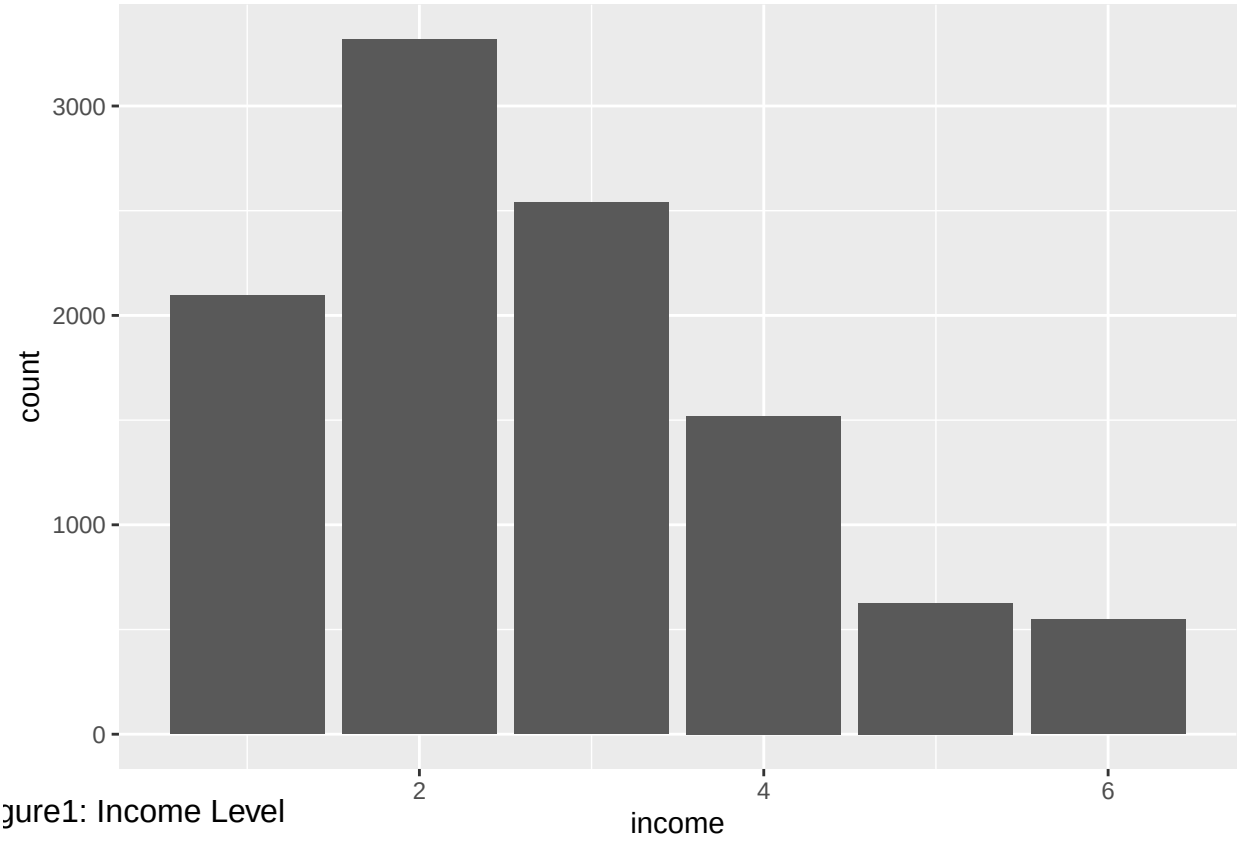
Influence of Education, Occupation, Hours of work, and Class of work on Income Level

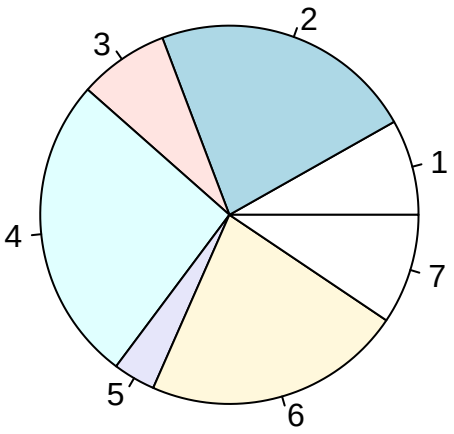
Tianbao Kuang (1004497724), Siyi Ma (1004554190), Xiaoke Zeng (1004005266)

Oct 18, 2020

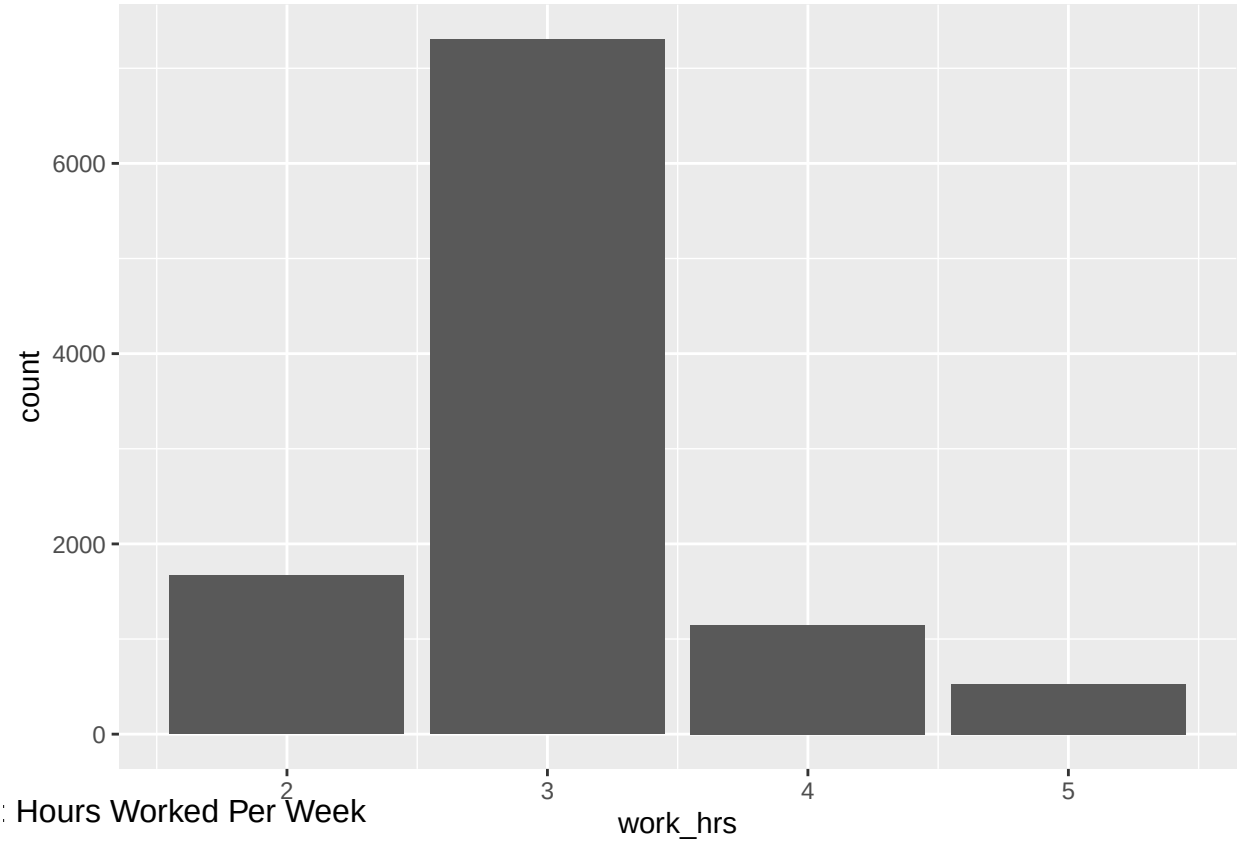
```
## -- Attaching packages ------------------------------------------------------------------------ tidyvers

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.0
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts --------------------------------------------------------------------------- tidyverse_con
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

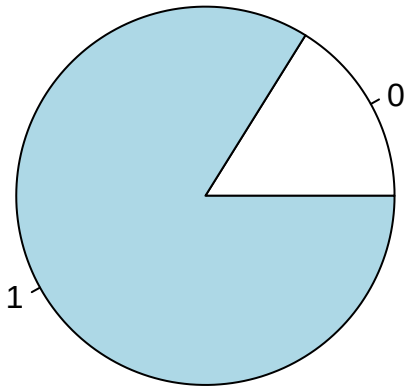


gure1: Income Level

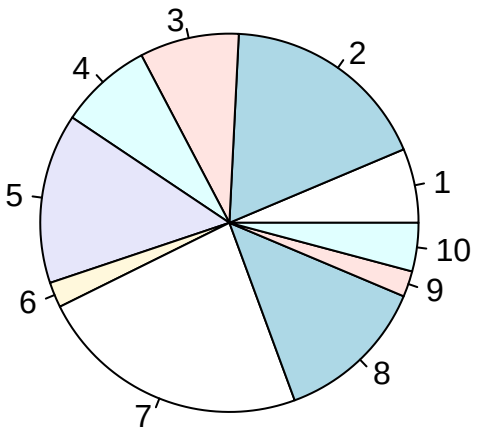


Education Level Distribution


Hours Worked Per Week
count
6000
4000
2000
0
2
3
4
5
work_hrs



ure4: Class of Work



gure5: Occupations

```
##
## Call:
## lm(formula = df6$income ~ df6$perm + df6$highsch + df6$trade +
##     df6$college + df6$uni + df6$bach + df6$above + df6$f30_40 +
##     df6$f40_50 + df6$over_50 + df6$manage + df6$bus + df6$sci +
##     df6$health + df6$edu + df6$art + df6$sales + df6$trans +
##     df6$res)
##
## Coefficients:
## (Intercept)     df6$perm  df6$highsch    df6$trade  df6$college      df6$uni
##      14.009       13.783        5.214       12.749        9.308       17.798
##    df6$bach    df6$above   df6$f30_40   df6$f40_50  df6$over_50   df6$manage
##      20.627       30.947       14.441       21.769       28.337       27.592
##     df6$bus      df6$sci   df6$health      df6$edu      df6$art    df6$sales
##       1.541       15.768        4.904        5.874       -7.269       -9.560
##   df6$trans      df6$res
##       7.641        2.164
```

Abstract

Maximizing annual income has always been a goal for everyone, but not many people are incentivized to discover the factors that may influence income. This report uses Multiple Linear Regression Model to obtain the estimated effects of education, occupation, hours of work and class of work on income level. We find significant differences in the income level between each factor.

Introduction

