Analyzing Canadian Income Against Age, Education and Occupation

Ruize Liu (1003320499), Yi Lei Feng (1003139356)

October 19th, 2020

Code and data supporting this analysis is available at https://github.com/Raymastered/STA304-PS2.

Abstract

In this paper, we assessed Canadian respondent incomes in relation to age, education level, and occupation to gain an understanding of how these factors contribute to modern economic success. We used a multiple linear regression model to map the relationship between our input parameters and our outcome of interest (income) and supplemented our findings with independent calculations by measuring each parameter individually. Ultimately, we found our multiple linear model to be a decent fit that had income scaling the most heavily with occupation type, and our supplemental data conveyed similar results for the strata we observed. We compared this against modern trends and found many similarities, but also that some limitations which could be addressed to finetune the results.

Introduction

When contemplating on the topic of societal balance and conditions, one of the defining characteristics that always comes to the forefront is individual salaries. This has been a longtime discussed subject matter and most individuals in society work towards improving in certain areas they believe will result in higher percentage chances of increasing their income. While there exist general conceptions of the various factors that contributes to income numbers, it is important to back up these theories with statistical evidence and analysis.

With this context in mind, our goal in this report is to identify clear social trends through an approximated statistical model that helps us interpret which of these factors significantly contributes towards individual incomes. To supplement this, we will be predicting income by deriving a regression formula for the model. The dataset we will be drawing from in this paper is the Canadian General Social Survey (GSS), an annual survey designed to gain insight on the social living conditions of Canadians.

The GSS has a sample population of approximately 20,000 respondents (a number that has changed over time), targeting people over the age of 15 in the ten provinces of Canada. The sampling frame consists of telephone numbers registered in Statistics Canada's common telephone frame. GSS collects its data through a combination of voluntary online questionnaires as well as telephone interviews, which utilizes volunteer-based sampling and simple random sampling techniques, respectively.

Returning to our outcome of interest and objective of this report, a few of the common factors that are believed to be associated with income includes highest level of education attained, work/volunteer experience (that comes with age) and job professions. Intuitively, it makes sense that all three of these variables will have some sort of correlation with income. For instance, we would expect that higher levels of education and increased experience should hypothetically result in higher income levels. Meanwhile, higher paying job

professions are more of a subjective matter, but our expectation is that the higher educated professions such as doctors, professors and researchers to be topping the salary charts.

By the end of this report, we will be able to confirm whether these hypotheses hold true by using these three variables as parameters for constructing a multiple linear regression model with respondent income as the outcome of interest. Our goal is to evaluate the estimates of this model to test if each of the specified parameters have the expected influence levels on income.

The report will consist of the follow sections: a discussion and justification on the dataset selected, a description of the selected regression model, model results including independent analysis measuring income against each of the isolated parameters, and a detailed discussion of our model.

Data

The data we selected was from the 2017 GSS statistical compilation. Per the official documentation link in the references, the study was carried out across a period of 9 months in 2017 to identify the general family well-being of Canadian societies. The target population was all Canadian households that contained a member over the age of 15, particularly for certain parameters that would have been difficult to assess under that range. The sampling frame was all eligible households that had an available telephone entry under the address system across Canada. The actual sample was around 21,000 households that ended up responding to the call, which is a decently large sample size in general. However, considering the context of the survey being national, it's debatable if an even larger sample size would have been ideal.

The study was conducted across all of Canada, but exempted the territories as well as full time residents of institutions. The data was conducted over 10 stratas which represented each province respectively. Overall, the study was quite successful as the respondent rate was around 91%. Non-respondents were typically just not accounted for, like that of phone numbers related to corporations or institutions.

As for the data collected, the featured parameters were mostly pertaining to general societal information such as occupation, region, and family size. We decided to focus on a set of parameters that corresponded with the idea of success in our modern society. Our primary parameters were that of **age**, **education level**, **occupation** and **income of the respondent**.

We felt this topic was an appropriate selection because understanding what factors into income baselines is a key topic that's widely discussed and analyzed. If the data is accurate, it would be a great aid for people at a certain level to understand what they need to do to advance to the next, whether it be education, occupation type, or other.

As for the parameters, we felt the three above would be the most interesting to consider, due to their significance in influencing income based off of anecdotal evidence. Occupation is a rather intuitive one. Education is also very easy to understand, but it's also interesting to know if there are diminishing returns of some sorts; the question of 'is college really worth it' and such. Age is a bit less intuitive, because we would expect it to steadily increase as one aged. However, with the modern industry boom and salaries for certain fields skyrocketing out of the gate for young recruits, it's interesting to observe if perhaps that notation is different from what we expect.

