Predicting probability of life satisfaction level among adult Canadians with multiple factors in life

Steven Tran, Nayoung Kim

October 19th 2020

Code and data supporting this analysis is available at: https://github.com/st-tran/STA304-Problem-Set-2.git

Abstract

This report analyzes how mature Canadians feel about their life, provided certain factors in life. Deriving data from the 2017 General Social Survey, this analysis utilizes a dependent variable that has two categorical levels- whether the respondent individual answered feeling generally happy, or the respondent was overall not content with life. Income of the respondent illustrated to play a role in influencing the life rating, whereas other factors in life did not portray a significance in affecting the rating. This was a meaningful discovery as higher income was found to be associated with higher life happiness level in general.

Introduction

The goal of this analysis is to discover a relationship between the outcome variable of contentedness of life and the auxiliary variables of factors in life by utilizing a logistic 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 relationship with the outcome variable. Feelings about life were measured as generally happy if the rating was six or higher out of ten, while they were grouped as overall not content if the rating was lower than six. The chosen independent variables are age at first birth, current age of the respondent, income of the respondent, place of birth whether it is within or outside of Canada, and future intention of having any children. While there were numerous variables in the data that were available to select, these five specific variables were intuitively picked to check the hypothesis that these will have an association with life rating. The importance of the work is that through the logistic regression model, an analysis can be built about how adult individuals in Canada feel about their life, considering the input variables. More precisely, the probability of an adult Canadian having a contented life can be calculated. Ultimately, the main goal is to observe if any of these life factors affect life rating, and if so, how large the influence is. In the subsequent sections, the model section is covered and explained using tables and the results section through a graph, while weaknesses and next steps are also discussed.

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 relationship between the y and x variables. All of the data were pulled from the General Social Survey - Family (GSS) that was conducted in 2017. The 2017 survey data set was selected because it was a relatively recent set of data, therefore more interesting and purposeful to observe as it reflects comparatively newer and current responses. Survey data includes a target population of non-institutionalized 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.

20,602 respondents had responded to the survey, and the number of respondents is also the sample size. This survey is a cross-sectional design and uses a two-stage sampling design. During the first stage, the sampling groups are the groups of telephone numbers, and the second stage units are people in the identified households. 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, handled both in English and French. Since responding to the survey was voluntary, there were non-responses, and survey respondents were allowed to have non-responses within their survey answers as well. Also, although there were households without telephone services- generally, these exclusions are small, and so introduced biases were expected to be minor too.

In the data set, there were numerous variables and responses, and some variables were very similar to one another. For instance, there were two questions regarding income, one for income of the respondent and another for income of the entire family. Since income of the respondent would usually be more direct and influential to the respondent, this variable was chosen instead of the other. There were no new variables that were constructed by combining multiple others, as each variable by itself normally had several response options, and it would be confusing to mix up the variables. The key features of the survey are that it covers various important life events and factors that influence life, and so the primary objectives of the survey are well covered and represented. Its strengths are that therefore, the survey is fairly accurate in terms of representing all the variables that affect life, and the large number of respondents and the appropriate survey designs also add to the accuracy of the survey.

A drawback to the data is that because of the numerous questions and responses in the survey and gathered data, it was time consuming to review all these to choose our variables. 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 which should be placed before the other. Furthermore, some weaknesses about the survey in general are that the questions could have been phrased more inclusively, and that the answer options could have been broader. For example, the variable "age at first birth" does not include males, and all males responded with "NA" to the question. However, males are broadly still affected by first child's birth, as they participate in nurturing the child. A more precise question would be with the phrase "age at first child's birth," which would permit males to answer. Moreover, the variable "sex" only represents female or male- such option for "intersex" is not included. Perhaps another category could have been added for "gender," in which respondents could have more comfortably represented themselves within the LGBTQ+ group.

Figures 1-6 show various visualizations of the raw data. We see that most respondents have a positive feeling about life with a bimodal (at 8 and 10), left-skewed distribution of `feelings_life`. An interesting result seen in Figure 2 is that most respondents are older than 50 and no respondents were younger than 20, despite the target population including persons 15 years of age or older. Moreover, most respondents make less than the national mean of around $50,000 per year as seen in Figure 4. Most respondents were born in Canada and are not confident about whether they will have children in the future.

Figures 7-9 show the aggregated means for all respondents across the variables `feelings_life`, `age`, and `age_at_first_birth` by province. We see from the limited range of the variables for each plot that the provinces are not too different from one another for these factors.

Figure 10 is a 3D scatterplot of `age` and `age_at_first_birth` against `feelings_life`, where the points are colour-coded corresponding to the value of `feelings_life`. We see that most of the responses are clustered towards a high value for `feelings_life`, low value for `age_at_first_birth`, and medium-high value for `age`.

Figure 1: Distribution of feelings_life

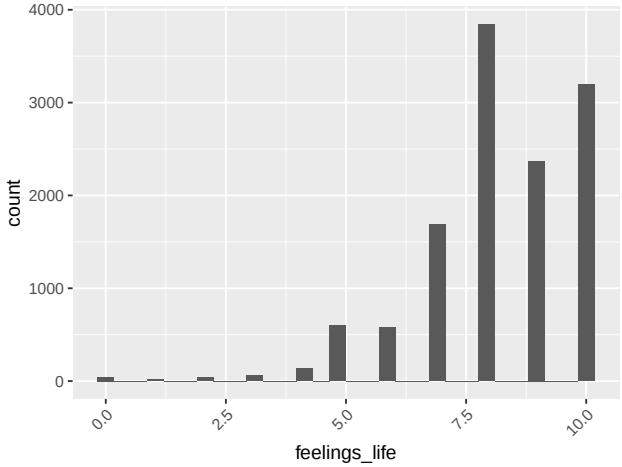


Figure 2: Distribution of age

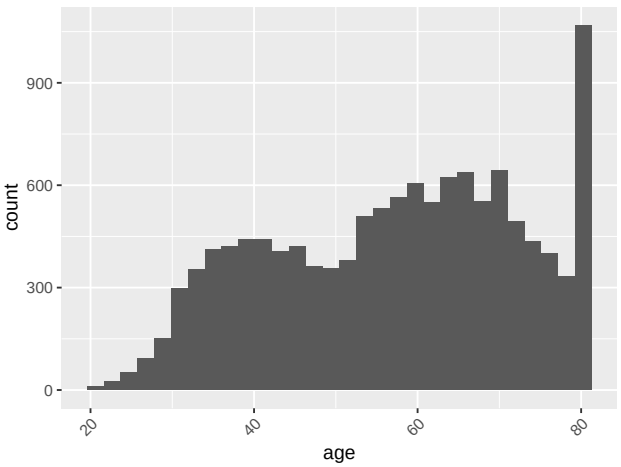


Figure 3: Distribution of age_at_first_birth

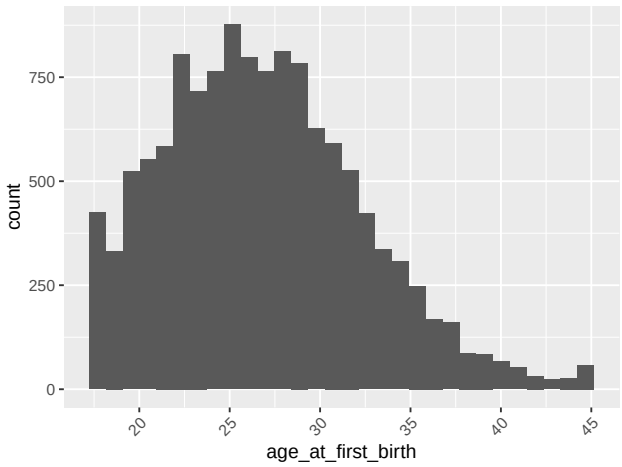


Figure 4: Distribution of income_respondent

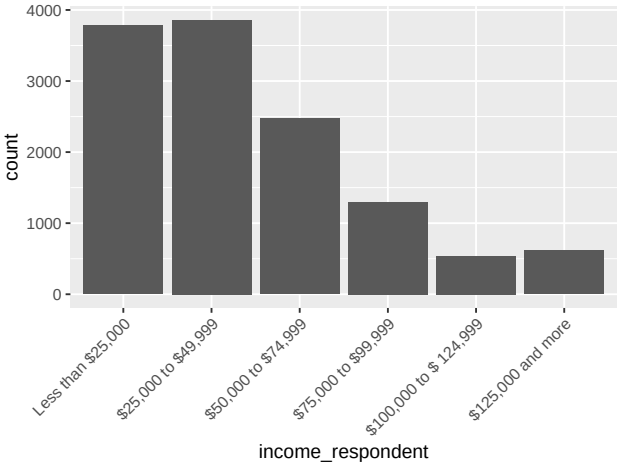


Figure 5: Distribution of place_birth_canada

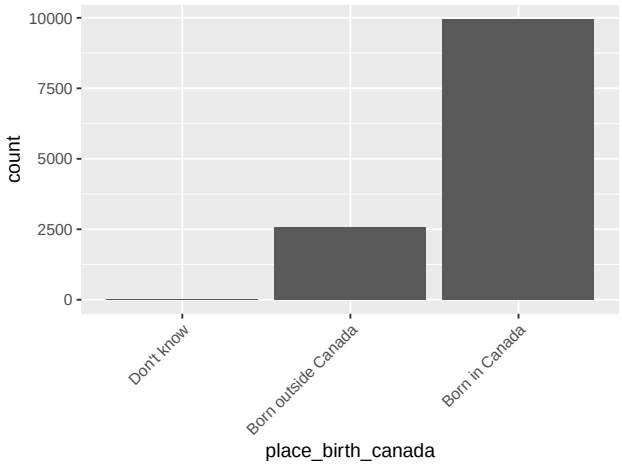


Figure 6: Distribution of future_children_intention

