Analysing the Salary Difference Between Genders

Shuk Yin Chung (1003384964)

December 10, 2020

Github repo link: https://github.com/Shuk-Yin-Chung/STA304-Final-Project.git

Abstrat

This report aims to investigate the salary difference and gender inequality between males and females in Canada, according to ages, education level, and weekly working hours. By using a multiple linear regression model, it shows that males' salaries are generally higher than females' salaries.

Keywords: Salary Difference, Gender Inequality, Ages, Education Level, Weekly Working Hours, Linear Regression Model

Introduction

Nowadays, gender discrimination is a global issue. It is often said that male earns higher salary than female, although they may have the same working hours and other characteristics etc. Then, it raises the public concern of salary inequality between males and females as it violates human rights and leads to other social problems (such as higher crime rates or poor health in females).

It is important to make an inference between salary and gender. In this report, it aims to investigate the salary gap between Canadian males and females based on their weekly working hours, ages, and education level. Salary is the response variable, whereas sex, age, education level, and weekly working hours are the predict variables.

The data is derived from 2017 Canadian Income Survey for analyzing if there is salary inequality between genders in Canada. The size of data set is 49295 after the cleaning-up. In the methodology section, a multilevel linear regression model is used to simulate the data and analyze the salary difference. Then, in the result section, it will include plots and quantitative analysis and indicate the result of the model. Finally, in the discussion section, it concludes the report and discusses the weakness and future steps of this project.

Methodology

Data

The size of the original data is 49295. For the sake of concise modeling, this report would randomly select 1000 samples from the data to perform the model. Here is the table providing baseline characteristics of the data:

```
## Rows: 1,000
## Columns: 6
## $ CASEID                <int> 19219, 6454, 25215, 5776, 20941, 11407, 38179,...
## $ age                  <chr> "40-49", "50-59", "40-49", "30-39", "40-49", "...
## $ sex                  <chr> "Female", "Male", "Female", "Female", "Female"...
## $ education            <chr> "Non-university postsecondary certificate or d...
## $ weekly_working_hours <dbl> 37.5, 40.0, 37.5, 40.0, 40.0, 80.0, 2.5, 35.0,...
## $ salary               <int> 22000, 38000, 60000, 70000, 27000, 160000, 0, ...
```

The tables shows the variables and the types of variables. CASEID and salary are which stands for integer. Working hours is which means double Both integer and double are numeric values. Sex and education are that stands for character. They are strings of letters. Age is splitted into groups, so it is a character and a string of numbers.

Model

This report uses a multiple linear regression model because it models the correlation between salary and certain predictors. Salary is the response variable because the goal of this report is to find out the salary difference. Sex, age, education level, and weekly working hours are the predict variables that affects the amount of salary. The equation of the full model is:

$$Salary = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{age} + \beta_3 x_{education} + \beta_4 x_{weekly_{w}okring_{h}ours} + e$$

where β_0 is the intercept of the model, β_1,…,β_4 are the estimated regression coefficients of particular predictors, and e is the error terms.

However, sex, age, and education level are categorical variables in the survey. So, this report uses dummy variable coding that alters the responses to binary. A new variable x_{male} replaces x_{sex}. Variables of $x_{age_{20-29}}$,$x_{age_{30-39}}$, $x_{age_{40-49}}$, $x_{age_{50-59}}$, $x_{age_{60-69}}$, and $x_{age_{70+}}$ represent different age groups. Also, variables $x_{education_{Lessthanhighschoolgraduation}}$,$x_{education_{Non-universitypostsecondarycertificateordiploma}}$, and $x_{education_{Universitydegreeorcertificate}}$ represent education levels. Then, the full model can be expanded into:

$$Salary = \beta_0 + \beta_1 x_{male} + \beta_2 x_{age_{20-29}} + \beta_3 x_{age_{30-39}} + \beta_4 x_{age_{40-49}} + \beta_5 x_{age_{50-59}} + \beta_6 x_{age_{60-69}} + \beta_7 x_{age_{70+}}$$

$$+\beta_8 x_{education_{Lessthanhighschoolgraduation}} + \beta_9 x_{education_{Non-universitypostsecondarycertificateordiploma}}$$

$$+\beta_{10} x_{education_{Universitydegreeorcertificate}} + \beta_{11} x_{weeklywokringhours} + e$$

If the respondent is male, x_{male} equals to 1. Otherwise, x_{male} equals 0. This logic can also be applied to the age groups and education levels.