There were other parameters we found very interesting as well such as region, birth place, and family size. Another parameter we were also interested in was income of entire households compared to individuals, which could offer a very solid comparison point between the two and offer insight into why a certain parameter might model a particular way. However, to limit the scope and prevent the study from becoming too general, we narrowed it down to these three main inputs.

Ultimately, calculating a multiple regression on these variables will prove to be quite useful, as it gives a side by side comparison of which factors have a higher influence in relation to others, and to what extent they do in their respective sub-clusters.

Model

Our main model will be a **multiple linear regression model** to analyze the extent to which the parameters for age, education, and occupation affects the average value for respondent income. That is, we assume these three variables will linearly scale together with the probability of getting a higher income.

$y_i = B_0 + B_1 x_{1i} + B_2 x_{2i} + B_3 x_{3i} + e, i = 1...N$

where x_1 = age, x_2 = education, x_3 = occupation, y = income bracket, N = occurrences observed

Please note that we will not be performing analysis for family income. For the purposes of our analysis, the respondent income is definitely the more important parameter.

The function we will be using to compute the model is glm(), a general linear modeling function which computes a regression over all inputs and gives a respective weight to our y, expected value of the respondent income. The strength of this function is it's intuitive and gives useful analytical data such as R-scores and Std. Error, which we can factor into our findings afterwards.

The weaknesses are mostly that while the variables can be analyzed in such a manner, it may not be the most nuanced fit as the parameters may not necessarily have strong relationships with one another. For instance, if age factored into the output value and education level did not, we can only see that age is more significant and education level is less significant in this relative comparison. However, if we had education level, degree type, and university rating, we could see the dependencies more clearly and have a stronger confidence of the weights as a result.

We felt that using this model was appropriate for a few reasons. The first was that the model would be easy to interpret and understand based on multiple inputs from us. We could concisely create a formula that mapped out the relationship between all parameters together. As output was numerical and not binary, we could not use a logistic model. We certainly could have attempted a Bayesian model, as there are pre-existing biases for our parameters that we could consider as our priors. However, as the data is now somewhat old and we are not strongly versed in the subject matter, we felt it would be better to avoid assumptions beforehand.

Additionally, to support this data, we will also perform analysis on each factor independently. This will provide insight on any details that could be overshadowed by the main model. Age vs income will be represented as a **linear regression model**, as it is a numeric field.

$y_i = B_0 + B_1 x_{1i} + e_1, i = 1...N$

where x_1 = age, y = income bracket, N = total age range

For the latter two, we could use categorical regression, but this would be difficult in the case of education (e.g. numeric difference between high school and college versus college and a PHD) and impossible in the case of occupation. As such, we will use **box plots** to represent the data. Having categorical parameters contains some weaknesses for analysis in this regard when we're using a linear model. For instance, we can compare the groups against one another, but we can't extrapolate the data for other groups like we could for numerical parameters like age.

Finally, it's also important to note that our parameter of interest, income, is not a numeric field but rather a string range. This would be difficult to perform linear regression on, so we must opt to convert them into numeric fields by assigning it the median of the range.

For instance, the range "$50,000 to 74,999" would be classified as the value 62,500.

This comes with some definite downsides as we'll see later, but it's a decent estimator for the data and will help simplify some factors of our data. While it would have been ideal to have numerical values, it's understandable as many people cannot easily discern their exact incomes, nor would they necessarily be comfortable with it. It's also easily to just have multiple choice for such a large scale study, particularly for data compilation and filtering.

Results

First, let's set up the data. We'll be importing the dataset 'gss.csv' and proceed to filter out the parameters we need.

age	education	occupation	income_respondent
52.7	High school diploma or a high school equivalency certificate	Sales and service occupations	37500
51.1	Trade certificate or diploma	Trades, transport and equipment operators and related oc...	12500
63.6	Bachelor's degree (e.g. B.A., B.Sc., LL.B.)	0	37500
80.0	High school diploma or a high school equivalency certificate	0	62500
28.0	College, CEGEP or other non-university certificate or di...	Sales and service occupations	12500

