

Analysis on key factors affecting life satisfaction

Woolim Kim, Yena Joo, Guemin Kim

Oct 19, 2020

Analysis on key factors affecting life satisfaction

Woolim Kim, Yena Joo, Guemin Kim

Oct 19, 2020

Abstract

We use the 2017 General Social Survey(GSS) data to study the response of their _____ to _____. The target population includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. The frame population is everyone who is registered combining both landline and cellular with Statistics Canada's address registers, and the sampled population is whoever was reached via telephone. The target populations were stratified into 27 strata by geographic areas—mostly CMA— and simple random sampling without replacement of records was performed in each stratum.

We find. . . .

1. a strong
2. no overall effects in

Together, these results suggest that.

Introduction

Here is where you should give insight into the setting and introduce the goal of the analysis. Here you can introduce ideas and basic concepts regarding the study setting and the potential model. Again, this is the introduction, so you should be explaining the importance of the work that is ahead and hopefully build some suspense for the reader. You can also highlight what will be included in the subsequent sections.

****INTRO (1)** In recent periods, COVID-19 crisis has changed many things. Millions of people have lost their jobs, untold numbers of people fear they could lose their incomes any day, and people are having constant worries for their relates health and their own.

There are many factors that determines people's satisfaction and well being of their life. It could be people's health, income, family, religion, or other elements that makes them feel satisfied. Specially, in situation where we are living in full of uncertainty, people are feeling more stressed and those concerns give negative influence on their own's well being. Under the circumstances we wanted to take a close look on which factors of people's life has the most impact on people's satisfaction of their lives.

The data set we chose for this assignment contains responses of the General Social Survey conducted in 2017. The 2017 GSS data are monitoring changes in the living conditions and well being of Canadians over time. The 2017 GSS data includes variety of different variables such as, age, sex, education, feelings life, self rated health, income, etc. With the collected data we can analyze how people's satisfaction score is dependent and has significant relationship with different factor variables.

The following sections will use statistical methods to build new models and analyze the data. It includes cleaning the data into simpler model, linear regression modeling to find the significance of satisfaction score

and predictor variables, comparing and creating graphical visualizations.

Data

Introduce the data, explain why it was selected. Make sure to comment on important features and highlight any potential drawbacks to the data.

**** Guemin**

The data set we chose for this assignment contains responses of the General Social Survey conducted in 2017. The contents of the survey include some characteristics of diverse families in Canada, their socio-economic status, as well as other subjective information such as the respondent's life satisfaction and health conditions.

Since we want to identify some key factors affecting one's life satisfaction as mentioned previously, the focus of our analysis will be "Health and subjective well-being".

The reason for choosing the 2017 GSS is because it is the most recent¹ survey that includes the "Health and subjective well being" concept.

Since the original data set contains too many variables that are not necessary, we are going to clean the data set prior to analysis by removing them.

This is what our new data looks like:

```
## # A tibble: 6 x 9
##   ID   age sex  satisfaction_sc~ self_rated_heal~ self_rated_ment~
##   <dbl> <dbl> <chr>          <dbl>          <dbl>          <dbl>
## 1     1  52.7 Fema~             8             5             5
## 2     2  51.1 Male             10            3             3
## 3     5   28  Male             8             3             3
## 4     9  63.8 Fema~             8             4             4
## 5    11  15.7 Male             10            5             5
## 6    12  40.3 Fema~             6             4             3
## # ... with 3 more variables: family_income <chr>, average_hours_worked <chr>,
## #   self_rated_total_health <dbl>
```

Our data contains 8 variables: ID, age, sex, satisfaction_score, self_rated_health, self_rated_mental_health, family_income, and average_hours_worked.

Detailed descriptions of some variables:

* satisfaction_score indicates the life satisfaction score on a scale of 0(very dissatisfied) to 10(very satisfied).

* self_rated_health and self_rated_mental health are the physical and mental health ratings, respectively, on a scale of 1(poor) to 5(Excellent) given by the respondent.

* average_hours_worked indicates the average number of hours worked per week.

Since we want to observe how the life satisfaction score is related to potential factors such as health or financial conditions, the response variable of our analysis will be satisfaction_score and the predictors will be the potential factors: self_rated_health, self_rated_mental_health, family_income, and average_hours_worked.

Model

Introduce the selected model here. It is expected that you will use some mathematical notation here. If you do please ensure that all notation is explained. You may also want to discuss any special (hypothetical) cases of your model here, as well as any caveats.

¹As it is stated in the documentation of the GSS, one of the primary objectives of the General Social Survey is to monitor the well being of Canadians over time. As a result, every survey conducted so far contain the responses related to the questions asking for the respondents' well being, and the 2017 GSS is the most recent survey with such responses.

Now, we are going to fit a multiple linear regression model in order to find linear associations between `satisfaction_score` and other predictor variables.