To predict the salary difference, it is useful to create subsets for males and females. Then, the model for males is:

$$Salary_M = \beta_0 + \beta_1 + \beta_2 x_{age_{20-29}} + \beta_3 x_{age_{30-39}} + \beta_4 x_{age_{40-49}} + \beta_5 x_{age_{50-59}} + \beta_6 x_{age_{60-69}} + \beta_7 x_{age_{70+}}$$

$$+\beta_8 x_{education_{Lessthanhighschoolgraduation}} + \beta_9 x_{education_{Non-universitypostsecondarycertificateordiploma}}$$

$$+\beta_1 0 x_{education_{Universitydegreeorcertificate}} + \beta_1 1 x_{weekly_{w}okring_{h}ours} + e$$

The model for females is:

$$Salary_F = \beta_0 + \beta_2 x_{age_{20-29}} + \beta_3 x_{age_{30-39}} + \beta_4 x_{age_{40-49}} + \beta_5 x_{age_{50-59}} + \beta_6 x_{age_{60-69}} + \beta_7 x_{age_{70+}}$$

$$+\beta_8 x_{education_{Lessthanhighschoolgraduation}} + \beta_9 x_{education_{Non-universitypostsecondarycertificateordiploma}}$$

$$+\beta_1 0 x_{education_{Universitydegreeorcertificate}} + \beta_1 1 x_{weekly_{w}okring_{h}ours} + e$$

Results

Summary of males' weekly working hours and salaries:

```
##  weekly_working_hours     salary
##  Min.   :  4.0        Min.   :     0
##  1st Qu.: 37.5        1st Qu.: 16000
##  Median : 40.0        Median : 44000
##  Mean   : 39.8        Mean   : 55728
##  3rd Qu.: 44.0        3rd Qu.: 76250
##  Max.   :105.0        Max.   :750000
```