Age vs Income: Simple Linear Regression

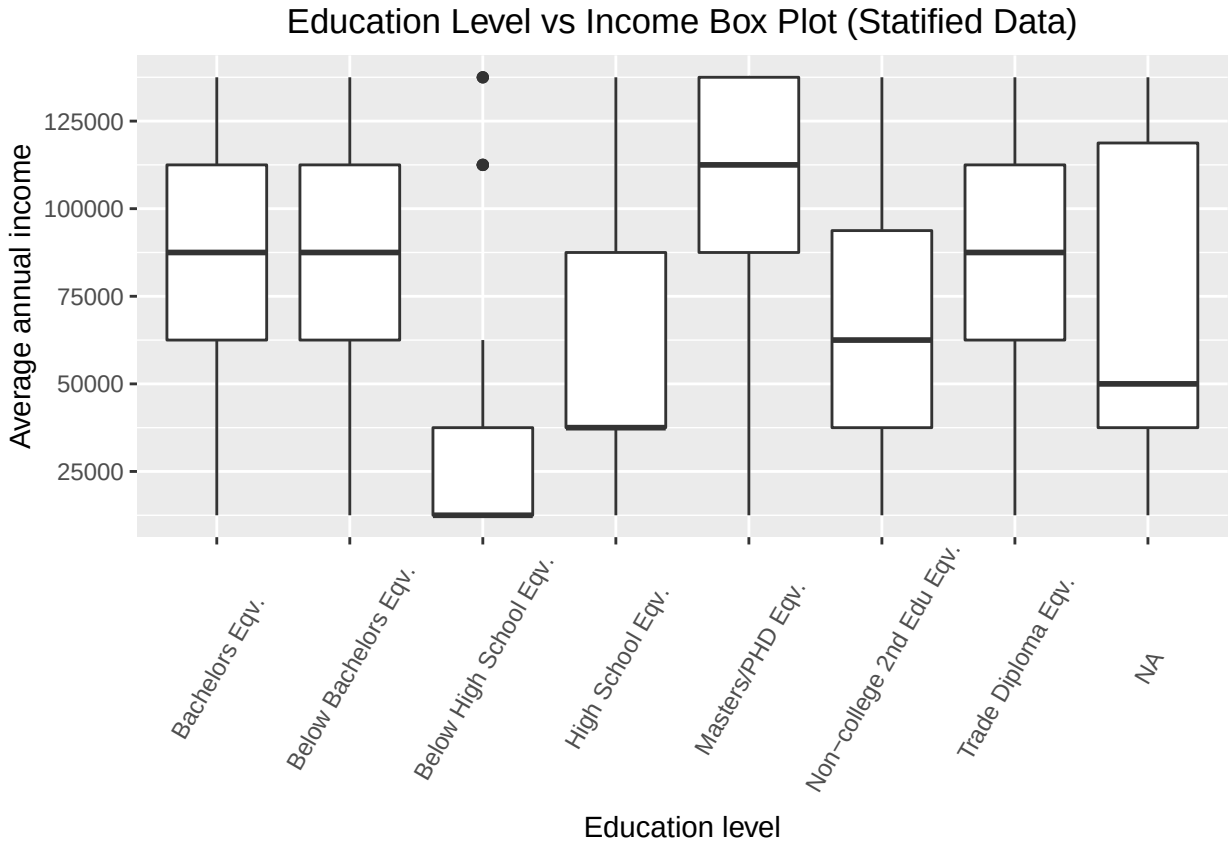
For visualization purposes, we'll perform a stratified sampling method where we'll take the first x entries of each of the 6 income brackets, based on their position from the top of the frame. Otherwise, there would be so many points that it would be impossible to discern anything from the plot. In our modeling, we chose $x = 100$ which seemed like a sizable number that was easier to visualize.

As we can see, a linear regression model really doesn't work that well for age and income bracket, the line of best fit is almost horizontal and the graph itself is all over the place. However, we can get some more meaningful information with a curve of best fit.

That's somewhat better, although the variance is still quite large. The general result is that the income for an individual peaks when they're around 40-45 years old, and declines from there on either side.

