A Study on the Canadian Society's Health Level

Cindy Gao(1005223410), Yanrong Huo(1004720965), Aiting Zhang(1004926066)

October 19th,2019

Code and data supporting this analysis is available at :https://github.com/aitingzhang47/STA304-Group33-Problem-Set2.git

Abstract

Although Canada has always been ranked as one of the highest for the best quality of life in the world, its consistent rapidly changing demographic profile, along with its social, political and economic issues, has led to many serious concerns regarding one's health and well-being in the country. Canada, as one of the most developed countries in the world, is in a very fast-paced environment that feeds off of the innovation and determination from its people, thus potentially leading to a highly stressed and unhealthy society. As the upcoming generation in this country, this has led us to want to further investigate this potential health and wellness problem in Canada. We want to see if there indeed is a large amount of people who are stressed in this economy, and if so, what are the major causes and/or factors. Throughout this investigation, we will be building a linear regression model through some stratified sample work, to see if there are indeed specific factors that play a role in one's health. From the end results, we can get a rough idea of which factors can improve or deteriorate one's health, thus letting us and the society know what can be done to hopefully improve the society's health levels as a whole.

Introduction

Our goal in this study is to find out and investigate which personal and lifestyle factors have an impact on one's health. For this investigation, we will be obtaining data from the 2016 General Social Survey (GSS), which is a program designed as a series of independent, annual, cross-sectional surveys, each covering one topic in-depth. In 2016, the GSS was on Canadians at Work and Home. Using six independent factors found in the 2016 GSS, including one's age, sex, household size, and how long they work in a week, if they play a sport, how often they are online, their eating habits and their stress levels, we will create a linear regression model to predict one's health levels. From there on, we will be evaluating our findings to see which factors actually have a direct effect on one's health levels, and hopefully providing some tips on what can be done to improve the health level in Canada as a whole.

Data

```
## -- Attaching packages ---------------------------------------------------------------------------- ti
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
## -- Conflicts ------------------------------------------------------------------------------ tidyvers
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Loading required package: grid

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loading required package: survival

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart

## Loaded glmnet 4.0-2

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

##
## -- Column specification ----------------------------------------------------------------------------
## cols(
##   caseid = col_double(),
##   age = col_double(),
##   sex = col_double(),
##   household_size = col_double(),
##   work_perweek = col_double(),
##   regular_sports = col_double(),
##   frequency_technology = col_double(),
##   eating_habits = col_double(),
##   stress_level = col_double(),
##   province = col_character(),
##   health_condition = col_double()
## )
## tibble [19,609 x 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ caseid             : num [1:19609] 1 2 3 4 5 6 7 8 9 10 ...
##  $ age                : num [1:19609] 5 5 6 6 2 3 4 3 1 4 ...
##  $ sex                : num [1:19609] 0 0 0 0 1 1 0 1 0 0 ...
##  $ household_size     : num [1:19609] 2 1 1 2 2 5 1 4 2 6 ...
##  $ work_perweek       : num [1:19609] 3 3 NA NA 3 3 3 3 4 3 ...
##  $ regular_sports     : num [1:19609] NA 0 NA NA NA 0 NA NA 1 NA ...
##  $ frequency_technology: num [1:19609] 5 5 5 1 5 5 4 5 5 5 ...
##  $ eating_habits      : num [1:19609] 3 2 4 3 5 4 5 4 4 3 ...
##  $ stress_level       : num [1:19609] 5 3 5 3 4 2 3 3 2 3 ...
##  $ province           : chr [1:19609] "Alberta" "Ontario" "Manitoba" "Quebec" ...
##  $ health_condition   : num [1:19609] 3 3 4 2 3 4 4 3 4 3 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   caseid = col_double(),
##   ..   age = col_double(),
##   ..   sex = col_double(),
##   ..   household_size = col_double(),
##   ..   work_perweek = col_double(),
##   ..   regular_sports = col_double(),
##   ..   frequency_technology = col_double(),
##   ..   eating_habits = col_double(),
##   ..   stress_level = col_double(),
##   ..   province = col_character(),
##   ..   health_condition = col_double()
##   .. )
```

All data chosen and used for this study is from the 2016 GSS survey on Canadians at Work and Home, conducted through August 2nd 2016 to December 23rd 2016. The target population, which is all of those who are qualified for the 2016 GSS survey, was all persons in Canada above the age of fifteen, excluding residents of Yukon, Northwest Territories, Nunavut, and as well as anyone who was a full-time resident of institutions at that time.