Here is the summary of the multiple linear regression model:

```
##
## Call:
## lm(formula = satisfaction_score ~ as.factor(self_rated_health) +
##     as.factor(self_rated_mental_health) + family_income + average_hours_worked,
##     data = life_satisfaction_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.1029 -0.7224  0.0663  0.8971  6.1204
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.92502    0.29485  13.312 < 2e-16
## as.factor(self_rated_health)2    0.51471    0.10379   4.959 7.17e-07
## as.factor(self_rated_health)3    0.83475    0.09740   8.571 < 2e-16
## as.factor(self_rated_health)4    0.98584    0.09793  10.067 < 2e-16
## as.factor(self_rated_health)5    1.17522    0.09973  11.784 < 2e-16
## as.factor(self_rated_mental_health)2 1.63274    0.11629  14.040 < 2e-16
## as.factor(self_rated_mental_health)3 2.69951    0.10889  24.790 < 2e-16
## as.factor(self_rated_mental_health)4 3.24499    0.10932  29.684 < 2e-16
## as.factor(self_rated_mental_health)5 3.72145    0.11004  33.818 < 2e-16
## family_income$125,000 and more    0.04865    0.03822   1.273 0.203047
## family_income$25,000 to $49,999 -0.36600    0.04275 -8.561 < 2e-16
## family_income$50,000 to $74,999 -0.14543    0.04231 -3.437 0.000589
## family_income$75,000 to $99,999 -0.12047    0.04285 -2.811 0.004939
## family_incomeLess than $25,000 -0.47951    0.05232 -9.166 < 2e-16
## average_hours_worked0.1 to 29.9 hours 0.29608    0.27262   1.086 0.277474
## average_hours_worked30.0 to 40.0 hours 0.23252    0.27176   0.856 0.392231
## average_hours_worked40.1 to 50.0 hours 0.32056    0.27335   1.173 0.240935
## average_hours_worked50.1 hours and more 0.39476    0.27448   1.438 0.150400
##
## (Intercept) ***
## as.factor(self_rated_health)2 ***
## as.factor(self_rated_health)3 ***
## as.factor(self_rated_health)4 ***
## as.factor(self_rated_health)5 ***
## as.factor(self_rated_mental_health)2 ***
## as.factor(self_rated_mental_health)3 ***
## as.factor(self_rated_mental_health)4 ***
## as.factor(self_rated_mental_health)5 ***
## family_income$125,000 and more
## family_income$25,000 to $49,999 ***
## family_income$50,000 to $74,999 ***
## family_income$75,000 to $99,999 **
## family_incomeLess than $25,000 ***
## average_hours_worked0.1 to 29.9 hours
## average_hours_worked30.0 to 40.0 hours
## average_hours_worked40.1 to 50.0 hours
## average_hours_worked50.1 hours and more
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.298 on 12927 degrees of freedom
## (62 observations deleted due to missingness)
## Multiple R-squared:  0.2733, Adjusted R-squared:  0.2724
## F-statistic: 286 on 17 and 12927 DF, p-value: < 2.2e-16
```

As we can observe from the output, estimated `satisfaction_score` increases as each of `self_rated_health`, `self_rated_mental_health` and `family_income` increases; however, the slope estimates for `average_hours_work` are found to be quite inconsistent, because there is a decrease in slope estimates from 0.29608 to 0.23252 in the first two intervals of `average_hours_work`, but it increases again in the third and fourth intervals. Furthermore, p-values for `average_hours_work` that are much greater than the significance level of 0.05 provide us with more evidence that there are no linear relationship between the average working hours and the life satisfaction score.

Therefore, the regression model suggests that only the physical health condition, mental health condition and financial condition have positive linear relationships with the life satisfaction score.

Now that we've found out the response variable has linear relationships with some of the predictor variables, ask yourself: are the predictor variables equally important in terms of explaining the variability in `satisfaction_score`?

The answer is 'No'.

Although they all have positive linear relationships with the response variable, not all variables may contribute significantly in explaining the variability of the response variable.

Hence, we would now like to identify the most important predictor variable in this regression model.

There are two ways we can identify the most important predictor²

The one that we are going to use in this analysis is by comparing the changes in R^2 for the last variable added to the model. This is a valid method for identifying which predictor explains the most variability in the response variable, because when a newly added predictor variable is the only difference between the two models, the associated change in R^2 will represent the 'goodness-of-fit'.

To be more specific, we are going to begin with fitting a linear model with only one predictor, and then add one predictor to the model at a time to see how much R^2 changes when each variable is added.

Key point in this method is to identify the predictor variable with the largest increase in R^2 when it is the last variable added to the model.