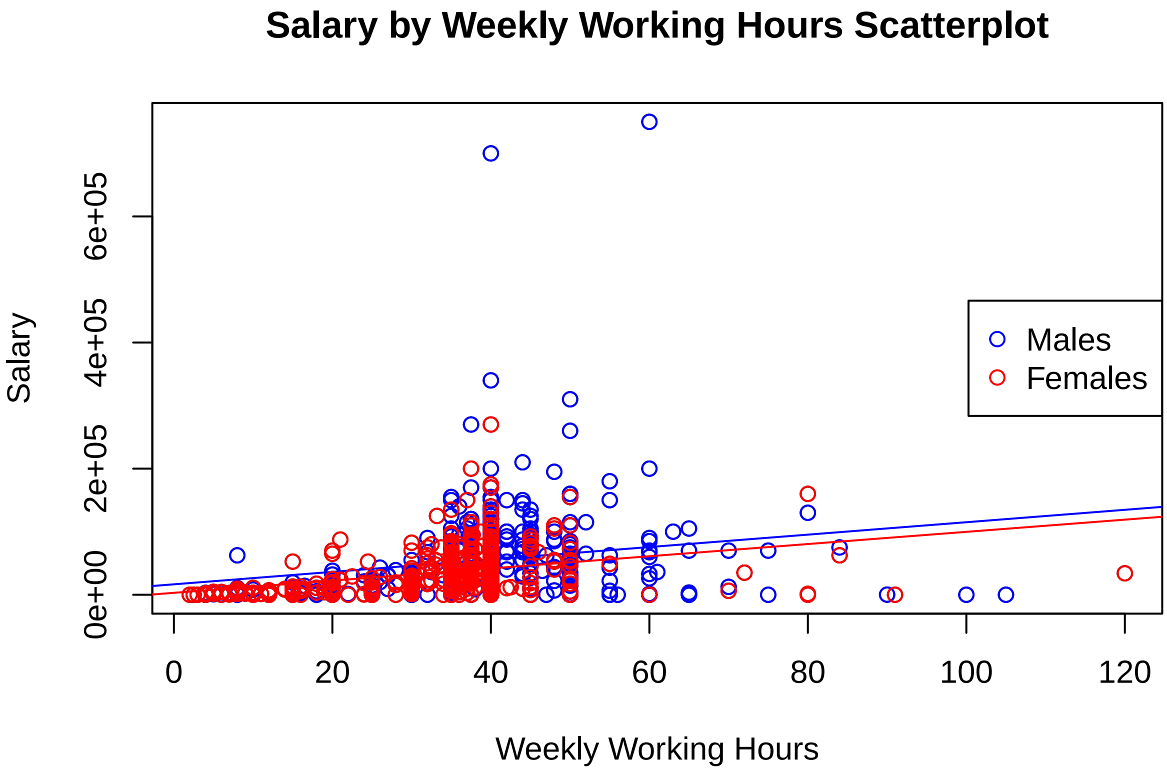
The weekly working hours of males vary between 4 to 105 hours. In average, males work 39.8 hours per week. The maximum salary for males is \$750000 per week, and the average salary is \$55728 per week.

Summary of females' weekly working hours and salaries:

```
##  weekly_working_hours     salary
##  Min.   :  2.00       Min.   :     0
##  1st Qu.: 30.00       1st Qu.:  8250
##  Median : 37.50       Median : 28000
##  Mean   : 34.16       Mean   : 36075
##  3rd Qu.: 40.00       3rd Qu.: 55000
##  Max.   :120.00       Max.   :270000
```

Females' weekly working hours vary between 2 to 102 hours, and the average is 34.16 working hours. Although females' average working hours per week is slight lower than males', but females' salaries are remarkable lower than males' salaries. It is because the average salary for female is \$36075 and the maximum salary is \$270000 only.

Use a scatter plot to support this assumption and illustrate the correlation between salary and weekly working



hours:

The scatter plot shows a positive correlation between salary and weekly working hours. The blue line represents the data of males while the red line represents the data of females. Both lines show positive correlations, but the blue line is above the red line. It implies that males' salaries are higher than females' salaries even they work the same amount of time.

Next, I will analyze the models for full model, males and females based on all predictors.

A list of p-values for full model:

```
##                                                            summary.model..coefficients...4.
## (Intercept)                                                                    8.512814e-02
## sexMale                                                                        6.269442e-08
## age20-29                                                                       9.062817e-01
## age30-39                                                                       6.106586e-02
## age40-49                                                                       1.683351e-02
## age50-59                                                                       2.268599e-03
## age60-69                                                                       7.981349e-01
## age70+                                                                         9.878916e-01
## educationLess than high school graduation                                      3.554366e-01
## educationNon-university postsecondary certificate or diploma                   1.418786e-02
## educationUniversity degree or certificate                                      4.141585e-12
## weekly_working_hours                                                           1.225164e-07
```

The p-valus of full model shows that males, having university degree or certificate, and weekly working hours are the most useful predictors that affects the amount of salaries. In other words, people may have higher chances to get high salaries if they have these characteristics. It is because their p-values are the lowest three in the full model. It means they are very significant in statistics.

