
8                    TRUE      TRUE            TRUE       TRUE    FALSE        TRUE
9                    TRUE      TRUE            TRUE      FALSE     TRUE        TRUE
> 
```

As indicated in the summary above, when only two variables are selected, intercept and sex are to be chosen. It is still quite confusing to include all the geographic provinces. As such, a reduced model with only sex, age and interaction between sex and age is evaluated with the result output shown below.

2) Reduced model

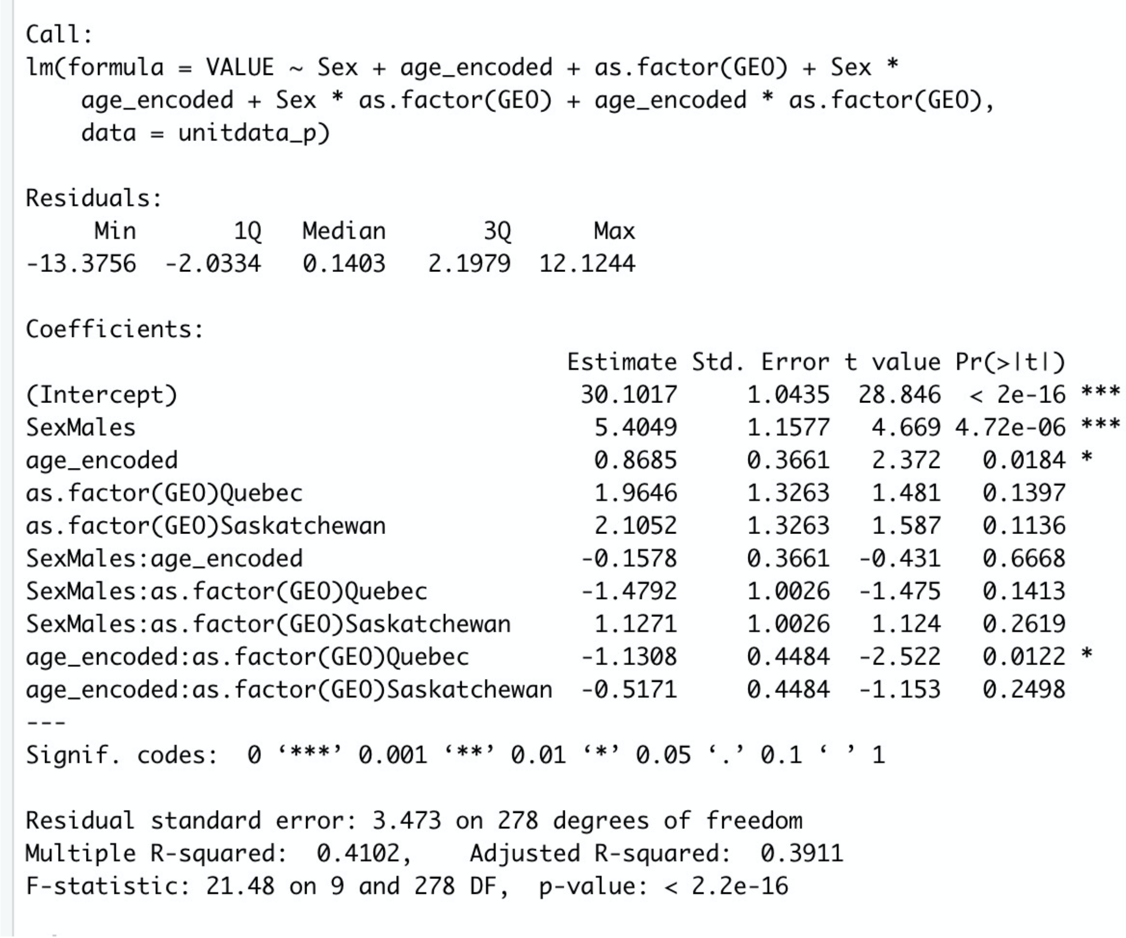