This is the table with R^2 values in each model, and the change in R^2 as more variables are added:

```
## # A tibble: 3 x 2
##   R_squared Change
##   <dbl> <dbl>
## 1    0.117 0.117
## 2    0.240 0.122
## 3    0.254 0.0140
```

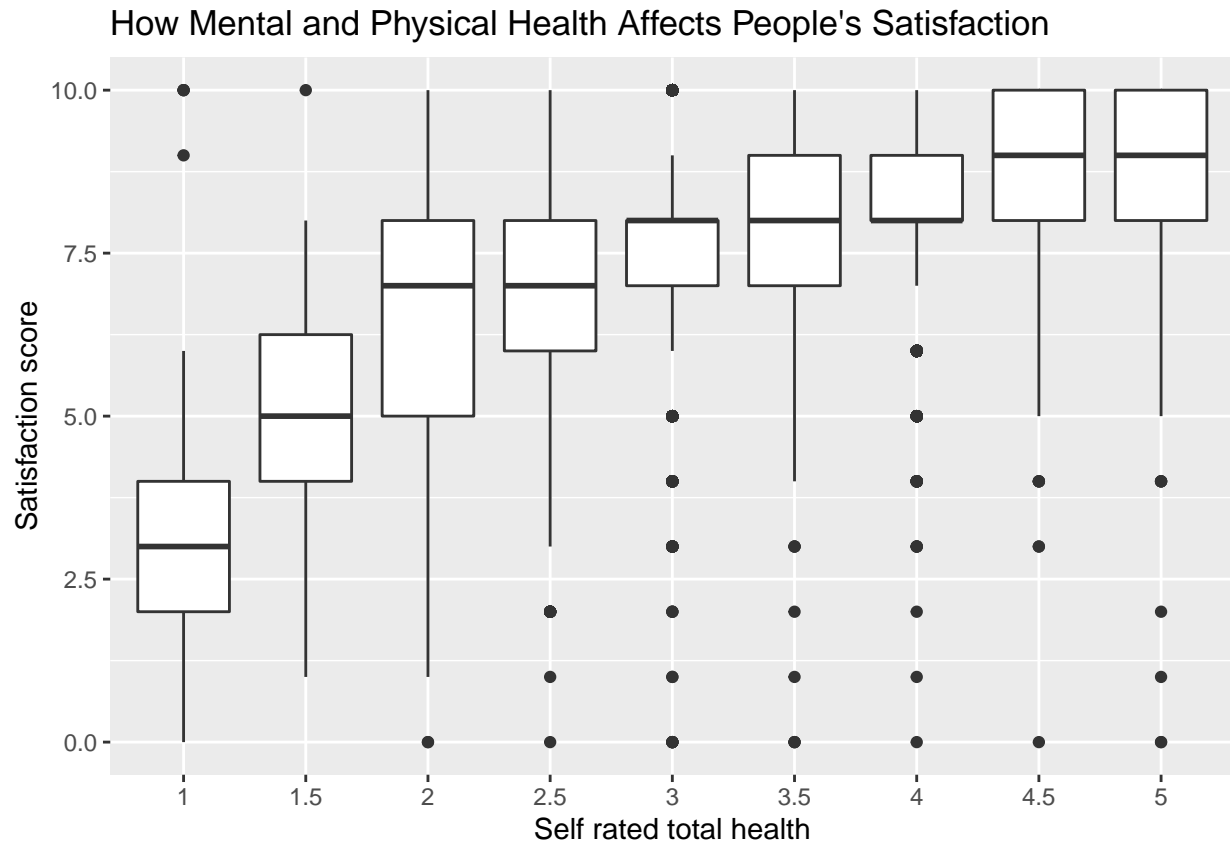
As it is shown in the table, there is a greatest increase in R^2 when the second predictor, `self_rated_mental_health` is added to the model. This increase is quite close the R^2 of our initial simple linear regression model with predictor `self_rated_health`.

Hence, the result suggests that both physical and mental health are significant predictors, however, mental health is the most important predictor in explaining the variability of the life satisfaction score.

²One way is to compare the standardized regression coefficients and the other way is to compare the increases in R-squared when each predictor is added to the regression model. For this analysis, specifically, we could not use the first method, because our model contains categorical predictor variable which cannot be standardized.

```
life_satisfaction_data %>%
  ggplot(aes(x = as.factor(self_rated_total_health), y = (satisfaction_score))) + geom_boxplot() +
  ggtitle("How Mental and Physical Health Affects People's Satisfaction") +
  ylab("Satisfaction score") +
  xlab("Self rated total health")
```

Warning: Removed 62 rows containing non-finite values (stat_boxplot).



Results

Here you will include all results. This includes descriptive statistics, graphs, figures, tables, and model results. Please ensure that everything is well formatted and in a report style. You must also provide an explanation of the results in this section. You can overflow to an Appendix if needed.

Please ensure that everything is well labelled. So if you have multiple histograms and plots, calling them Figure 1, 2, 3, etc. and referencing them as Figure 1, Figure 2, etc. in your report will be expected. The reader should not get lost in a sea of information. Make sure to have the results be clean, well formatted and digestible.

Discussion

Here you will discuss conclusions drawn from the results and comment on how it relates to the original goal of the study (which was specified in the Introduction).

Weaknesses

Here we discuss weaknesses of the study, data, analysis, etc. You can also discuss areas for improvement.

Next Steps

Here you discuss subsequent work to be done after this report. This can include next steps in terms of statistical analysis (perhaps there is a more efficient algorithm available, or perhaps there is a caveat in the data that would allow for some new technique). Future steps should also be specified in terms of the study setting (eg. including a follow-up survey on something, or a subsequent study that would complement the conclusions of your report).

References

1. GSS Data:
2. Data Cleaning Code: