How do Educational Level and Gender Difference Affect Individual's Income

Yutong Wu 1004861955, Yao He: 1004955328, Yujie Wu 1003968904, Xinxuan Lin 1002467877

October 19, 2020

Abstract

Finding a job or pursing a higher educational level after graduation from high school, college or university is an essential concern that young people care about. In this report, we will investigate the relationship between educational condition and income based on the 2017 General Social Data Family conducted by Statistics Canada. Furthermore, how gender difference influences one's income under the same educational background will be included in our discussion as well. Since we filtered the sample group in Ontario area, the conclusion is especially aiming for Ontario people. Our founds displayed a positive relationship between education and income and it also suggest that gender pay gap for women does exist in Canada.

Introduction

While nowadays teenagers tend to feel uncertain about their future, income temporarily seems to be a decent short-period target and foundations to achieve their lifelong goals. Our data is given by General social survey Family 2017 containing personal information of each family unit that was selected from most province(strata) in Canada. By filtering the observation data, we found that Ontario area had the most samples of family that offered the least NA information. So we determined to study the group of respondents in Ontario who's 15 years old and above to explore our topic:"How do Educational Level and Gender Difference Affect Individual's Income".

We first plotted figure 1 box plot to visualize the frequency of each education level and income in different genders. It is clearly that the amount of individuals who graduated from university approximately doubles the the amount of individuals who gets a degree higher than bachelor regardless of gender. While the majority of male's highest educational level was trade certificate, getting a high school diploma is the most common phenomenon among female education. The barplots figure 2-9 and boxplots figure 10 & 11 suggests that women included to earn less them men in general. The logistic regression model we will introduce in detail also implied the same conclusion. However, what led to this gender pay gap? Discrimination, differences in hours worked or differences in years of experience? We have no clue based on the data we are given. So a further investigation of the working condition of males and females is worthy attempting. Additionally, figure 10 and figure 11 suggested some outliers representing that a low educational level could also generate an decent income for both male and female. Thus, a legitimate inference would be our respondents inherited a fortune from their parents which didn't require themselves a higher educational level. But this will need a series of survey aimed at the degree of respondents' parents to explain the possibility and reason behind the outliers.

Data

About Gss The data (General social survey on Family (cycle 31), 2017 (i.e. gss.csv)) was provided by Statistic Canada. It contains personal information of each family unit that was selected from each strata in order to update people's understanding of families in Canada.

Methodology ***Target Population*** The target population for "gss.csv" were all people who are older than 15 years old in Canada, but not including people who lived in Yukon, Northwest Territories, and Nunavut; and full-time residents of institutions.(statistic Canada, 2020, p.11)

Frame The survey frame in this dataset is the combination of the list of telephone number in use and the address registered within provinces. The telephone number were linked to associated home address. For the address that linked to multiple telephone numbers, they were reordered by type of the telephone numbers (i.e. landline telephone would be considered as best telephone to reach, and cellphone would be the last). Those telephone numbers that were not linked to associated address would also be included on the survey frame.

Sampling Approach This dataset(gss.csv) was sampling by using stratified sampling method; samples were divided into strata based on the geographic area, in total 27 strata. In each strata, samples were collected by using method of simple random sample without replacement.

Sample Sample in this dataset were households that have at least one person older than 15 years old in the selected reached households. The target sample size of "gss.csv" was 20,000, but the actual number of observations is 20,602.

Method of Reaching Sample Data were collected by interviewing these selected family units via Computer assisted telephone interviews. All interviews took place from 9:00 a.m. to 9:30 p.m. Mondays to Fridays. If the selected household was inconvenient at the timing of the first call-in interview, they could also reschedule the interview to Saturdays from 10:00 a.m. to 5:00 p.m. and Sundays from 1:00 a.m. to 9:00 p.m.. To encourage the participation of the household who refused the interview, they would be reached up to two more times to explain the importance of the survey. Households who missed the call would be contacted numerous times.

Date Collected Method **education:** The responses of "education" were collected from respondent by using binary questions; for example, press one if the respondent (is attending/attended) trade school or college. Also, collected by asking the types of the highest diploma or certificates they have completed. **income_respondent:** The income of respondent in "ggs.csv" was derived from respondent income tax files in calendar year 2016. It represents the total income of respondent received in calendar year 2016 before income taxes and deductions. The income of respondent who were in 15-year-old would be considered as 0, because his/her tax files could not be linked. **sex:** The gender of the respondents were derived from the household roaster. The sample contains 9,399 males, and 11,203 females.

Strength Interviewers made multiple call backs to people who refused to participate the survey, and explain the reasons and importance of the survey which could provide reassurance to them, and encourage them to do the survey. Numerous call backs were also made to those people who were not at home, it could provide more chances to those people to participate the survey as the way of increasing the response rate. In order to keep the accuracy of the data, instead of only asking once, the survey contains multiple questions related to the same topic as the way to prevent non-thinking answers (i.e respondent randomly select an option without thinking). For example, instead of only asking what intuition that response attended/is attending, the survey also asked them what the highest certificate, diploma or degree that they have completed is. By comparing the answers of these two questions, we could find out the accuracy of responses. The income of respondents were derived from their tax file instead of asking them during the interview, which could increase the accuracy and preventing the false or non-response answers from respondents.

Weakness The survey is a long phone interview which would cause the non-meaningful response. For example, when asking the education level of the respondent, there were 4 unit selected the option "Don't know" which is unrealistic that the person does not know his/her education level.

The length of the the survey could also lower the response rate. The response rate of gss.csv was 52.4%.

Moreover, when asking the education level, the question used "What type of educational institution (are you attending/did you attended)?" followed by some types of institution. But the options for the interviewees are "Yes/No". It is unclear for the respondent to use a binary option to answer an open type questions. It would be better to make the question like "Are you attending/did you attended …" followed by some types of institution.