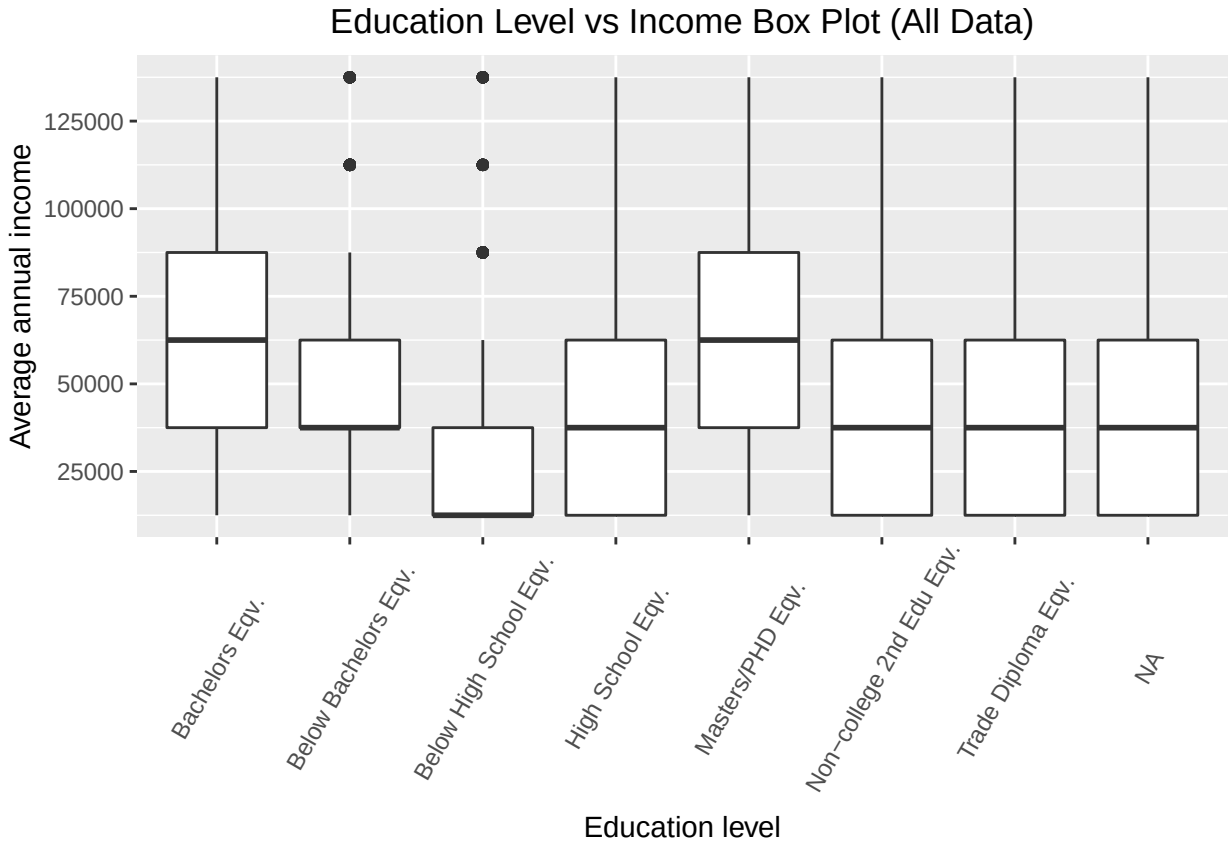
Education Level vs Income: Box Plot

Next, we'll plot the categorical relationships between education and income. A good way to visualize this is in the form of a box plot. As with the age variable, we'll stratify the large pool of data and take 100 of each strata for analysis.



The data above shows that for education, income scales well with the level of education. PHD/Masters are at the top, followed by the Bachelor equivalents, all the way down to below high school levels where it's really quite low. What's interesting to not is that as education level increased, the range of income became more stable as well. With high school, while the range of income scaled up to 87500, the vast majority of the entries were stuck at 37,500. This is not the case for most of the other higher education entries.