Taking and working with data from the 2016 GSS is a great strength for our study as there is a full-set of data and so many possible variables to work with. Along with that, the six independent factors/variables all had detailed responses and were not just simply yes or no answers. One drawback of using the 2016 GSS survey now is that it may be outdated as we are now four years past 2016.

To carry out this study, the method chosen to reach the maximum potential amount of the wanted target population was through the lists of telephone numbers in use provided by Statistics Canada, and the Address Register (AR), thus making those on these two lists the sample frame, which are the people we can actually potentially reach. Henceforth, the sample population was the pool of the qualified survey takers who were actually successfully reached and completely responded to the survey.

Model

To further investigate this case, the same stratified sampling technique was used as in the 2016 GSS Survey. Stratified sampling is a method of sampling, where the target population is divided into subgroups. From then on, in each subgroup, we perform random sampling, which is where each individual and their responses were picked and gathered at complete random. In this case, we divided the subgroups by the desired Canadian provinces, and from there on performed random sampling within each province. In our code, we calculated each subgroup/province's sample size by researching each province's respective population and then subtracting the amount of people aged below fifteen, and those who reside in insitutitions.

With the data obtained through our stratified sampling design, we were able to build a linear regression model to predict one's health level. From the gathered data in the survey, we then picked a few more specific variables that we wanted to further study to see if they do indeed have an effect on one's health. The selected variables were the individual's age, sex, household size, how often they work a week, if they practice sports regularly, their amount of technology usage, their eating habits and their self-ranked stress levels.

One's age and sex were included in our model, as we believe we should include some basic and fundamental information to get a quick insight to each individual. At the same time, age and sex can both play a role in one's health. Our hypothesis is that in general someone in their teens is thought to be more healthy than someone in the seniors category, as well as perhaps males are less healthy than females, as men tend to pass away earlier than women. Through this linear model, we can inspect to see if there is indeed a relationship between age and sex and one's health and if there is how we predicted it to be.

Next, we chose to take a deeper look at how often one works a week, how often they are online and if they play a sport. We specifically chose these three variables to further investigate because we believe they serve as a measure of one's leisure time. This was an important factor for us to consider because our hypothesis is that the more time spent at ease, the more happiness and thus result in a healthier person. At the same time, each of these three definitely can directly affect one's body, for example, working too much and too much technology is bad for one's mind and eyes, whereas playing sports regularly is shown to improve your overall body and health. Therefore, through our model, we can investigate if there is a direct relationship between these three variables and one's health, and what kind of relationship it is.

Lastly, we also included one's household size, eating habits, and stress levels in the model, as these are direct factors that can also potentially directly affect one's health level. Our assumption is that the healthier one's diet is, the smaller one's household size is, and the lower the one's stress level is, it will result in a healthier body. Once again, we can use our linear regression model to see if there does exist a linear relationship between these factors and one's health, and how they affect one's health.

Using the gathered data and the specially picked variables, we built a linear regression model. Most of the picked variables were categorical variables, posing a threat to our linear regression model. Therefore, prior to building our linear regression model, we changed all our categorical variables into numerical variables. For most of the variables, we were able to rank them on a number scale, beginning at 1, where 1 usually represents the lowest amount/poorest given choice for the category, and the highest number for each variable represents the highest amount/best choice for the category. For example, a 1 in one's household size represents one person per household, and 6, which is the highest choice in the household category, represents six or more people in a household. The two variables that we did not treat with an order are sex and whether the individual plays a sport or not. Instead we treated them as dummy variables, where 1 represents a male and 'Yes', and 0 represents a female and 'No'.