Summary of males’ model:

```
##
## Call:
## lm(formula = salary ~ age + education + weekly_working_hours,
##     data = subset_M)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -135408  -26608   -4396   17681  642901
##
## Coefficients:
##                                                               Estimate
## (Intercept)                                                   -6222.6
## age20-29                                                       5116.9
## age30-39                                                      22413.4
## age40-49                                                      28236.8
## age50-59                                                      35423.2
## age60-69                                                       8231.7
## age70+                                                        -1848.4
## educationLess than high school graduation                     -7311.0
## educationNon-university postsecondary certificate or diploma  12617.4
## educationUniversity degree or certificate                     35433.3
## weekly_working_hours                                            707.7
##                                                               Std. Error t value
## (Intercept)                                                      13873.6  -0.449
## age20-29                                                         13904.1   0.368
## age30-39                                                         14348.4   1.562
## age40-49                                                         14291.6   1.976
## age50-59                                                         14184.8   2.497
## age60-69                                                         14565.2   0.565
## age70+                                                           19334.1  -0.096
## educationLess than high school graduation                         9878.4  -0.740
## educationNon-university postsecondary certificate or diploma      6849.3   1.842
## educationUniversity degree or certificate                         7442.6   4.761
## weekly_working_hours                                               232.1   3.050
##                                                               Pr(>|t|)
## (Intercept)                                                    0.65397
## age20-29                                                       0.71302
## age30-39                                                       0.11889
## age40-49                                                       0.04872 *
## age50-59                                                       0.01283 *
## age60-69                                                       0.57221
## age70+                                                         0.92387
## educationLess than high school graduation                      0.45958
## educationNon-university postsecondary certificate or diploma   0.06604 .
## educationUniversity degree or certificate                     2.52e-06 ***
## weekly_working_hours                                           0.00241 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 58870 on 508 degrees of freedom
## Multiple R-squared:  0.1471, Adjusted R-squared:  0.1303
## F-statistic: 8.762 on 10 and 508 DF,  p-value: 2.593e-13
```

The model reveals that most of the predictors have high p-values (p > 0.05), which means they are not statistically significant. The useful predictors to explain males' salaries are ages of 40-49, ages of 50-59, education level with university degree or certificate, and weekly working hours because they have low p-values (p <= 0.05) and they are significant in statistics.

Summary of females' model:

```
##
## Call:
## lm(formula = salary ~ age + education + weekly_working_hours,
##     data = subset_F)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -89090 -17891  -1669  12768 210891
##
## Coefficients:
##                                                               Estimate
## (Intercept)                                                    -7365.9
## age20-29                                                       -1640.0
## age30-39                                                       11264.6
## age40-49                                                       13140.7
## age50-59                                                       14850.1
## age60-69                                                         129.6
## age70+                                                         18155.0
## educationLess than high school graduation                      -1309.2
## educationNon-university postsecondary certificate or diploma   6009.1
## educationUniversity degree or certificate                      23352.6
## weekly_working_hours                                             749.5
##                                                               Std. Error t value
## (Intercept)                                                       6574.8  -1.120
## age20-29                                                          7146.5  -0.229
## age30-39                                                          7045.2   1.599
## age40-49                                                          7171.5   1.832
## age50-59                                                          6902.5   2.151
## age60-69                                                          7133.4   0.018
## age70+                                                           13114.4   1.384
## educationLess than high school graduation                         6096.9  -0.215
## educationNon-university postsecondary certificate or diploma     3819.6   1.573
## educationUniversity degree or certificate                         3945.7   5.919
## weekly_working_hours                                               122.2   6.134
##                                                               Pr(>|t|)
## (Intercept)                                                     0.2632
## age20-29                                                        0.8186
## age30-39                                                        0.1105
## age40-49                                                        0.0675 .
## age50-59                                                        0.0320 *
## age60-69                                                        0.9855
## age70+                                                          0.1669
## educationLess than high school graduation                       0.8301
## educationNon-university postsecondary certificate or diploma   0.1163
## educationUniversity degree or certificate                     6.27e-09 ***
## weekly_working_hours                                          1.82e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 31060 on 470 degrees of freedom
## Multiple R-squared:  0.2434, Adjusted R-squared:  0.2273
## F-statistic: 15.12 on 10 and 470 DF,  p-value: < 2.2e-16
```

The model shows that the predictors: ages between 50-59, education level with university degree or certificate, and weekly working hours are useful to explain females' salaries. It is because they have low p-values (p <= 0.05) that means they are statistically significant. The rest predictors are not significant for analyzing females' salaries as their p-values are higher than 0.05.