Our Data We were interested in the relationship between different individuals' education level and income. Therefore, our data was collected by drawing out the columns of "Education", "Income_Respondent", and column of "Sex" from the dataset "gss.csv". In order to better determine the relationship in education level and income, we eliminated the non-response answer from "education" column.

However, due to the difference between consumer price in different cities, we focus on information that provided by respondents who lived in Ontario instead of all individuals. So we used responses from Ontario respondents to form a new dataset called *"processed_data"*. Hence, it would not be able to represent all province in Canada.

Since "income_respondent" was recorded as an interval, and it is impossible to find the actual amount of income that each respondents have. Therefore, we treated the mid-point of each interval as the actual value of income for each respondent in that category. As drawbacks, it caused the bias of income, because it is unrealistic that respondents who are in the same category have same amount of income. **Population:** All people in Ontario who are in/older than 15 years old. **Frame:** The list telephone number in use, and the list of address registered in Ontario. **Sample:** Households that have at least one person older than 15 years old in the selected reached Ontario households.

Variable ·**Education**: Excluding the non-response answers, we have seven different categories for education level, which were: 1. "Less than high school diploma or its equivalent"; 2. "High school diploma or a high school equivalency certificate"; 3. "Trade certificate or diploma"; 4. "College, CEGEP or other non-university certificate"; 5. "University certificate, diploma or degree above the bachelor's level"; 6. "University certificate, diploma or degree below the bachelor's level"; 7. "Bachelor's degree (e.g. B.A., B.Sc., LL.B.) ·**Income_Respondent**: The income of respondent in this datatset(gss.csv) were the total income of respondent received in calendar year 2016 before income tax and deductions. ·**Sex**: sex in"gss.csv" were considered as two different genders: male and female. The sample in our data ("processed_data") contains 2,489 males and 3042 females.

Below is the bar graphs showing the frequency between each education level and income in different genders:

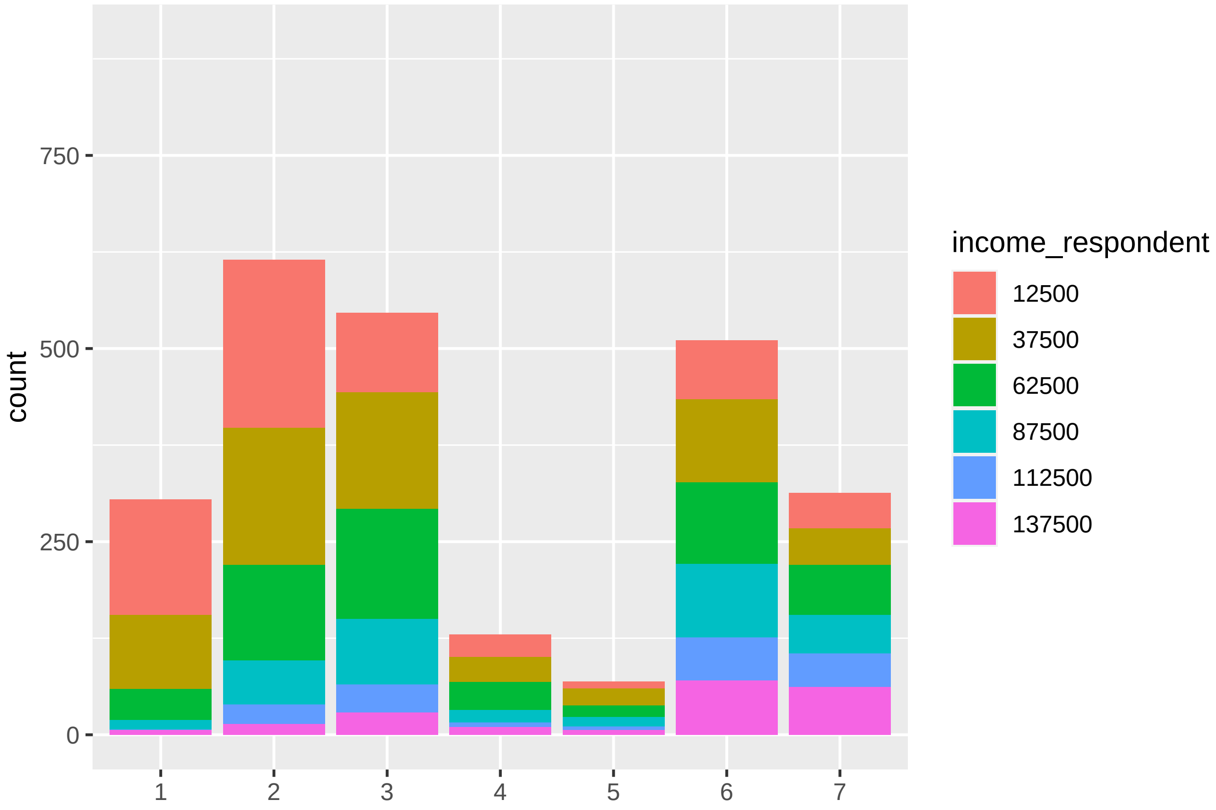

figure 1. male

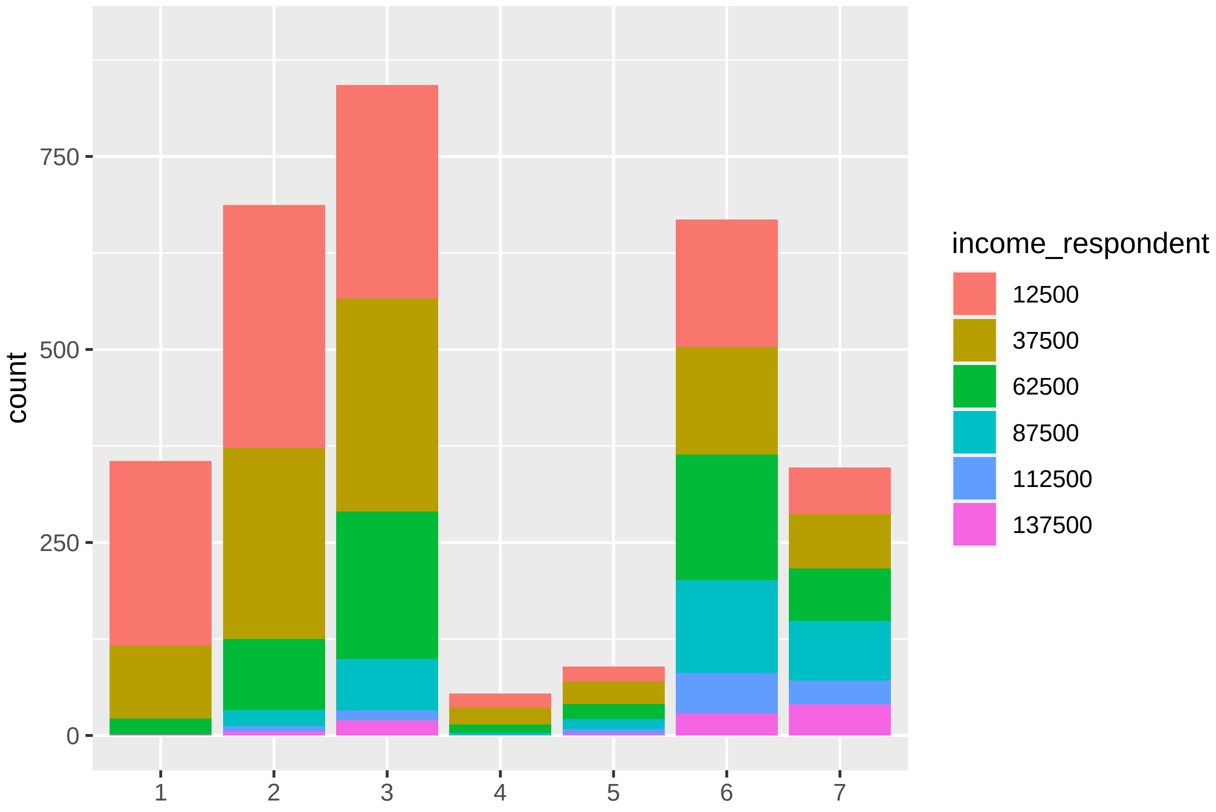

figure 1. female

1: “Less than high school diploma or its equivalent” 2: “High school diploma or a high school equivalency certificate" 3: "College, CEGEP or other non-university certificate or di…" 4: "Trade certificate or diploma" 5: "University certificate or diploma below the bachelor's level" 6: "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" 7: "University certificate, diploma or degree above the bach…"

Model

As mentioned above, in order to determine the relationship between education level and income, we used "education", "income_respondent"; and because it included male and female respondents, we also include the column of gender. Since income was collected as interval, we used mid-point of interval as the actual amount of income which turned income as numeric data. In other words, we have numerical response variable and binary predict variables.

Based on figure 3 to figure 9 (figures in Results and Discussion), we could see that all education categories have more females than males except in Trade School category, however in each category the number of male respondent who has highest income were more than the number of female respondent who has highest income. Therefore, gender could affect the relationship between education level to income.

Therefore, it is appropriate to use multiple linear regression for our data.

Our hypothesis is the higher the education level is, the higher the income would be, and gender would affect the relationship between education level and income.

The model is:

$$Income_{respondent} = \beta_0 + \beta_1 Xc + \beta_2 X_h + \beta_3 X_{lh} + \beta_4 X_t + \beta_5 X_{ub} +$$
$$\beta_6 X_{ua} + \beta_7 X_m + \beta_8 X_c X_m + \beta_9 X_h X_m + \beta_{10} X_{lh} X_m + \beta_{11} X_t X_m$$
$$+\beta_{12} X_{ub} X_m + \beta_{13} X_{ua} X_m + \epsilon_i$$

$Income_{respondent}$ represents the total income (before income tax and deductions) that the respondent received in calendar year 2016.

X_{lh}: vector of respondent whose education level is less than high school diploma or its equivalent. X_h: vector of respondent whose education level is high school diploma or a high school equivalency certificate. X_t: vector of respondent whose education level is Trade certificate or diploma. X_c: vector of respondent whose education level is college, CEGEP or other non-university certificate. X_{ua}: vector of respondent whose education level is university certificate, diploma or degree below bachelor's level. X_{ub}: vector of respondent whose education level is university certificate, diploma or degree below bachelor's level. X_m: respondent who was indicated as male in column of sex.

When all X values were zero, this model shows the average income for female respondent who has bachelor's degree, which is same as what $beta_0$ represents to. β_0 represents the average income of female respondent who has Bachelor's degree. β_1 to β_6 represents the relationship between income and corresponding education level. If β_i is negative, it means that income is inversely proportional to the corresponding education level. Positive β_i represents that income is proportional to corresponding education level. β_7 shows the relationship between gender and income. Negative β_7 means that male has negative effect on income. Conversely, positive β_7 indicates male has positive effect on income. β_8 to β_{13} represents the interaction effect between education level and gender. Negative β_j, where j is between 8 to 13, represents that male and corresponding education level have negative effect on income. Conversely, a positive β_j indicates that male and corresponding education level have positive effect on income.

ϵ_i indicated as the error term in the model.

term	estimate	std.error	statistic	p.value
(Intercept)	22429.5775	1718.080	13.0550253	0.0000000
education2	9975.8083	2115.918	4.7146475	0.0000025
education3	19939.7812	2048.493	9.7338799	0.0000000

term	estimate	std.error	statistic	p.value
education4	15070.4225	4728.334	3.1872584	0.0014443
education5	29958.0630	3837.425	7.8068147	0.0000000
education6	34119.8237	2126.145	16.0477378	0.0000000
education7	44969.5580	2443.696	18.4022712	0.0000000
sexMale	10234.3570	2527.350	4.0494416	0.0000520
education2:sexMale	998.4687	3101.079	0.3219746	0.7474842
education3:sexMale	4722.2917	3090.507	1.5279991	0.1265701
education4:sexMale	8419.4892	5818.378	1.4470509	0.1479395
education5:sexMale	-121.9974	5774.800	-0.0211258	0.9831460
education6:sexMale	3348.3358	3163.362	1.0584739	0.2898858
education7:sexMale	-516.8790	3571.448	-0.1447253	0.8849330

We used R to simulate the model of the relationship between education level and income; we got the above summary table. Based on this summary table, we could get the estimated model as followed:

$$\widehat{Income}_{respondent} = 56549.401 - 14180.043 \cdot X_c - 24144.043 \cdot X_h - 34119.824 \cdot X_{lh} - 19049.401 \cdot X_t - 4161.761 \cdot X_{ub} + 10849.734 \cdot X_{ua} + 13582.693 \cdot X_m + 13582.693 \cdot X_c \cdot X_m - 2349.867 \cdot X_h \cdot X_m - 3348.336 \cdot X_{lh} \cdot X_m + 5071.153 \cdot X_t \cdot X_m - 3470.333 \cdot X_{ub} \cdot X_m - 3865.215 \cdot X_{ua} \cdot X_m$$

Interpretation: X is the dummy variable which means the value of X would be 1 if respondent fell into the corresponding categories, otherwise the value of X would be 0. For example, if the respondent is male and has high school diploma or a high school equivalency certificate, then $X_h = 1$, $X_m = 1$, and other variables of X would be 0 Otherwise the value of X_h, and X_m would be 0, and the associated X variable would be 1. By substituting these values, we could get the estimated number for average income of this respondent. From this model, we could see that the coefficient of X_lh (i.e. respondent whose education level was less than high school) is -34119.824 which is also the smallest number among all the coefficient. It shows that education level which less than high school has the largest negative effect on income of respondent. The coefficient associated with X_m is 13582.693, which shows that male has positive effect on income. More details would be explain in Results and Discussion section.

Caveats As mentioned in Data section, income was collected as the mid-point of the interval due to the difficulty in determining the actual number, which would result people who were in the same category all have the same amount of income. It would affect the accuracy of income estimation; and also could cause the low R^2 value, and hence affect the accuracy of the model.

Results and Discussion

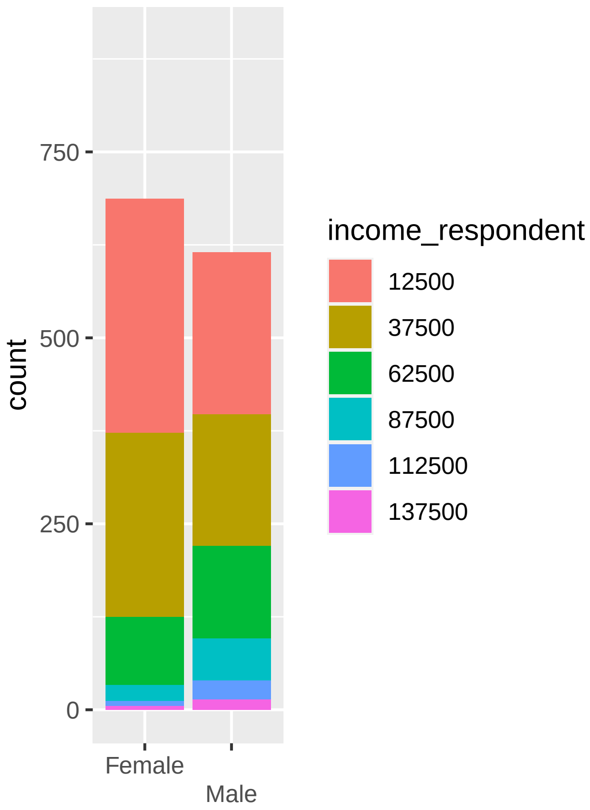

figure 3.

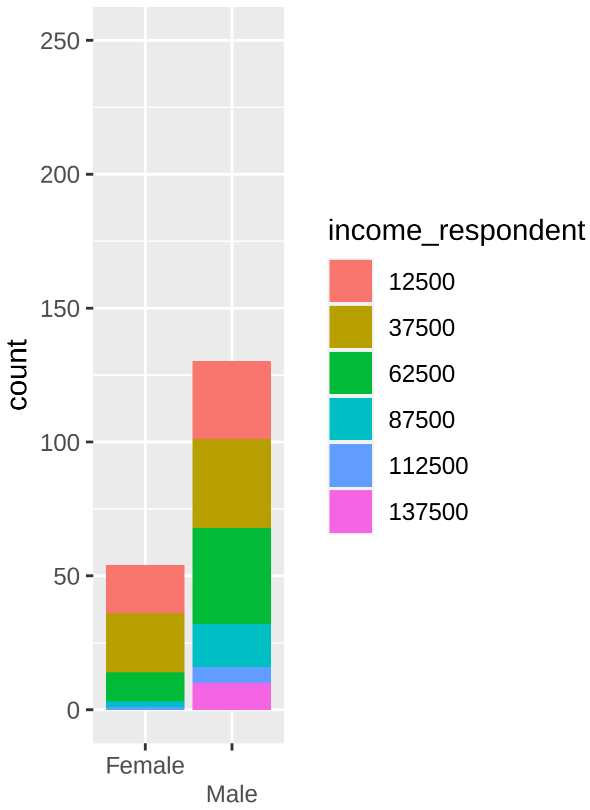

figure 4.