```
Call:
lm(formula = VALUE ~ Sex + age_encoded + Sex * age_encoded, data = unitdata_encode)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.3374  -2.0065   0.1528   2.2414  13.7626 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          33.28375    0.37117  89.673   <2e-16 ***
SexMales              4.53458    0.52491   8.639   <2e-16 ***
age_encoded          -0.33658    0.13553  -2.483   0.0132 *  
SexMales:age_encoded  0.08067    0.19167   0.421   0.6740    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.32 on 956 degrees of freedom
Multiple R-squared:  0.3426,	Adjusted R-squared:  0.3405 
F-statistic: 166.1 on 3 and 956 DF,  p-value: < 2.2e-16
```

For this summary, we see a significant coefficient in sex, age_encoded, with the p-value less than 0.05. However, the adjusted R-squared has dropped significantly from 0.4087 to 0.3405. It could not be considered as a better model compared to the full one.

3) Adjusted-province model

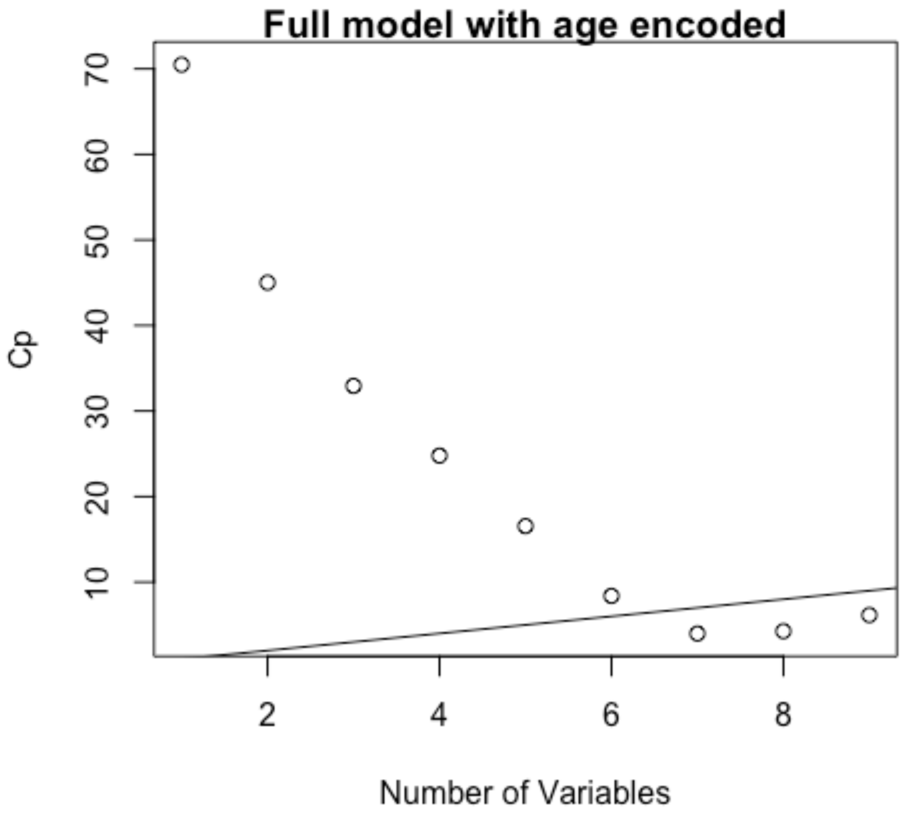
It has been discovered from the previous reduced model that simply discarding the ten provinces will decrease the adjusted R-squared remarkably. As a result, the group decided to include some of the provinces to see if the new model could better explain the predicted value of working hours. The dataset was then filtered to only three provinces, Saskatchewan, Quebec and Newfoundland, the three variables with lowest p-value indicated in the full model. Then a new linear model was fit and summarized below.



```
Call:
lm(formula = VALUE ~ Sex + age_encoded + as.factor(GEO) + Sex *
    age_encoded + Sex * as.factor(GEO) + age_encoded * as.factor(GEO),
    data = unitdata_p)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3756  -2.0334   0.1403   2.1979  12.1244

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                            30.1017     1.0435  28.846  < 2e-16 ***
SexMales                                5.4049     1.1577   4.669 4.72e-06 ***
age_encoded                             0.8685     0.3661   2.372   0.0184 *
as.factor(GEO)Quebec                    1.9646     1.3263   1.481   0.1397
as.factor(GEO)Saskatchewan              2.1052     1.3263   1.587   0.1136
SexMales:age_encoded                   -0.1578     0.3661  -0.431   0.6668
SexMales:as.factor(GEO)Quebec          -1.4792     1.0026  -1.475   0.1413
SexMales:as.factor(GEO)Saskatchewan     1.1271     1.0026   1.124   0.2619
age_encoded:as.factor(GEO)Quebec       -1.1308     0.4484  -2.522   0.0122 *
age_encoded:as.factor(GEO)Saskatchewan -0.5171     0.4484  -1.153   0.2498
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.473 on 278 degrees of freedom
Multiple R-squared:  0.4102,	Adjusted R-squared:  0.3911
F-statistic: 21.48 on 9 and 278 DF,  p-value: < 2.2e-16
```

As shown above, the model with only three provinces is better than the reduced one. However, the adjusted R-squared is still smaller than the original full model. Also, the p-value for most of the geographic variables are not significant. As a result, according to the adjusted R-squared and p-value, the best model of all is the original full model.

4) Comparison between three models



It could be clearly seen in the Mallow's Cp statistics figure, the model with seven variables is the best. In fact, the model with seven, eight or nine explanatory variables are all good to select due to the fact that their Cp values are smaller but also close to the p. This confirms that the full model is better in explaining the response working hours.

To further compare between the three models, root mean square error has been calculated for the actual and predicted response variable of working hours. In all models, 75% of the data were used to train and the rest was left to test the difference. The full and reduced model have RMSE of 0.056 and 0.071, respectively. The model with three provinces included, on the other hand, has a RMSE of 0.670, which is quite large compared to the first two. The RMSE value for the three models further confirmed the best model to be the full model.

3. Conclusion

Based on the previous analysis of the dataset, we could conclude that although the age and geographic province variables are not quite significant in predicting the response variable of working hours, it is still a better model to have all of them into the model instead of just having the most significant ones. The full model with all explanatory variables included is the best by all mean, as was discussed in the previous analysi