Then, substitute the coefficients into the Salary_M and Salary_F:

$$Salary_M = -6222.6332 + 5116.9x_{age_{20-29}} + 22413.4x_{age_{30-39}} + 28236.8x_{age_{40-49}} + 35423.2x_{age_{50-59}} + 8231.7x_{age_{60-69}} - 1848.4x_{age_{7}}$$

$$-7311x_{education_{Lessthanhighschoolgraduation}} + 12617.4x_{education_{Non-universitypostsecondarycertificateordiploma}}$$

$$+35433.3x_{education_{Universitydegreeorcertificate}} + 707.7x_{weekly_wokring_hours}$$

$$Salary_F = -7365.9 - 1640x_{age_{20-29}} + 11264.6x_{age_{30-39}} + 13140.7x_{age_{40-49}} + 14850.1x_{age_{50-59}} + 129.6x_{age_{60-69}} + 18155x_{age_{70+}}$$

$$-1309.2x_{education_{Lessthanhighschoolgraduation}} + 6009.1x_{education_{Non-universitypostsecondarycertificateordiploma}}$$

$$+23352.6x_{education_{Universitydegreeorcertificate}} + 749.5x_{weekly_wokring_hours}$$

By comparing Salary_M and Salary_F, most coefficients of Salary_M are larger than coefficients of Salary_F except for the coefficients of $x_{age_{70+}}$, $x_{education_{Lessthanhighschoolgraduation}}$, and $x_{weekly_wokring_hours}$. However, the p-values of $x_{age_{70+}}$ and $x_{education_{Lessthanhighschoolgraduation}}$ are higher than 0.05, which means they are statistically insignificant to explain the salary. For $x_{weekly_wokring_hours}$, the coefficients' difference between males' and females' model is not large enough to affect the salary. Therefore, males earn higher salaries than females based on the coefficients of models.

Discussion

Summary

In this report, I used a multiple linear regression model to predict the salary difference between Canadian males and Canadian females based on ages, education level, and weekly working hours. This report also uses scatter plot, summary, coefficients of models, and p-values to analyze the model and discuss the result.

Conclusion

The result shows that Canadian males' salaries are higher than Canadian females' even males and females have the same characteristics. According to the full model, males who have university degree or certificate and have longer working hours may probably get higher salaries.

It raises the concern of gender inequality because females are disadvantaged while comparing to males. This report shows that the labor market in Canada suppresses females' salaries. It is because females have less accesses to education and economic opportunities. It may also lead to other social problem like poor health in females.

Weakness & Next Steps:

As shown in the scatter plot, the data involves outliers and leverage points that influence the result of the model. The next step could be removing the leverage points, so it would also reduce the residuals and error terms. Besides, the R-squared values of males' model and females' model are 0.1471 and 0.2434 respectively. The values are relatively low and imply some predictors are not useful to predict the salaries. The next step could be increasing the R-squared value because it would improve the accuracy of the model. To adjust R-squared value, it is possible to use AIC and BIC to add or remove predictors that would result in improving the model.

The future steps of analyzing salary difference genders could be adding more potential predictors to the model. For example, salary for part-time jobs vs full-time job, employee benefits and all sources of personal incomes.

References

1. "CHASS Microdata Analysis and Subsetting with SDA, Canadian Income Survey (CIS), 2017." Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/cis.htm
2. "Canadian Income Survey, 2017: Public Use Microdata File Data dictionary" Statistics Canada. University of Toronto Data Library Service, 2001. Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/cis/2017/more_doc/2017CIS_Codebook.pdf