Here is a list briefly describing how we ranked and converted all our categorical variables to numerical variables: -eating habits is ranked from 1-5, from poor to excellent, -stress level is ranked 1-5, from extremely stressful to not at all stressful -amount of time spent online is ranked from 1-5, from no usage in the past month to daily -amount of hours worked per week is ranked from 1-4, from 0-15 hours to 41hours+ -age is ranked from 1-7, from 15-24 years of age to 75 years and above -household size is ranked from 1-6, from one person per household to six and more people per household (as mentioned as above) -sex is classified as a 1 for males and 0 for females (as mentioned as above) -whether the individual plays a sport or not is classified as a 1 for 'Yes' and 0 for 'No' (as mentioned as above) -health level (our variable of interest) was ranked from 1-5, from poor to excellent

From there, we were able to build our linear regression model with our chosen variables, and derive a formula of the following: Health = 0 + 1(age) + 2(as.factor(sex)) + 3(household_size) + 4(work_perweek) + 5(as.factor(regular_sports)) + 6(eating_habits) + 7(stress_level). In our formula, 0 is the intercept of our linear model, which is 1.647083. This means that given there was no information provided from the six chosen independent variables, one's health level will be 1.647083 out of a max scale of 5. The rest of the betas (1 … 7) are all the coefficients of the respective chosen variables.
Lastly, to verify if there does indeed exist a linear relationship between one's health level and the six chosen independent variables, and if our linear model actually works, we computed the R2, which is the coefficient

of determination, and an F statistics table. The coefficient of determination is the percentage of variation that can be explained by the model, therefore, in our model a high R2 means the variable of interest, one's health level, can be explained very well by our six chosen independent variables, and a low R2 will mean that our six chosen independent variables are not very useful in determining and/or explaining one's health level. This will be thoroughly discussed below in the Results section.

Results

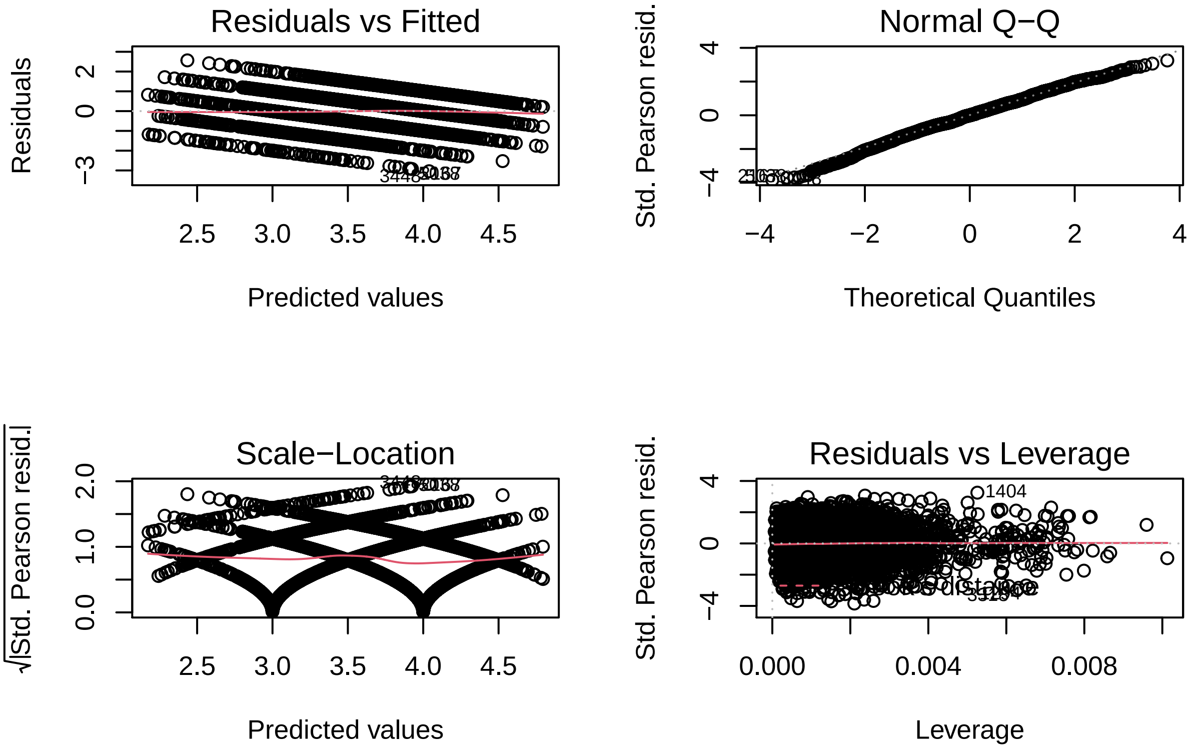
```
## Warning in summary.lm(svyglm_stratified): residual degrees of freedom in object
## suggest this is not an "lm" fit
```

```
##
## Call:
## svyglm(formula = health_condition ~ age + as.factor(sex) + household_size +
##     work_perweek + as.factor(regular_sports) + frequency_technology +
##     eating_habits + stress_level, design = design_stratified,
##     family = "gaussian")
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6822 -0.4577 -0.0009  0.4540  2.9525
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.647083   0.097871  16.829  < 2e-16 ***
## age                       -0.032753   0.007911  -4.140 3.52e-05 ***
## as.factor(sex)1           -0.038551   0.021453  -1.797  0.07239 .
## household_size            -0.006840   0.008186  -0.836  0.40341
## work_perweek               0.043681   0.013100   3.334  0.00086 ***
## as.factor(regular_sports)1  0.206493   0.023342   8.846  < 2e-16 ***
## frequency_technology       0.018149   0.012721   1.427  0.15371
## eating_habits              0.411799   0.011517  35.757  < 2e-16 ***
## stress_level               0.147715   0.011249  13.131  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7921 on 5966 degrees of freedom
## Multiple R-squared:  0.228,  Adjusted R-squared:  0.2258
## F-statistic: 220.3 on 8 and 5966 DF,  p-value: < 2.2e-16
```