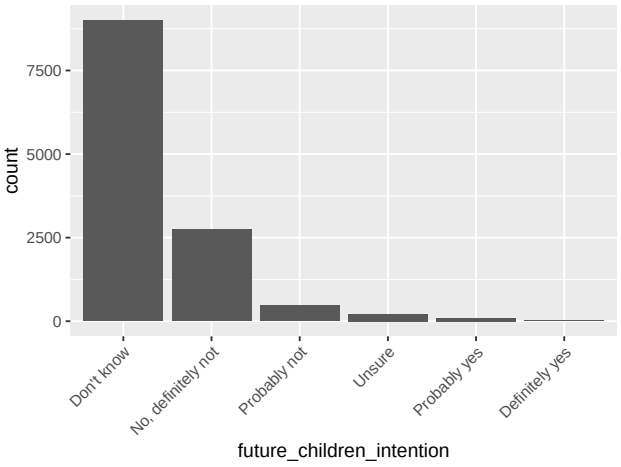


Figure 7: Mean feelings_life by province

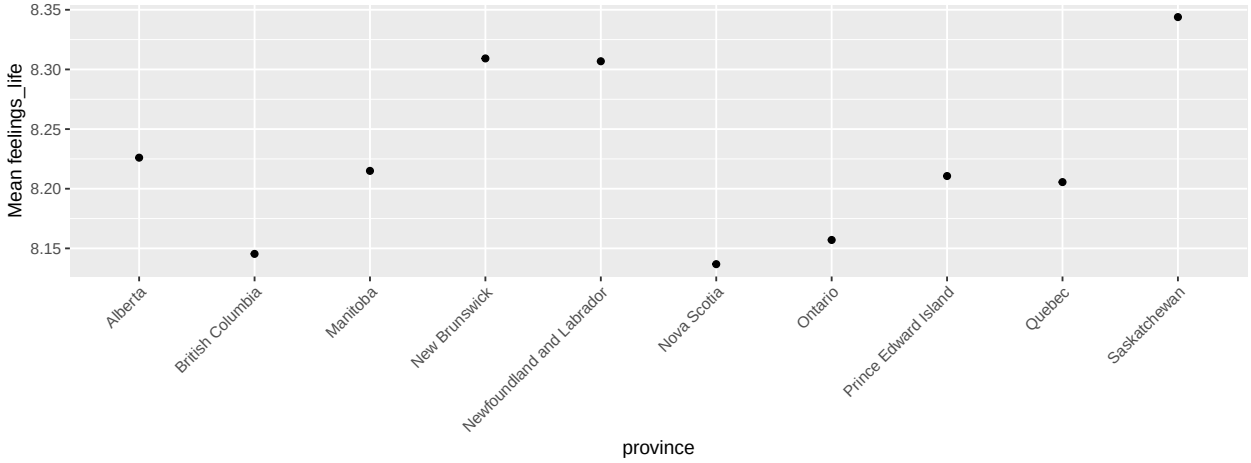


Figure 8: Mean age by province

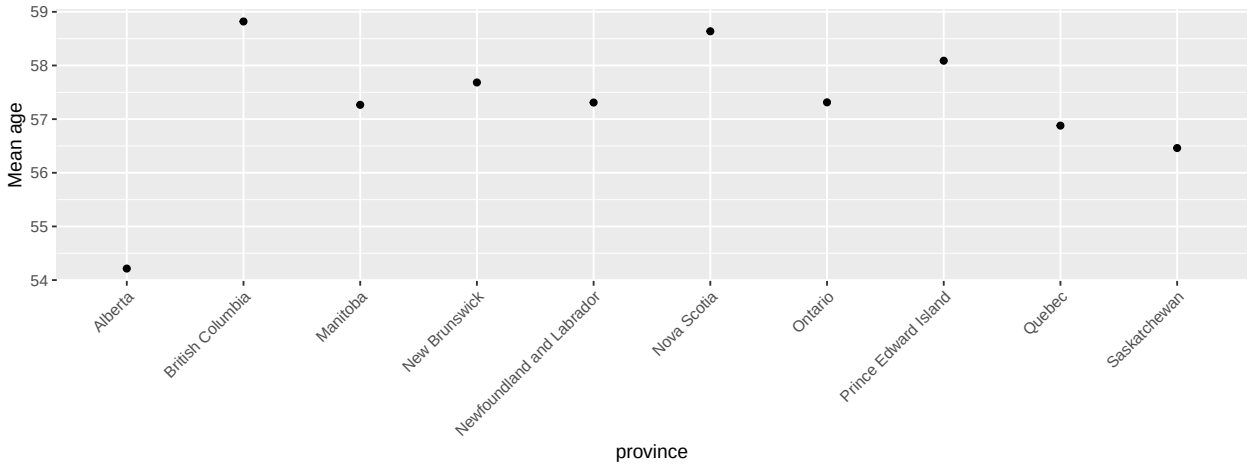


Figure 9: Mean age_at_first_birth by province

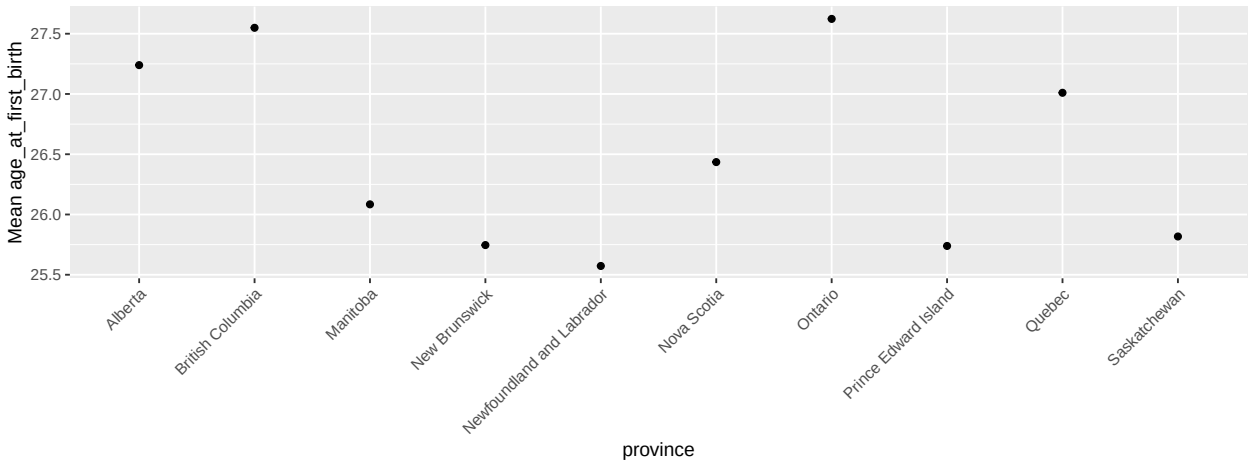
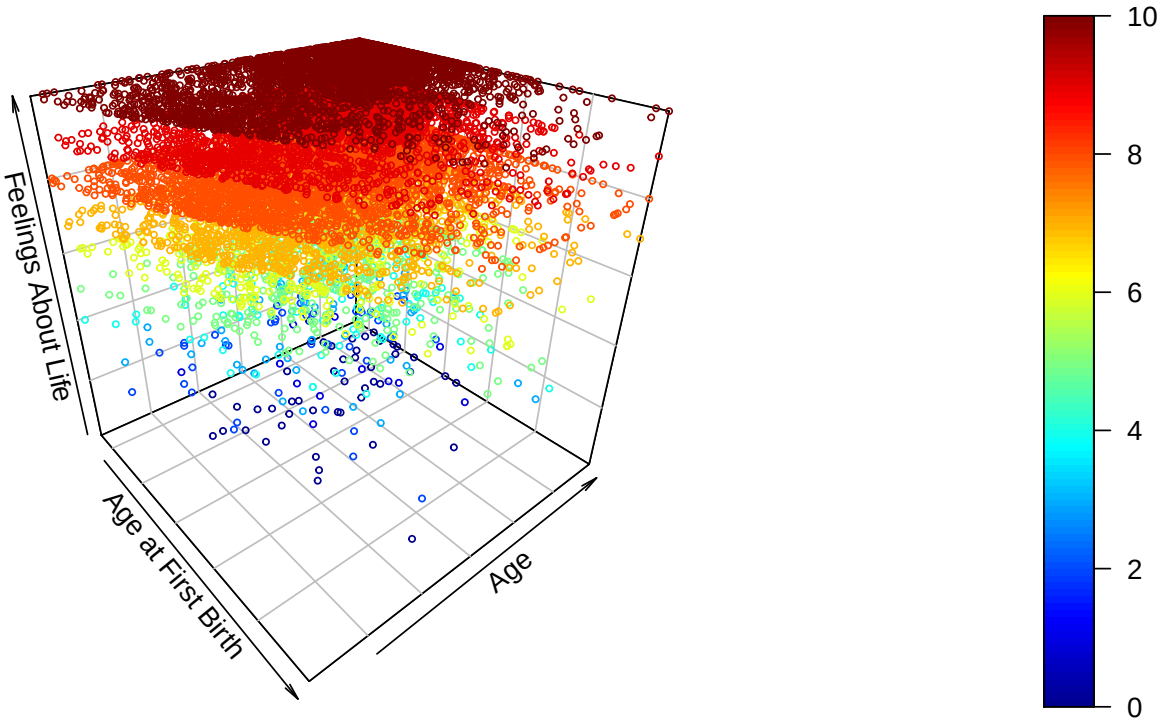


Figure 10: Age at First Birth & Age vs Feelings About Life



Model

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. This yielded 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` (h). 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.