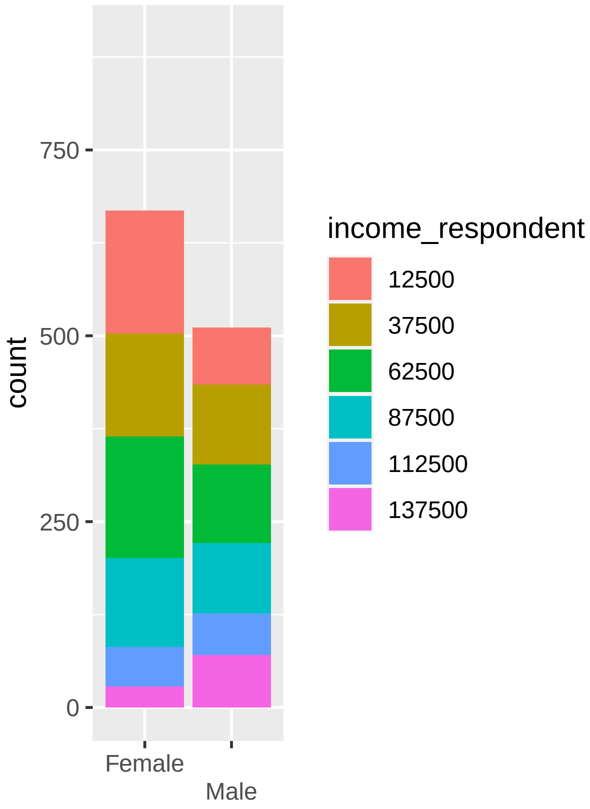

figure 5.

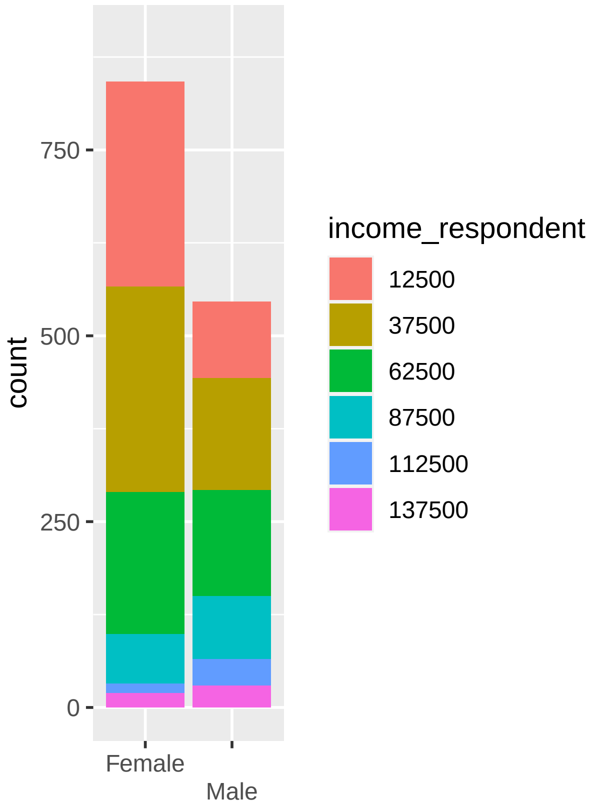

figure 6.

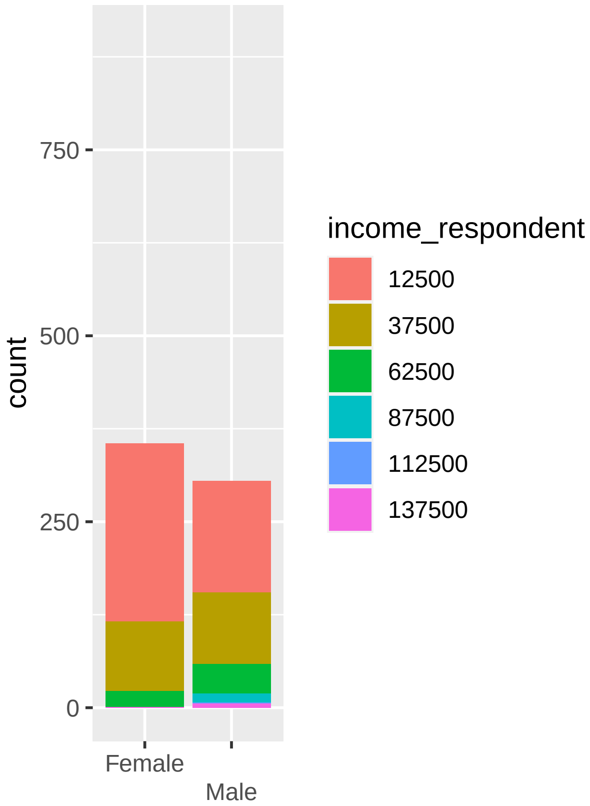

figure 7.