The goal is to analyze how the four factors influence people's annual income, including respondents' education level, occupations, whether the job is permanent or not, and the average number of hours worked per week. Even though we are aware of the general direction of the association between certain factors and income, it still needs further analysis of the association's strength and how different the income will be for various categories within each factor. We will reveal the answer in the result section.

Data

The dataset we choose for the project is used to analyze the relationship between personal income and its five factors. Annual personal income is the response variable and it contains six different categories. The predictors we choose to build the multiple linear regression model includes education level, hours worked per week, permanent/not permanent job, and occupation. We made a hypothesis that education, hours worked and permanent job position are linearly correlated with someone's income level. We aim to figure out how much influence these factors each has on our response variable income level.

Generally, we dropped all the data and its corresponding respondents that are categorized into"N/A", including but not limited to "Valid skip", "Don't know", and "Refusal". This is because we are not sure which category they belong to and its presence as a separate category will affect our entire model estimation.

From the Figure 2 of education level, we can observe that the majority of the respondents holds a high school diploma(23.9%) or college diploma(22.5%) or a bachelor's degree(18.5%).

The Figure 1 of personal income is right skewed which means its mean value will be above the median value and the mode. The categories of the income variable are income intervals such as "$25,000 to $49,999". We used the median values of the income intervals. It is reasonable as the estimation is in alignment with our research of income distribution in Canada [1].

The Figure 3 of working hours shows that most of the respondents work for 30 hours to 40 hours per week. The category "0 hour of work per week" is dropped out from the dataset. It makes little sense when people do not work at all while being employed and getting paid. Also, it only accounts for 0.2% of the dataset. Therefore, it is safe to assume this is an outlier and may be a result of collection error.

The Figure 4 shows the distribution of permanent vs non-permanent workers. We can observe that most of the people surveyed are working in a permanent job position. It is expected that people who work in permanent positions earn more income than people who are non-permanent workers on average.

