Title of Your Report

Names of your Group Members

October 19th 2020

Abstract

This data set analyzes how mature Canadians feel about their life, provided certain factors in life. The original Canadian General Social Survey on family was conducted in 2017. Survey data includes a target population of individuals who are 15 years of age or older in the 10 provinces of Canada. The survey's primary objectives are to gather data on social trends to monitor changes in the living conditions of Canadians throughout life, and to provide information on specific social policy issues of current interest. 20602 respondents had responded to the survey and is the sample size. The sampling frame is Statistics Canada's common telephone frame, which combines landline and cellular telephone numbers that are from the Address Register. Survey data were collected through self-completed online questionnaires and telephone interviews. A linear regression model was applied in order to observe if a linear relationship exists between the study variable and the auxiliary variables.

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.

The goal of this analysis is to discover a linear relationship between the outcome variable and the auxiliary variables by utilizing a general approach of linear regression model. In this analysis, the study variable is feelings about life, and five auxiliary variables were selected to test if they indeed have a linear relationship with the outcome variable. The significance of the work is that through this model, an analysis can be built about how adult individuals in Canada feel about their life, considering the input variables. In the subsequent sections, results and discussion are covered to discuss weaknesses and next steps.

Data

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

The selected output variable is feelings about life, and the auxiliary variables are age of the respondent, age at first birth, place of birth (whether it was in Canada or not), income of the respondent, and intention of having future children. We chose this study variable in the intent to scrutinize the life satisfaction rating of Canadians, whether this is dependent on certain factors in life. Also, we decided on the specific input variables because intuitively, they are the most suitable data for our hypothesis and to prove that indeed, there is a linear relationship between the y and x variables. A drawback to the data is that because of the numerous 'NA' responses, our population data size greatly decreased from the original sampled population number. Another drawback to the data is that there are numerous vague responses such as "Don't know" or "Unsure", which are difficult to place in order.

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.

To predict the probability of a person being happy, which we define as:

$$\text{Prob}(h) := \begin{cases} 1, & \text{if feelings\_life} \geq 6 \\ 0, & \text{otherwise} \end{cases}$$

we fit a logistic regression model with some independent/predictor variables.

Table 1: Categorical Variables

Income of Respondent	Future Children Intention	Born in Canada?
Less than $25,000	Don't know	Don't know
$25,000 to $49,999	No, definitely not	Born outside Canada
$50,000 to $74,999	Probably not	Born in Canada
$75,000 to $99,999	Unsure	
$100,000 to $ 124,999	Probably yes	
$125,000 and more	Definitely yes	

The other predictor variables are `age_at_first_birth` and `age`.

Using the GSS data, we replicated the approach used in the original survey. A single-stage stratified sampling approach by applying finite population correction to the sample was employed, adjusting each observation by the corresponding provincial population to reduce the variation. Then, we fitted a logistic model to the survey design, yielding the following model:

$$\begin{aligned} \log\left(\frac{h}{1-h}\right) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &+ \underbrace{\beta_{3a}x_{3a} + \beta_{3b}x_{3b} + \beta_{3c}x_{3c} + \beta_{3d}x_{3d} + \beta_{3e}x_{3e}}_{\text{dummy coding for income}} \\ &+ \underbrace{\beta_{4a}x_{4a} + \beta_{4b}x_{4b}}_{\text{dummy coding for place birth Canada}} \\ &+ \underbrace{\beta_{5a}x_{5a} + \beta_{5b}x_{5b} + \beta_{5c}x_{5c} + \beta_{5d}x_{5d} + \beta_{5e}x_{5e}}_{\text{dummy coding for future children intention}} \end{aligned}$$

The functional form of the logistic model gives the logarithm of the odds of the outcome variable – in this case, the binary variable `happy`. For each of the categorical variables in Table 1, we use dummy variable coding with the variable at the top representing the baseline, in order to be able to assess what effect, if any, moving to a category would have compared to the baseline.

Results

The coefficients fitted using the logistic regression model are given in Table 2. The *OR* column gives the odds ratio, *Beta (SE)* gives the logarithm of the odds ratio and corresponding standard error of the estimate. *P* gives the p-value for the significance test which evaluates the probability of encountering data as or more significant than the Wald test statistic.