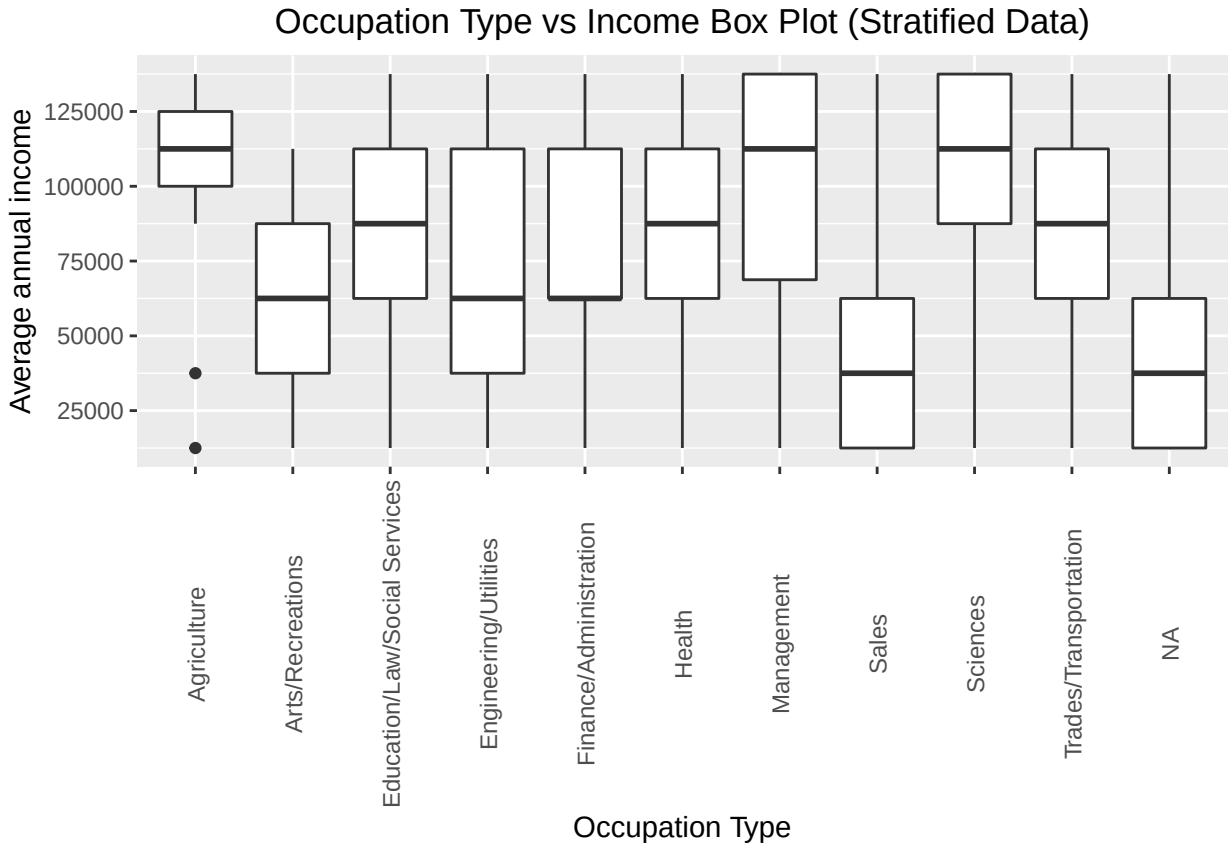
However, now let's try and model this with all the data points, how will it change?



It's a bit unfortunate that our model only models the boxes at certain intervals, but the data presented is still plain to see. When we take a particularly large sampling size, the expected value of income becomes far less drastic across each education level. Of course, Bachelors and Masters/PHD does still have an edge up, but a lot of the others, such as trades, begin averaging out to the same levels as high school graduates. While the nuances are not necessarily accounted for, it's still a bit of a surprise to see.

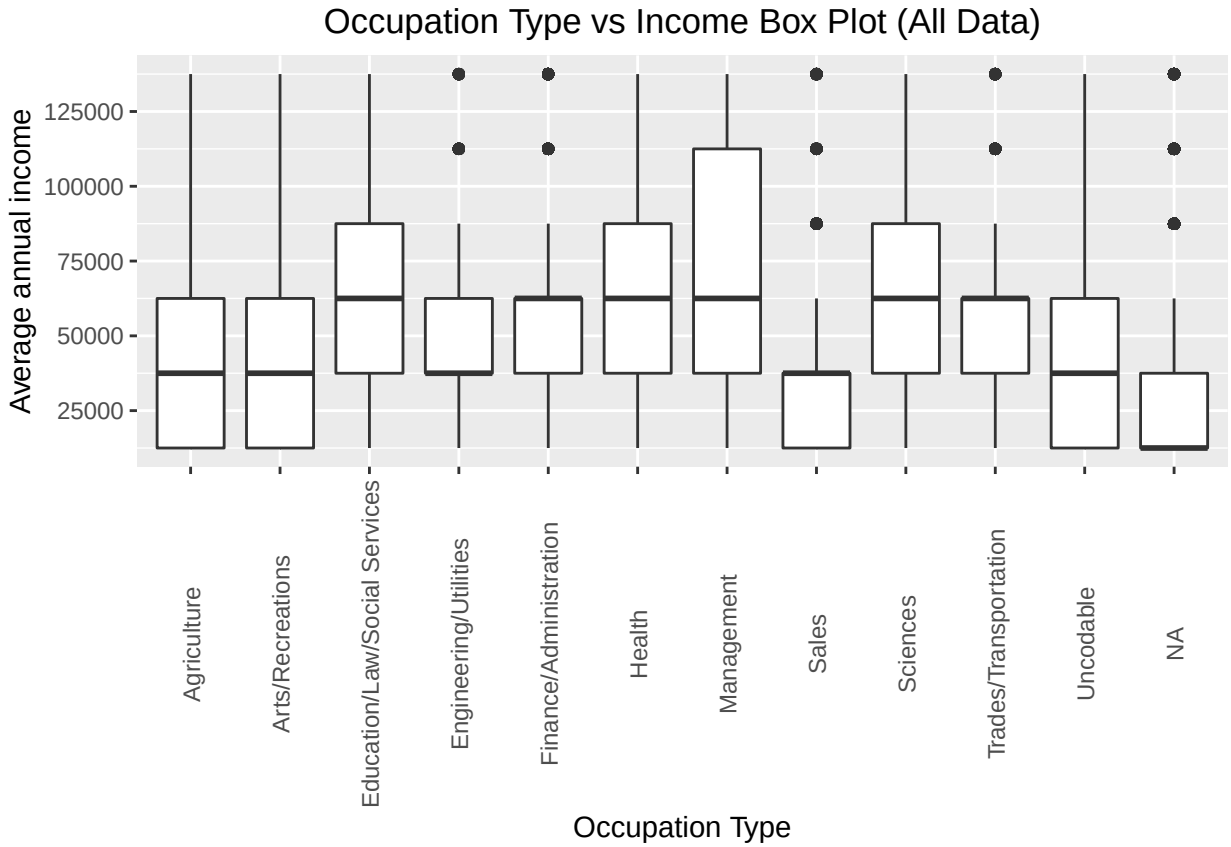
Occupation Type vs Income: Box Plot

Next, we'll graph a similar box plot for the occupation parameter. Again with the stratified data first, followed by the entire dataset.



This box plot shows that there is a decent spread across the occupations, some expected, and others not as much. For instance, sciences average in the high range of 112,500+, which makes sense by today's standards. However, others like agriculture, for instance, scale up to the same level which is quite surprisingly. However, the range is very small, meaning it's likely a bias from our stratified pickings.

Now let's plot the entire data.



Like with education, aside from a few fields like sciences or health, most have scaled down in income on average and is in line with most of the other occupations. However, there are outliers. Notably, management has an extremely large range, with the highest estimates scaling far above the other fields. To picture this, imagine CEOs of large corporations versus CEOs of smaller shops on smaller streets.

Age, Education, and Occupation vs Income: Multiple Linear Regression

Finally, we can perform our multiple linear regression on the 3 parameters above for income together. We will not be able to plot this easily, so we will just represent it as a function based on the results of the calculations. We've summarized the outputted coefficients below as a reference:

Parameter Label	Estimates
Intercept	8416.8632
Age	409.9854
Bachelors Eqv. (Education)	10781.8593
Non-college 2nd Edu Eqv. (Education)	-1735.3966
High School Eqv. (Education)	-6496.3748
Below High School Eqv. (Education)	-14374.0913
Trade Diploma Eqv. (Education)	-751.1717
Below Bachelors Eqv. (Education)	5212.8568
Masters/PHD Eqv. (Education)	21396.6922
Finance/Administration (Occupation)	24896.3991
Health (Occupation)	29017.7139
Management (Occupation)	38871.6244
Sciences (Occupation)	39837.7869
Agriculture (Occupation)	28896.1813
Arts/Recreations (Occupation)	7063.7567
Education/Law/Social Services (Occupation)	25922.0904

Parameter Label	Estimates
Engineering/Utilities (Occupation)	25945.9995
Sales (Occupation)	13062.6810
Trades/Transportation (Occupation)	31806.4142
Uncodable (Occupation)	10818.2223

With the above results, we can formulate our regression model formula of:

$y_i = 8416.86 + 409.99x_{1i} + 10781.86x_{2_1 i} - 1735.40x_{2_2 i} ... + 21396.69x_{2_7 i} + 21396.69x_{3_1 i} + 29017.71x_{3_1 i} ... + 10818.22x_{3_{11} i}$,

where y_i = income for i, x_{1i} = age for i, $x_{2_1 i}$ = edu category 1 for i, etc. , $x_{3_1 i}$ = occ category 1 for i, etc.

This equation will be interpreted later in the discussion section.

Discussion

In this section, we will evaluate the models and graphs created in the previous sections while also reflecting back on our goals and hypotheses stated near the beginning of the report.

Let's begin with the age versus income relationship. We expected some type of association to show as people aged and gained more experience with the amount of income they generate. Contrary to this expected behavior, there is no clear linear correlation between these two variables as it was not appropriate to fit a simple linear regression model. The high variance from this graph is a direct result of the approach we took to transform the categorical income variable into numerical values, something we will further discuss below in the limitations of the model. Taking a closer look at the results, it is evident that there are clear peaks in age when it comes to predicting incomes. The middle-age class ranging from roughly ages 35-50 accounts for the highest percentage of the upper percentile of income values $> 100{,}000$.

When connecting this to the real world, it is possible to explain these results and the curve of best fit (shown in the parametric model) as follows. Individuals less than the age of ~22 have minimal work experience and are very likely to be in the process of building their fundamental knowledge through their studies. Working a part-time job while in school is also prevalent, and these two factors combined help explain why the curve bottoms out for the cluster of points in the 0-24,999 income range for respondents in this age group.

As expected, the average respondent income rises after this age group as students transition into the workforce and gain valuable experience, eventually peaking in the aforementioned peak age group. At last, we expect people to hit the retirement stage, in which their income will consist of pension plans, returns on long-term investments, retirement savings plans, etc. This accounts for the gradual decline of the curve given that the descent was more of a level-off as opposed to mirroring the sharp ascent that we saw for income values up to the age of ~40, in the opposite direction. All this real-world context can hopefully be helpful in justifying why the curve of fit behaves the way it does and why plain linear regression failed to model this particular relationship.

Next, we look at our second predictor variable, education level. There is not a ton of additional real-world context that we need to put into place here, as the results of the box plot do confirm our initial expectations that there is a strictly positive correlation between respondent income and level of education attained.

Looking at a relevant report by Joseph Berger and Andrew Parkin studying the "Value of a Degree", the level of attained education degree was measured against income earnings in linear and rate of change models with respect to isolated demographic groups (e.g. college males and females, university males and females, Aboriginals vs Non-Aboriginals). The end conclusion of the report argued that the benefits of completing a post-secondary degree has in fact been growing in recent decades, with "degree- and diploma-holders [being] financially better off now relative to non-graduates than they were 25 years ago". This is all despite it being less scarce than any other time in Canadian history to hold a post-secondary degree. The report also detailed the inverse holds true, as the growth of unemployment rates for 25-34 year old individuals was highest by a

wide margin for people with lower levels of education, resulting in the conclusion that the positive rate of growth of education against earnings conversely ties into the overall strength of the Canadian economy.

The final parameter we analyzed was plotting occupation against income. Unlike education level, each category of occupation types is entirely independent of each other. It is far more difficult to quantify the hierarchy of occupation types as easily as education levels which does possess this hierarchical structure. The primary reason for why this is the case, as can be seen from the box plot, is that there are wide ranges of income levels for each quantified occupation type. We observe that the the highest paying occupations include education/law, health, management, and sciences, which is further supported by data from Statistics Canada citing highest average annual salaries by profession as follows - physicians, dentists, engineers, managers, professors, and so on.

When relating to the real-world, an interesting thing to note for these higher paying jobs is that it is well-known that careers in health care, education and scientific research absolutely require the highest level of education (masters/PhD degree). Thus, we can conclude that the parameters we picked are highly dependent of each other, and both education and the specified groups of occupations scale with an increase in income. This would justify our variables as being reasonable choices of parameters to estimate income in the regression formula.

Finally reaching our multiple linear regression equation estimating income, recall we obtained:

$y_i = 8416.86 + 409.99x_{1i} + 10781.86x_{2_1 i} - 1735.40x_{2_2 i} ... + 21396.69x_{2_7 i} + 21396.69x_{3_1 i} + 29017.71x_{3_1 i} ... + 10818.22x_{3_{11} i}$,

where y_i = income for i, x_{1i} = age for i, $x_{2_1 i}$ = edu category 1 for i, etc. , $x_{3_1 i}$ = occ category 1 for i, etc.

We can interpret the overall results as follows:

At **age 0** with **no education** and **no occupation**, a person is estimated to have an **average annual income of $8416.86** (obviously unrealistic but such is the nature as being the intercept value).

For every year an individual ages, it is estimated that their average annual income increases by approximately **$409.99**.

Depending on a person's highest level of education attained, their average annual income takes a modifier between $ **-14374.09 to +21396.69**.