From the Figure 5 of occupation, we can find out that most of the samples in the survey work in the service and business industries, some of other people work in education and trade industries and only few people work in other industries.

One of the potential drawbacks which may affect our model estimation is that some factors which have great impact on income level may not be included in the GSS dataset, and therefore they are not included in our model. Another drawback is that some of the data points are dropped out from the dataset due to the undefined values of the variable. It may help improve the accuracy of our estimation if we could correctly categorize them into existing datasets.

Model

For this project, we used a multiple linear regression model to simulate the linear relationship between response variable y and several explanatory variables x.

The full model is y= B0+B1X1+B2X2+B3X3+...BkXk+e

The {X1,X2,X3...Xk} are the explanatory variables which predict the result of response variable y. B0 is the intercept and {B1,B2,B3...Bk} are the slope coefficients which measures the change in response variable

for one unit change in explanatory variable. In our model, we have four different factors which may affect the estimation of income level and each factor is broken down into several categories to indicate its separate impact on income level as following: income = B0 + B1 permanent + B2 highschool + B3 trade + B4 college + B5 university + B6 bachelor + B7 above + B8 workhour30_40 + B9 workhour40_50 + B10 hourover_50 + B11 manage + B12 bus + B13 science + B14 health + B15 education - B16 art + B17 transport + B18 res

For example, variable education has seven different categories. Among these categories, "less than high school diploma" is considered as a reference. Other categories like "high school diploma" and "bachelor's degree" are made into binary predictors. Under each column, the value is "1" when the respondent has the corresponding education level, and "0" otherwise. The coefficients of those six binary predictors are elaborated as the average difference of income earned between those who have certain education levels and those who have less than high school diploma.