In this study, we want to investigate whether the linear regression model we built is good or not. First, we computed R2, which is the coefficient of determination and F statistics. As stated, the coefficient of determination is the percentage of variation that can be explained by the model, therefore, in our model a high R2 means the variable of interest, one's health level, can be explained very well by our six chosen variables, and a low R2 will mean that our six chosen variables are not very useful in determining and/or explaining one's health level. The R2 for our linear regression model is 23.91%, which is quite low, but however in fact very realistic. Although our coefficient of determination is on the lower side, our p-value from the F statistics for our linear regression model is less than 2.2*10^(-16). This is an incredibly small p-value, thus meaning we can reject the null hypothesis, which is an original default assumption that there does not exist a relationship in between one's health levels, and each independent variable respectively. Therefore, we can say that our linear regression model is indeed significant and that there does indeed exist a linear relationship between one's health and our six chosen independent variables. However, the small R2 does show that there is definitely room to adjust and improve our model's performance.

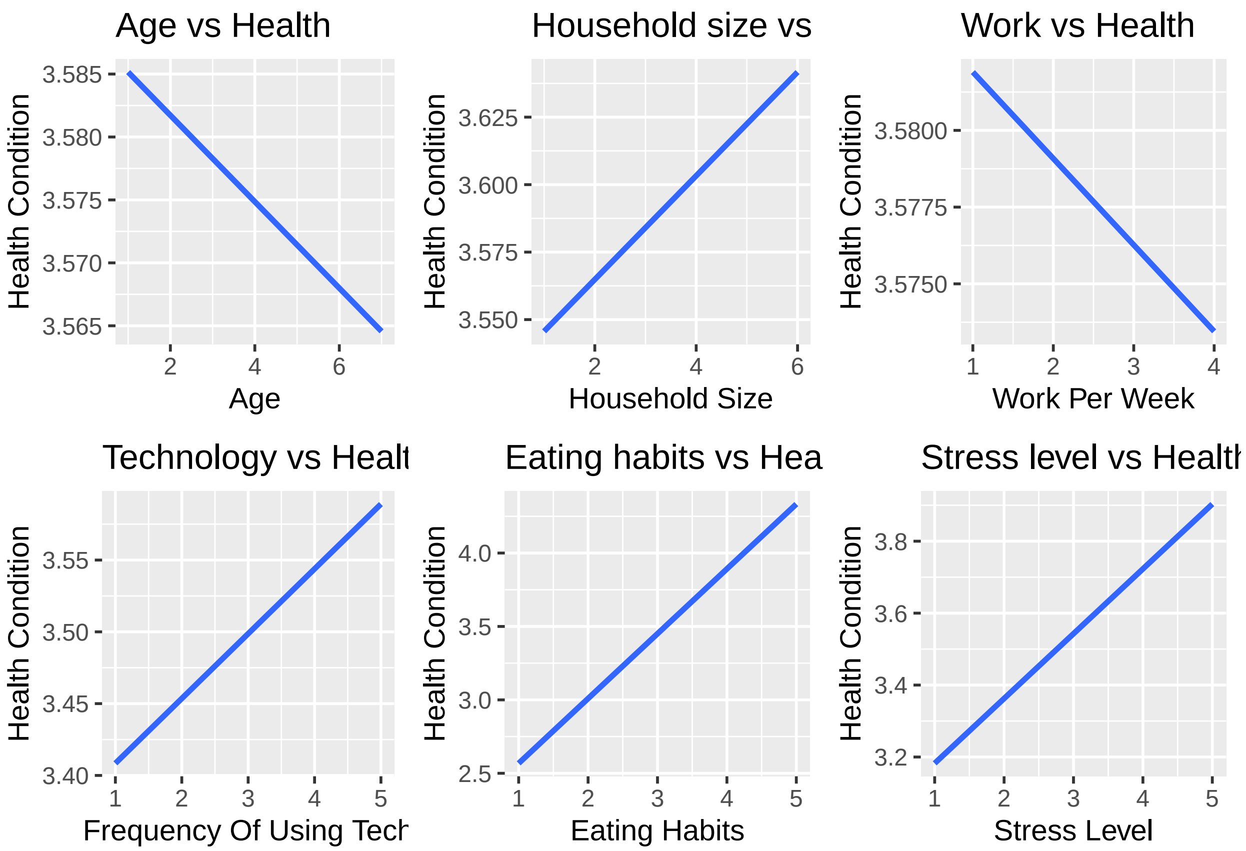
```
##                      age           as.factor(sex)            household_size
##                 1.209009                 1.094067                  1.103627
##             work_perweek as.factor(regular_sports)      frequency_technology
##                 1.088320                 1.064961                  1.122145
##            eating_habits             stress_level
##                 1.067995                 1.061053
```

We also checked our model's multicollinearity by using the function of vif. From this we discovered that the model does not have multicollinearity and all predictors in it make sense. Therefore, we do not need to remove any variables from our model.



Then, we constructed a diagnostic plot including a Residual vs Fitted plot, Normal Q-Q plot, Scale-Location plot, and Residual vs Leverage plot. The Residual vs Fitted plot, tests to see whether the relationship between the variable of interest, health level, and our six chosen variables are linear. From our Residual vs Fitted plot, we can see that the residuals spread equally around a horizontal line without distinct patterns, therefore the linearity assumption is appropriate. From the Normal Q-Q plot, it can be seen that most of the residual points rests along the dashed diagonal line, meaning that our data is for the most part normally distributed, thus strengthening our linear model even more. The Scale-Location plot shows whether the residuals are spread out equally, and in the end if there is a horizontal line, it means that the constant variance assumption is valid and that once again our model is strong. In our Scale-Location plot, it can be seen there is an almost-horizontal line, thus supporting the strength of our model. Lastly, a Residual vs Leverage plot is supposed to help in identifying extreme values in the models. In our Residual vs Leverage plot, it can be seen that there does exist a few points floating around, however, it can also be seen that most of our points are packed together, thus once again supporting the idea that we do indeed have a strong model.

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Then, we drew six linear models, and as it can be seen through each graph, there does exist a linear relationship between each six chosen variables and one's health levels, which is our variable of interest. To support this finding even more, we can look at each independent variable's p-value. From each independent variable's p-value, we can use statistical data to verify if there truly is a relationship between our independent and dependent variables, and if we are actually able to use this linear regression model to predict one's health level. From the table, it is clear that each independent variable's p-value, except for household_size, is smaller than 0.05, which means that we can reject the null hypothesis for the rest. Therefore, it shows that most of our chosen independent variables do in fact play a role in determining one's health and well-being level.