Depending on an occupation type, an individual's average annual income takes another modifier between $ **+7063.76 to +39837.79**.

As an example, consider a person who is **25 year** olds who is holding a **bachelors degree** in the **sciences**. With this equation, we estimate that this particular individual is set to learn $8416.86 + 409.99 * 25 + 10781.86 + 39837.79 = 69286.26$ dollars, on average.

To summarize the findings based on the regression model, we can reach some of the following conclusions. From analysis of our model as well as drawing from other relevant studies, a lack of quality education will have the most significant negative effect on a respondent's income level (as can be seen by the large modifier value). What we can infer from this is that external factors such as social living conditions as well as family expectations associated with education levels will then translate to influencing income. In other words, an individual's initial environment and family circumstances can have a great effect on their chances at elevating their social status (income) because they have additional obstacles to overcome when setting out to acquire an education baseline. Ultimately, having a quality education is the foundation on which other influencing factors are built upon.

Weaknesses

One of the main weaknesses of this analysis that was briefly mentioned earlier was converting the outcome variable, income, from a categorical variable to defined numerical values. Although this is a necessary

step that was required to perform regression, one of the main weaknesses to this approach is the slight misrepresentations of the data, which is not 100% indicative of the original dataset.

Consider the following potential problem for the income range between 25,000-49,999 classified as a salary of 37,500: for this given range of salaries, a situation could arise where more income data points lie on the higher end of the spectrum and is underestimated at a value of 37,500. Given that the average income of Canadians over the age of 16 in 2017 was 47,800, it is clear that this is a very plausible scenario and a decent number of individuals end up being misrepresented.

Although this can be a clear weakness in the analysis, when looking at the alternatives, it is still the best available option to help us transform this data numerically. When considering other options taking any other value (mean, arbitrary number, highest frequency of each range) to classify income values of each range numerically would have far worse consequences. To see why this is the case, an arbitrary number would not make any sense as it would not be based on the data itself. In addition, taking the mean or highest frequency incomes to classify the range would disrupt the scale of the outcome variable, leading to poor modeling of the dataset when attempting to fit the data and performing regression calculations.

In regards to areas of improvement, one alternative option would be to take household/family income as the outcome of interest to replace the income of the respondents. It can be argued that using data for family income accounts for previously discussed unaccounted factors like household size, which impacts the cost of living and as a result, income numbers. These external factors are not taken into consideration when only taking individual income numbers. However, the major drawback with a new approach like this lies in the fact that we can no longer measure the parameters we picked in age, education, and occupation which contradicts the purpose of the report. Thus, despite presenting certain benefits, taking household income as the new outcome of interest would require a brand new model altogether, and would be at least interesting to consider.

For other potential improvements, testing for additional variables to sharpen the presented multiple regression model formula could be a valid strategy. However, it is ideal that parameters can be linked together as we have here so we ended up limiting the scope of the study to our three main inputs. Other parameters worth considering could include nationality, location, cultural background (race, religion, etc.) and mixing and matching between these parameters could lead to more interesting results and possibly a more precise regression formula. On a surface level, it would intuitively make some sense that these factors could influence income in some way and be worth experimenting.

Next Steps

As mentioned above in the areas of improvement, it would be worth to consider performing a different set of analysis with household income as the outcome of interest and contrast those results with the current model. A quick brainstorm of the parameters involved with a model predicting household income could feature factors such as household size, martial status, average age, and more. Such a model would seemingly complement the current study extremely well and may potentially produce intriguing results that we can take to re-evaluate our conclusions on individual respondent incomes.

References

Beaupré, P. (2020, April). General Social Survey. Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

Berger, J. & Parkin, A. (n.d.). The Value of a Degree: Education, Employment and Earnings in Canada. Retrieved from https://www.chs.ca/sites/default/files/uploads/value_of_education_with_degree.pdf

Financial Consumer Agency of Canada. (2019, February 01). Sources of retirement income. Retrieved from https://www.canada.ca/en/financial-consumer-agency/services/retirement-planning/sources-retirement-income.html

General Social Survey: An Overview, 2019. (2019, February 20). Retrieved from https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm

Government of Canada, Statistics Canada. (2020, February 24). Income of individuals by age group, sex and income source, Canada, provinces and selected census metropolitan areas. Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1110023901

Harris, P. (2018, February 15). Stats Can: The highest (and lowest) paying jobs in Canada for 2018. Retrieved from http://blog.careerbeacon.com/stats-can-the-highest-and-lowest-paying-jobs-in-canada-for-2018/