Hypothetical cases of the model are that if income is higher, individuals would be able to enjoy a more comfortable life not having to worry about their expenses, and therefore life rating would be higher. While there are age at first birth, age, income of the respondent, place of birth, and future children intention for independent variables and feelings about life as the dependent variable- the independent variables enter the model specifically because for instance, income of the respondent would more accurately portray the financial circumstance of him or her than income of the family. Also, responses to income of the respondent was categorized in an increasing manner for income. Caveats in the model are that male responses are not captured for the variable "age at first birth," and that the model assumes males to be the baseline of the variable. The logistic regression equation is determined by the addition of each independent variable calculated to find probability of having a happy life. Since the outcome of interest is binary- either the individual has a happy life or does not have a happy life, the model is suitable in the sense that the goal is to observe who is living a satisfied level and who is not, and if more people are living a contented life or not.

Because the log function of the logistic regression model is one-to-one, if the regression equation is of a high value, then it would mean the numbers for the variables are also of high values. The justification for applying this model is that the general goal of the study is to find if an average adult Canadian is having a happy or unhappy life, thus the logistic regression would work because of its binary outcome model. Although a linear regression model could have also identified if there is a relationship between a given independent variable and the dependent variable, this model would be limited in that it would only allow one independent variable to be considered and only illustrate a linear relationship. Working with an ordinal logistic regression model may allow more levels of life ratings for the independent variable as well, but this may be more complicated to use for dealing with a lot of variables. The chosen independent variables are a suitable fit for this study as some are categorical and some are numerical, showing both sides. If the variable responses were all categorical or numerical, the study might not have been able to cover a greater picture of the details in the data set.

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 under the assumption that the population parameter coefficient of interest is equal to 0. The Wald test is relevant because of the log-transformed odds ratio; the model summary summarizes the results of this test to each of the parameters.

Table 2: Logistic Regression Model Summary

	Variable	OR	Beta (SE)	P
β_0	Intercept	-	2.12 (0.88)	0.02
β_1	age_at_first_birth	1.02	0.02 (0.01)	0.01
β_2	age	1.00	-0.00 (0.00)	0.35
income cat. var	as.factor(income_respondent)			
income baseline	Less than $25,000 (ref)	-	-	-
β_{3a}	$25,000 to $49,999	1.62	0.48 (0.09)	<0.001
β_{3b}	$50,000 to $74,999	2.37	0.86 (0.12)	<0.001
β_{3c}	$75,000 to $99,999	2.82	1.04 (0.17)	<0.001
β_{3d}	$100,000 to $ 124,999	4.39	1.48 (0.31)	<0.001
β_{3e}	$125,000 and more	7.65	2.03 (0.35)	<0.001
birthplace cat. var	as.factor(place_birth_canada)			
birthplace baseline	Don't know (ref)	-	-	-
β_{4a}	Born outside Canada	0.61	-0.50 (0.82)	0.55
β_{4b}	Born in Canada	0.69	-0.37 (0.82)	0.65
children int. cat. var	as.factor(future_children_intention)			
children int. baseline	Don't know (ref)	-	-	-
β_{5a}	No, definitely not	0.95	-0.05 (0.13)	0.67
β_{5b}	Probably not	0.95	-0.05 (0.24)	0.84
β_{5c}	Unsure	0.75	-0.28 (0.30)	0.34
β_{5d}	Probably yes	1.20	0.18 (0.45)	0.69
β_{5e}	Definitely yes	188128.04	12.14 (0.25)	<0.001

The model has the following interpretations:

- The intercept term has no meaningful interpretation but it is statistically significant with a very small p-value. It shifts the overall model prediction by e^{β_0}.
- The continuous variables `age_at_first_birth` and `age` don't affect the model by much with odds-ratios slightly off from 1. β_1 is statistically significant but β_2 is not, with a p-value of 0.35, and so we fail to reject the null hypothesis that $\beta_2 \neq 0$.
- We see a very large prediction of the odds ratio for β_{5e} (`future_children_intention`=*Definitely yes*) which is likely caused by the large number of people who voted *Don't know*. The other parameters β_{5i} for other levels of `future children intention` are not statistically significant, which is also due to the same reason. However, we included this factor in the model because it is intuitive that living a happy life would be necessary to raising and taking care of children. Despite this, there are certainly outliers in reality in low-income households, wherein there are many children in the same household.
- `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.

Discussion

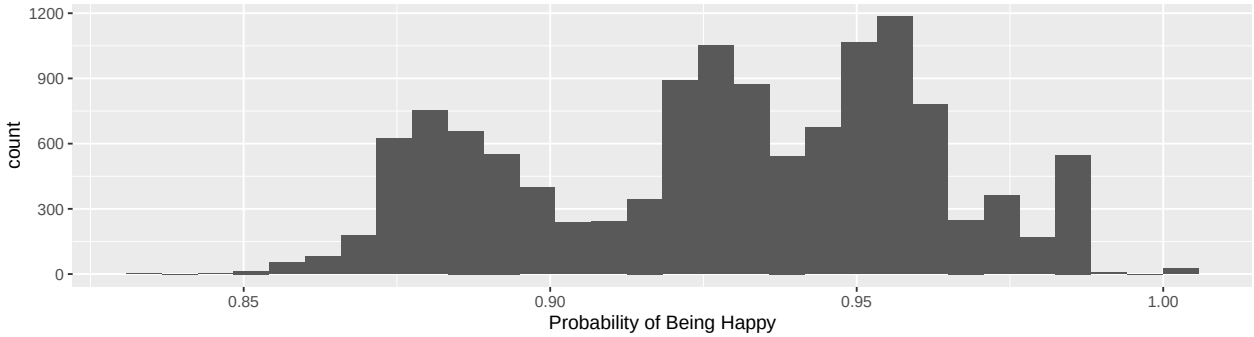
The assumptions used to fit the logistic regression model generally do not reflect our assumptions on real-world relationships between happiness and factors that affect one's standard of living. While some of the factors we set out to investigate, namely: age, intentions of having children in the future, and whether or not they were born in Canada, were all found to be statistically insignificant, it is possible that there are confounding factors that may reveal a relationship. These results did not fit with our initial hypothesis that they would all contribute to our measure of Canadians' happiness.

It was seen that most respondents responded that they are very content with their lives, with a mean rating of 9.2043156 which doesn't fall in line with reality. As we will discuss in the Weaknesses section of the report,

there is likely response bias that affected the General Social Survey. Due to this, our model was biased towards predicting the upper end of the range.

This can be seen in Figure 11. Using a variety of parameters with this model in which each combination has a uniform probability of being chosen, we obtain the following histogram that shows the class distribution of happiness across Canada. Note that the bins are lower-bound inclusive, explaining a number of predictions appearing beyond 1 (these probabilities are 1).

Figure 11: Predicted Distribution of Happiness



Indeed, there are further studies to be done for comparison to our results. Differences in the survey design such as a larger sample size, different sampling method, or change in the target population could affect the data that was analyzed. As shown in the Data section, most respondents were more than 50 years old. The other factors were also skewed, but perhaps a Bayesian model could be fit to the data with assumptions on the prior distributions of these variables to expose a more meaningful relationship. Alternatively, further post-stratification may be employed to make the sample more representative of the Canadian population, adding more factors that were available in the GSS data set.

Weaknesses

There were some weaknesses found in the data set and the study as well. Recalling one of the originally discussed weaknesses, the variable "age at first birth" does not capture valid responses for males who would have otherwise been able to answer a question with a phrase such as "age at first child's birth." Furthermore, the data may not represent all of adult Canadians correctly as those in the three territories were excluded. Also, because the survey responses were assembled voluntarily, again the sample from the target population of mature Canadians might not have been detained properly- it is possible that only those who are more eager to answer questions or those with more available time might have responded to the survey. And because the survey was provided in the form of an online questionnaire or a telephone interview, those who have low vision or are hard of hearing might not have been able to answer the survey even if they wanted to, and these conditions could generally apply more to seniors. These weaknesses could have made the results more biased for the analysis.

For instance, the data mostly had high ratings for feelings about life- this could have happened because if people are not satisfied with their life, they may not wish to pick up the phone and interact with another individual on the other line or not involve themselves in extra tasks such as fulfilling a non-mandatory survey. Similarly, even if they actually rate their life on the lower scale, they may falsely provide a higher rating, thinking that a higher life rating would be a desired response. Then the final data set would contain social desirability bias for response bias. On a similar note, due to some respondents possibly wishing to appear to be in a more favourable financial situation or to have a more advanced level of education, they might have given inflated responses, producing prestige bias. Areas for improvement would be to create a more inclusive survey overall- the population would be more representative if the three territories of Canada along with the ten provinces were included. Also, as noted previously, the questions and responses could be phrased in a more inclusive form or have wider options to avoid 'NA' responses.

Next Steps

After this report, it may be interesting to analyze the same study variable of life rating utilizing similar independent variables. For example, the variables "income of the respondent" and "grandparents still living" are two differently categorized variables. However, "income of the respondent" and "income of the family" are alike in that they both are "income." By replacing the independent variables used in the analysis with these closely related ones, the analysis can be repeated to illustrate some results. Alternatively, a very specific study can be completed by simulating a linear regression model incorporating just one independent variable and a dependent variable, such as if age at first marriage has a positive or negative linear relationship with life rating. A caveat in the data was that a lot of these independent variables to life rating had "NA" responses, so in a future study, it may be worthwhile to choose variables with the smallest number of "NA" responses to avoid nonresponse bias.

In a potential future follow-up survey, it may be useful to revisit the male respondents and ask the "age at first birth" question again, rephrasing it to "age at first child's birth." Then once those who have experienced first child's birth provide their answers, this variable's responses can be updated. Another interesting subsequent study may be to conduct the same study with the same respondents a few years later. This will fix some variable responses as they do not change, such as place of birth. However, age would have changed definitely, and responses to other questions such as income or future children intention may have changed. It would be notable to observe if there are any significant changes to life rating because of these changes in question responses. Finally, another more advanced model that can be applied is the ordinal logistic regression model. As it is an extension to the binary logistic regression model, it would be able to produce not just the "happy" and "unhappy" life rating outcomes, but each actual level of feelings about life.

References

Bibliography:

1. Children living in low-income households. (2017, September 13). Retrieved October 19, 2020, from Statistics Canada, Canada website: https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016012/98-200-x2016012-eng.cfm
2. 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
3. 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
4. Grimm, P.(2010). Social Desirability Bias. Wiley Online Library. https://doi.org/10.1002/9781444316568.wiem02057
5. Income of individuals by age group, sex and income source, Canada, provinces and selected census metropolitan areas. (2020, October 19). Retrieved October 19, 2020, from Statistics Canada, Canada website: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1110023901
6. Population and Dwelling Count Highlight Tables, 2016 Census. (2019, February 20). Retrieved October 17, 2020, from Statistics Canada, Canada website: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=101&S=50&O=A
7. Sheather, S. J. (2009). A modern approach to regression with R. Springer. https://books-scholarsportal-info.myaccess.library.utoronto.ca/en/read?id=/ebooks/ebooks0/springer/2010-02-11/1/9780387096087#page=1
8. Wiley, J. F., Pace, L. A. (2015). Beginning R: An Introduction to Statistical Programming. Apress. https://books-scholarsportal-info.myaccess.library.utoronto.ca/en/read?id=/ebooks/ebooks3/springer/2017-08-17/1/9781484203736#page=1

9. Wu, Changbao, and Mary E. Thompson. “Basic Concepts in Survey Sampling.” Sampling Theory and Practice. Springer, Cham, 2020. 3-160.