Results

We get a multiple linear regression model as the following: income = 14.009 + 13.783perm + 5.214highsch + 12.749trade + 9.308college + 17.798uni + 20.627bach + 30.947above + 14.441f30_40 + 21.769f40_50 + 28.337over_50 + 27.592manage + 1.541bus + 15.768sci + 4.904 health + 5.874edu - 7.269art + 9.560trans + 2.164 res

As a reference point, The average income among the 10645 respondents is $14,009 for someone who possesses a less than high school diploma or its equivalent, has a non-permanent manufacturing job, and works for less than 30 hours per week.

In terms of the impact of education, income earned generally increases as the education level increases. However, those who hold a trade certificate or diploma earn $3,441 ($12,749 - $9,308) more than those who hold a college degree. It makes sense since the trade certificate is tailored into specific skill sets that are in high demand in the job market. On average, whose who have a high school diploma or a high school equivalency certificate, trade certificate or diploma, college, diploma below the bachelor's level, bachelor's degree, diploma above the bachelor's level respectively earn $5,214, $12.749, $9,308, $17,798, $20,627, $30,947 more than those whose education level is lower than high school diploma or its equivalent.

Another important factor is occupation. The average income for different occupations is ranked from highest to lowest as the following: management, natural and applied sciences, trades and transport related, education and law, heath, natural resources and agriculture, business and finance, manufacturing and utilities, arts. Holding other factors constant, those who have a permanent job earn $13,783 more than those whose job is seasonal, temporary, term, or casual on average. As for the hours worked per week, it strictly follows a positive correlation. The increase of income is significant from working less than 30 hours per week to working 30 to 40 hours per week at $ 14,441. This can be explained by the different salary standard between part-time and full-time employees. As the working hours increase, however, the magnitude of increase in annual income decreases. The differences between those who work over 50 hours per and those who work 40 to 50 hours per week is $760 (($28,337 - $ 21,769)-($21,769 - $14,441)) less than the difference between those who work 40 to 50 hours per week and those who work 30 to 40 hours per week.

Discussion

In general, we confirmed our preliminary prediction of the associations' direction. At the same time, we get to quantify the differences in terms of income earned. However, some evidence is against our preliminary prediction, including the unexpectedly high income for trade certificate holders and the diminishing marginal return of hours worked. Different occupations' income rank is also informative for those deciding their career paths factoring the income level.

Weaknesses

Since we used the Multiple Linear Regression Model throughout our analysis, we need to be aware of the assumptions of MLR. Any violation of the assumptions could cause bias and inconsistency in our estimates. Therefore, we carefully filtered our data to avoid any violations of Zero Conditional Mean, No multicollinearity and IID data. Given our Model is linear, and we have randomly collected data, the IID data assumption will be satisfied. Then, we looked at the Zero Conditional Mean assumption. This assumption assumes the error term in the regression model is uncorrelated with the exogenous variables. We did not find any apparent terms that might be correlated with the exogenous variable during the analysis. However, we didn't do any exogeneity test, so it is still possible for the independent variable to be correlated with the error term. Even though we assumed multicollinearity does not exist in our Model, we did not come up with any evidence for such an assumption. Moreover, we noticed that the dependent variable in regression models should be continuous. Due to the nature of survey data, the dependent variable y (in our case is personal income) is discrete; we have to turn it into a continuous variable for estimator validity, making our estimates less accurate.

Next Steps

If we could continue to work on our report, we will use an exogeneity test and multicollinearity test to verify our assumptions. We will use the T-test to check for statistically insignificant variables and drop them to make our analysis more reliable. However, we were aware about the discrete dependent variable problem and wanted to make our analysis more accurate by changing the main model to the ordinal logistic regression. According to our research, this regression model will regress the dependent variable on the exogenous variable and thus solves the main issue in our report. Unfortunately, none of us had the knowledge about this regression model.

References

[1]"Income distribution in Canada in 2018, by income level", statista, 2020, [Online]. Available:https://www.statista.com/statistics/484838/income-distribution-in-canada-by-income-level

[2]R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[3] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[4] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.