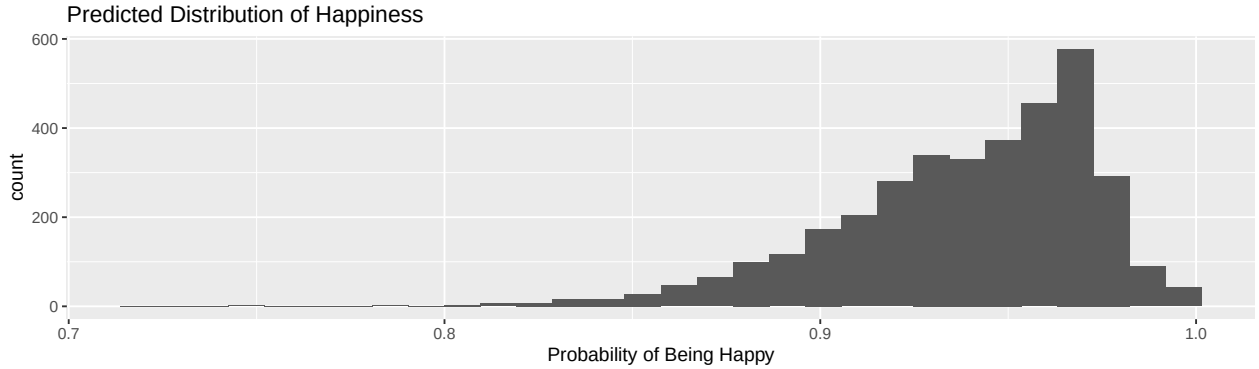
Table 2: Logistic Regression Model Summary

	Variable	OR	Beta (SE)	P
β_0	Intercept	-	30.70 (0.90)	<0.001
β_1	age_at_first_birth	1.05	0.05 (0.02)	0.002
β_2	age	0.96	-0.04 (0.01)	<0.001
income cat. var	as.factor(income_respondent)			
income baseline	Less than $25,000 (ref)	-	-	-
β_{3a}	$25,000 to $49,999	1.34	0.29 (0.19)	0.13
β_{3b}	$50,000 to $74,999	2.18	0.78 (0.23)	<0.001
β_{3c}	$75,000 to $99,999	3.03	1.11 (0.29)	<0.001
β_{3d}	$100,000 to $ 124,999	2.58	0.95 (0.39)	0.02
β_{3e}	$125,000 and more	4.96	1.60 (0.49)	0.001
birthplace cat. var	as.factor(place_birth_canada)			
birthplace baseline	Don't know (ref)	-	-	-
β_{4a}	Born outside Canada	0.00	-14.21 (0.46)	<0.001
β_{4b}	Born in Canada	0.00	-14.12 (0.44)	<0.001
children int. cat. var	as.factor(future_children_intention)			
children int. baseline	Don't know (ref)	-	-	-
β_{5a}	No, definitely not	0.00	-14.02 (0.52)	<0.001
β_{5b}	Probably not	0.00	-14.20 (0.55)	<0.001
β_{5c}	Unsure	0.00	-14.48 (0.58)	<0.001
β_{5d}	Probably yes	0.00	-14.10 (0.66)	<0.001
β_{5e}	Definitely yes	0.96	-0.04 (0.56)	0.94

The model has the following interpretations:

- The intercept term has no meaningful interpretation and it is statistically significant with a very small p-value.
- The continuous variables `age_at_first_birth`, `age` very slightly affect the model with odds-ratios slightly off from 1.
- The categorical variables `place_birth_canada` and `future_children_intention` make a large difference to the odds ratio with odds-ratios near 0, but this is probably due to a flaw in the survey design which we will discuss in the Weaknesses section of this report.
- `income_respondent` is a meaningful variable in this model. For all six levels, the odds ratio is greater than 1 which implies that each increase in the respondent's income results in an increase in the log odds of being happy compared to the baseline.

Using a variety of parameters with this model, we obtain the following histogram that shows the class distribution of happiness.



Discussion

Weaknesses

Next Steps

References

Bibliography:

1. General Social Survey: An Overview, 2019. (2019, February 20). Retrieved October 12, 2020, from Statistics Canada, Canada website: https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm
2. General Social Survey - Family (GSS). (2019, February 7). Retrieved October 12, 2020, from Statistics Canada, Canada website: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501
3. Wu, Changbao, and Mary E. Thompson. "Basic Concepts in Survey Sampling." Sampling Theory and Practice. Springer, Cham, 2020. 3-160.

sorry I'll cite this properly later!

4. Used for finite population correction across strata: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=101&S=50&O=A