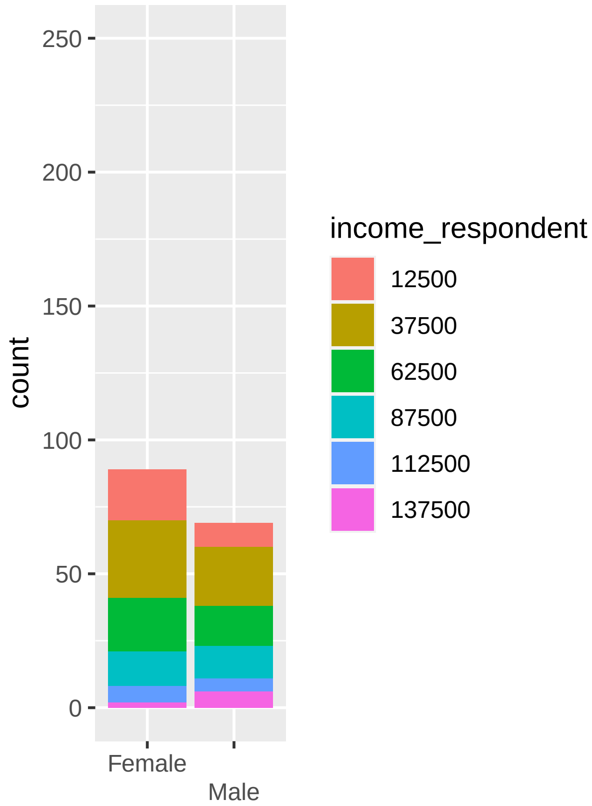

figure 8.

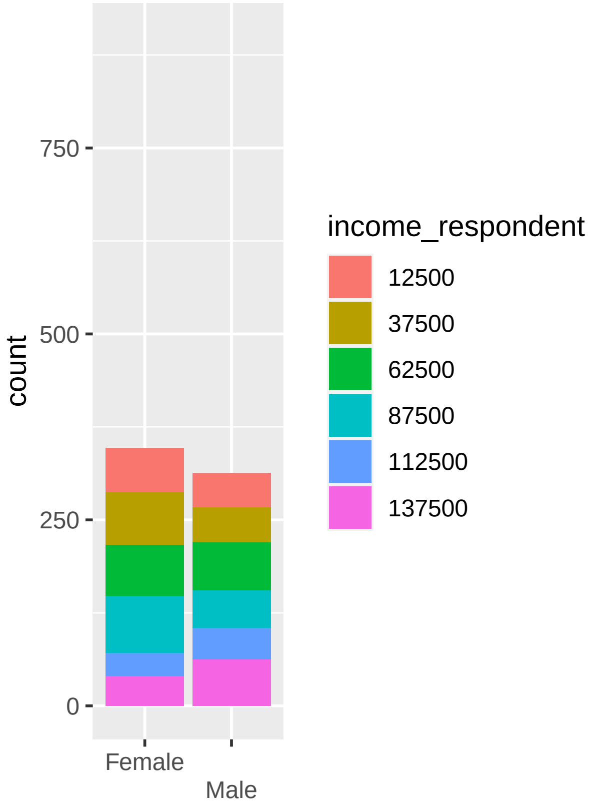

figure 9.

figure 3. High school diploma or a high school equivalency certificate figure 4. Trade certificate or diploma figure 5. Bachelor's degree (e.g. B.A., B.Sc., LL.B.) figure 6. College, CEGEP or other non-university certificate or di… figure 7. Less than high school diploma or its equivalent figure 8. University certificate or diploma below the bachelor's level figure 9. University certificate, diploma or degree above the bach…

The bar graphs illustrate the count of the respondents with different incomes, grouped by their education and degrees. The majority of the groups (shown in figure 3 and 5 to 9) are the groups where female respondents are more than male respondents while figure 2 is the only group where male respondents are more. If we put the respondents with income from 12500 to 62500 into the low income group and the respondents with income from 87500 to 137500 into the high income group, it is clearly shown in figure 1 to 7 that there are (slightly) more counts of males in the high income group than females with the same degree based on our data. Again, this is another illustration of the gender discrimination in careers.

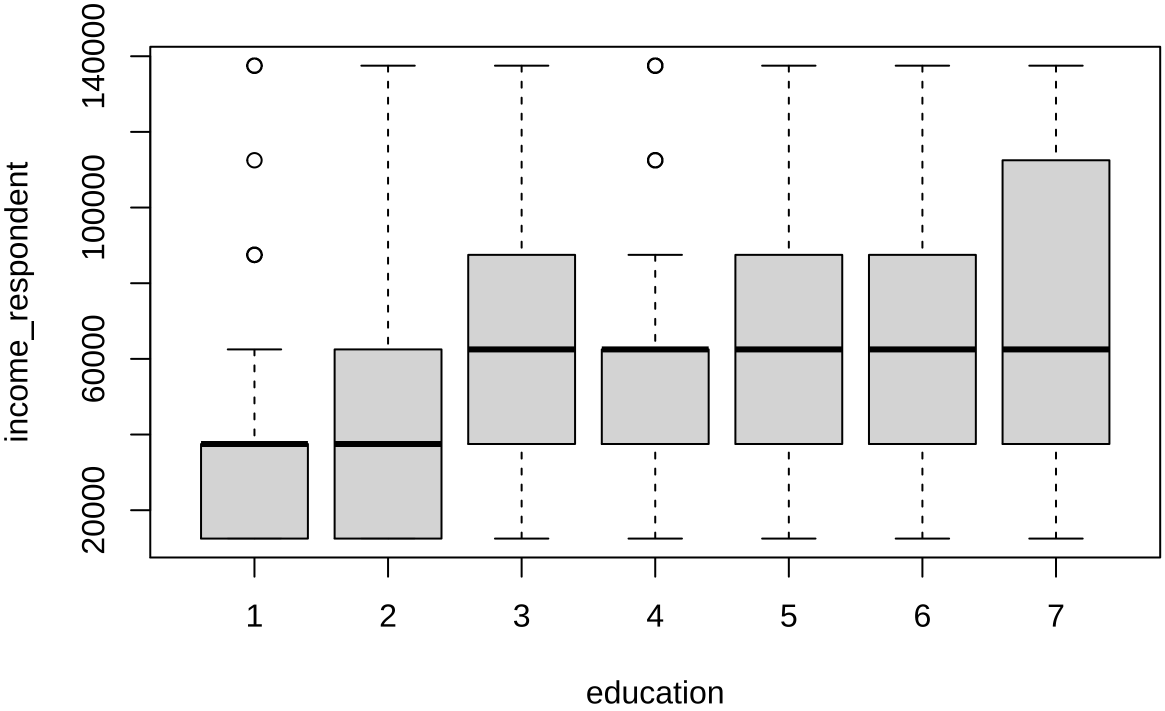

Figure 10:Male

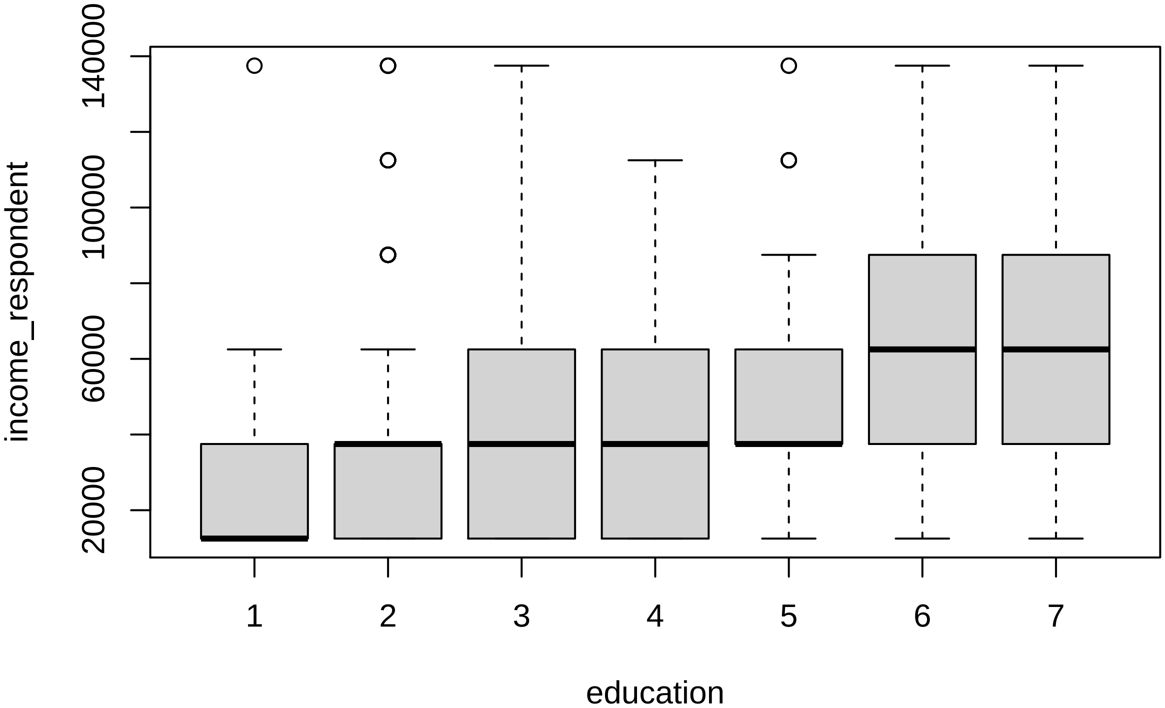

Figure 11:Female

1: "Less than high school diploma or its equivalent" 2: "High school diploma or a high school equivalency certificate" 3: "College, CEGEP or other non-university certificate or di…" 4: "Trade certificate or diploma" 5: "University certificate or diploma below the bachelor's level" 6: "Bachelor's degree (e.g. B.A., B.Sc., LL.B.)" 7: "University certificate, diploma or degree above the bach…"

The box plots illustrates the income for males and females by their degrees, with the first column being the lowest degree and the seventh column being the highest degree. We may see from figure 10, that as the degree of the male respondent goes up, the average income goes up as well. For respondents with college certificates and above, the average income of these groups are around 60,000. Another thing worth mentioning here is that each group tends to be distributed wider. In the group 1: Less than high school diploma or its equivalent and the group 4: Trade certificate or diploma, there are 3 and 2 outliers respectively. These respondents earn a incredibly high income than the average of other respondents in their group. This might be due to personal ability of these respondents. From figure 11, we may also see that as the degree of the female respondents increase, the average income goes up as well. However, compared to male respondents, female respondents with degrees of bachelor and above may have an average income of 60,000. For the groups of college certificate, trade certificate and university certificate, female respondents on average earn 20,000 less than male respondents. In the group 1: Less than high school diploma or its equivalent, the group 2: High school diploma or a high school equivalency certificate, and the group 5: University certificate or diploma below the bachelor's level, there are 1, 3, and 2 outliers respectively. These respondents earn a incredibly high income than the average of other respondents in their group. This might also be due to personal ability of these respondents. In conclusion, from the boxplots, we may see the increase trend of degree's effect on income. Another thing worth mentioning here is the gender discrimination shown through the income data. We may discuss more in the linear regression model.

For simplicity, let's interpret β_0 to β_6 under the scenario when $X_m = 0$. That is, assuming the respondent is a female.

As we may see from the summarizing table, when all X values were 0, the value of β_0 is 56549.This represents the average income of female respondent who has Bachelor's degree or above is 56549. The p-value for $\beta_0 < 0.05$, hence we have a strong evidence that the average income of female respondent who has Bachelor's degree is not 0 and supports our result.

β_1 to β_5 are negative, meaning that as the degree of the female respondent decreases, the average income tend to decrease. The p-value for female respondents who has a University certificate or diploma below the bachelor's level is greater than 0.05 while p-values for the other four groups of respondents are less than 0.05. We have a strong evidence that the average income of female respondent who has degrees that is below Bachelor's degree (except respondents who has a University certificate or diploma below the bachelor's level) is not 0 and supports our result. For respondents who has a University certificate or diploma below the bachelor's level, we fail to reject the these respondents tend to on average have a lower income than the respondents with a Bachelor or higher degree.

β_6 is positive, meaning that the average income for female respondents with a University certificate, diploma or degree above the bachelors level tend to be higher than the average income for female respondents with education above bachelors level. The p-value is less than 0.05. We have a strong evidence against the null hypthesis that the difference between the average income of female respondent who has degrees that is University certificate, diploma or degree above the bachelors level and the average income for female respondents with education above bachelors level is 0 and supports our result.

Another thing worth mentioning here is the extent by which the income increases and decreases as degree changes. As we may read from the summary table, having a degree that is above the bachelor's level increases the income most while the further away the degree is from the bachelor's level, the more decrease in the income.

In general, we may find that as a respondent's degree go above the bachelor's degree, the average income of the respondent tend to increase. This may be caused by the case that as the respondents pursue higher degrees, they learn more useful skills and knowledge, experienced more in the industries, and therefore may bring more value to the companies. Assuming income reflects the value created by the respondents, it is reasonable that as a respondent's degree go above the bachelor's degree, the average income of the respondent tend to increase. One exemption is the group of a University certificate or diploma below the bachelor's level. One possible explanation for that is these respondents might be awarded certificates for their hard skills and apprenticeships. With that being said, their skills may not require lots of theoretical knowledge and experiences matter more.

β_7 is 13583, which means that on avergae, a male with a bachelor or higher degree has a higher income than a female with a bachelor or higher degree. The p-value < 0.05. This indicates that there is strong evidence against the null hypothesis that there is no difference between males and females of the same degree, meaning on average, males earn more than females with same degrees. One explanation is that gender discrimination, shown through income differences, occurs in the careers.

β_8 to β_{13} represents the interaction effect between education level and gender. That is, for male respondents, that's an additonal effect of degree on their income other than the mere effect of β_1 to β_6 have on females and the effect of being a male, which is shown through β_7. In this case, college certificate and trade certificate have an additional positive effects on males' income, while the other degrees have an extra negative effects on males' income. For example, if a respondent is a male with a trade certificate, then his predicted income will be 56549 + (-19049) + 13583 + 5071 = 56154. While for a female respondent with a trade certificate, her predicted income will be 56549 + (-19049) = 37500. However, the p-value for all the interaction terms are greater than 0.05, which means there is no evidence against the hypothesis that there is no additional effect on a male's income with respect to his degree. With that being said, effect of degrees are the same for both males and females.

The overall R^2 for the model is 0.18, which means that there is a weak positive correlation between the variables.

Our study is to investigate the effect that degrees and genders may have on people's income. From our analysis, we may see that degrees do have a positive effect on people's income. That is, the higher a person's degree is, the higher his income would be. In addition, we also find that males tend to have a higher income on average than females. This is an illustration of gender discrimination that occurs in the real life. In the past, women may not have so many opportunities to study and to work. And there might be the sterotype that men work more and are smarter in some aspects. However, with the development of the society, the awareness of gender equality does arise and females are recognized, although on average females are still at the disadvantages. Our study may bring people's attention towards education and the gender equality. Hopefully in the future, citizens will have higher education level and the gender discrimination may disappear.

Weaknesses and Next Steps

Although we are able to analyze and draw conclusions from the data, there are still drawbacks in our dataset that reduces the accuracy of our model. Firstly, the most important weakness we encountered is that the income data was collected as a interval rather than a number. In order to model our data, everyone in the same income division will share the same amount of income, which is unrealistic. For instance, some teenager respondents aged 15 to 18 may have not started earning money, yet their income was automatically recorded as $12500 annually since they answered "less than 25000" when doing the original questionnaire. To improve the accuracy, a smaller interval or even better, a precise amount of income number is expected to be collected for our education and income relationship assessment.

Another weakness in this study was that family background of the respondents might be an interesting factor that caused outliers in the figure 10 and 11. For example, a person with a wealthy farm parents do not need to purse a higher degree to manage his(her) parent's business, though he(she) could still make lots of money by taking over the elder's farm. Although it may look like an outlier in the overall data set, it is a statistically possible exception that could happpen in the real life. Therefore, for a futural steps, we can construct an another survey asking about the occupation of the interviewers' family and develop another variable to explore the potential reason hidden behind the outliers.

After seeing the gender pay gap for women, it is hard to conclude that gender wage gap is simply gender discrimination. So it would be interesting to do a further study on "Why in general women earns less than men?" or "A Comparison of The Obstacle that men and women are facing in the same occupation".

Appendix

Github repo: https://github.com/wuyujie1/STA304PS2

References

1. Statistics Canada. (2017). General social survey on Family (cycle 31), 2017 - Canadian general social surveys (GSS). Retrieved October 19, 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm

2. Statistics Canada. (2020, April). General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User's Guide (Rep.). Retrieved 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

3. N/A. (2017). 2017 General Social Survey: Families Cycle 31 Public Use Microdata File PUMF (Rep.). Retrieved October 19, 2020, from Statistics Canada website: https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_Codebook.pdf

4. Alexander, R., & Caetano, S. (2020, October 7). Gss_cleaning.R. Retrieved October 19, 2020.

5. Robin Bleiweis. Quick Facts About the Gender Wage Gap. Retrieved 2020, from Center for American Progress: https://www.americanprogress.org/issues/women/reports/2020/03/24/482141/quick-facts-gender-wage-gap/