Overall, it can be seen through each evaluation that was conducted above that our model is indeed strong and does perform at a high level. However, at the same time, in many of our evaluations, such as the low R2, and the diagnostic plots, it is clear that there is room for improvement. This will be discussed more in detail in the Weaknesses and Next Steps section.

Discussion

From the six linear models, we can see a linear relationship between each of the six independent variables and one's health, thus proving that these six factors do indeed play a role in determining one's health. From the graphs, we can see that all have a positive relationship with one's health levels, except age and how long each individual spends at work. This makes sense and further strengthens our prior hypothesis. The only ones that surprised us a bit were the positive relationships between household size and amount of technology usage.

Our prior assumption with the amount of technology usage is that the less you use internet, the more active you are, thus the healthier you are. However, instead it can be seen that the more you use your devices, the healthier you are. This might be because, nowadays, people are on their devices as a form of entertainment and relaxation, thus increasing one's calmness and increasing health.

Similarly with the size of one's household, our prior assumption was that the more people you have in your family, the more stressed you are, thus the less healthy you are. However, from the graphs it can be seen that the bigger the household size, the healthier you are, which goes against our prior assumption. At this point, it is also important to keep in mind that from our investigation above, it was shown that one's household's size's p-value was greater than 0.05, thus meaning there potentially may be no relationship between one's household size and health.

With all findings in this study, it is safe to say that there are many various aspects that contribute to one's overall health status. From our study specifically, it can be seen that the amount of hours someone works per week has a major effect on one's health levels. With this information, a potential way to help create a healthier society is to implement lower working hours. Other possible solutions include enhancing more of the factors that have a positive relationship with one's health, such as having the Canadian Government encouraging a more healthy lifestyle, and releasing healthier food guides for its people.

Weaknesses

A small problem, already mentioned above, is that we gathered our data for the 2016 GSS Survey, which may be already outdated in 2020, as Canada's population grew and as stated before Canada has a rapidly changing demographic and economy.

Following the GSS Survey weaknesses, specifically for this study, the six independent variables were chosen out of our interest and our assessments on this topic. Along with that, in real life there may be hundreds of different aspects in life that have a potential influence on one's health, whereas in our simple linear regression we only used six of the potential factors. These are important weaknesses to consider as in reality there may be a bigger factor influencing one's health and we may have potentially completely missed it in our study. Another potential weakness is that each individual's health level was self-ranked. This may pose a threat because everyone may define "healthy" differently. For example, someone who has a healthy body but a very low mental health may rank themselves a 5, extremely healthy, whereas someone with a very high mental health but very weak body may also rank themself has a 5. These are some small weaknesses in our study, however, if more accurate data is wanted, these weaknesses must be fixed.

Next Steps

The first big steps to take after this study is to fix our weaknesses. First, we should send out a newer and updated version of the 2016 GSS Survey. Along with this, we can change the questions to be very detailed and precise, so that there is no confusion or misunderstanding between the researchers and the survey takers. This way we can reach out to everyone in all of Canada and have updated responses. This can be done through simply expanding the method of communications, which should be easily attainable in our day and age.

To combat our biases in choosing the independent variables, we should expand our perspective and research and investigate with more independent variables. This may be difficult because as said there may be hundreds of different factors affecting one's health. However, once again, we live in a very modern society with access to high-speed technology. Thus, instead of sticking to simple linear regressions, we can shift to more advanced softwares to aid us in analyzing more variables.

References

1. General Social Survey, cycle 30: 2016: Canadians at Work and Home. (n.d.).

2. Thevenot, S. (2020, January 22). Canada rated top country for quality of life in 2020: Canada Immigration News. Retrieved October 19, 2020, from https://www.cicnews.com/2020/01/canada-rated-top-country-for-quality-of-life-in-2020-0113587.html

3. Government of Canada, S. (2020, March 31). Census Profile, 2016 Census. Retrieved October 19, 2020, from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E