Predicting Self-Rated Mental Health Based on Demographic and Family Traits

James Bao, Alan Chen, Xinyi Zhang, Zidong Yang

10/19/2020

The code used to generate this RMarkdown file can be found at https://github.com/cindyzhang99/sta304_ps3/blob/main/sta304_ps3.Rmd.

Abstract

- An abstract is included and appropriately pitched to a general audience.
- The abstract answers: what was done, what was found, and why this matters (all at a high level).
- If your abstract is longer than four sentences then you need to think a lot about whether it is too long. It may be fine (there are always exceptions) but you should probably have a good reason.

Introduction

- The introduction is self-contained and tells a reader everything they need to know, including putting it into a broader context.
- Your introduction should provide a bit of broader context to motivate the reader, as well as providing a bit more detail about what you're interested in, what you did, what you found, why it's important, etc.
- A reader should be able to read only your introduction and have a good idea about the research that you carried out.
- It would be rare that you would have tables or figures in your introduction (again there are always exceptions but think deeply about whether yours is one).
- It must outline the structure of the report.

Data

The dataset we used in our modeling is the 2017 General Social Survey (Family cycle). The following sections will discuss how the data was collected, what the key features of the dataset are, and what the data looks like.

Data Collection

From February 1, 2017 to November 30, 2017, Statistics Canada gathered data on the Canadian family unit by conducting voluntary telephone interviews. Their target population was all non-institutionalized individuals living in Canada, aged 15 or older. Cross-sectional sampling was conducted in a two-stage design.

The stratified simple random sampling method was used in the first stage. Here, the sampling frame consisted of telephone numbers from the Census grouped as households using data from Statistic Canada's dwelling frame. Strata were formed at the census metropolitan area (CMA) level and at the province level (i.e., large CMAs formed their own stratum, smaller CMAs were grouped together, and the non-CMA regions of each province were grouped together), forming a total of 27 non-overlapping strata. Finally, households were sampled randomly from each stratum such that the number sampled units from each stratum corresponded to the population sizes of each stratum. To reiterate, the sampled population for this first stage was the chosen households from each stratum. The stratified simple random sampling method was also used in the second stage. Here, the sampling frame was a list of household members, aged 15 and older, from the households selected in the first stage. Then, one individual was randomly selected from each household, forming the sampled population. Approximately 43,000 individuals were contacted to participate in the survey.

Overall, the surveying method using two-stage simple random stratified sampling is effective in generating a sample that is geographically representative of the Canadian population. In addition to estimates about the Canadian population at large, the stratified sampling method also allows estimates to be made about subpopulations (at the province level).

Statistics Canada reported that the non-response rate was 47.6%. To reduce the effects of non-response bias, survey responses were adjusted based on the demographic characteristics of households that were non-responsive (by pulling their information from the 2016 Census). This ensures that the discrepancy between the target population and survey responses resulting from non-response is minimized. Furthermore, for the Family cycle of the GSS, responses were also adjusted for income and household size to make more accurate survey estimates for the variables of interest.

Statistics Canada did not disclose the true cost of conducting the survey but we can make some speculations based on the available information about their field work methodology. Surveying was conducted using Computer Assisted Telephone Interviewing (CATI) wherein interviewers read aloud the computerized questionnaire and immediately record the respondent's answers. Although this allows for a reduction in costs compared to traditional in-person surveying, labor costs still include time spent computerizing the survey, training interviewers, and having interviewers administer the questionnaire. Other labor costs include designing the questionnaire and surveying methodology as well as conducting quality control (data consistency was checked by the CATI system during surveying and unresolved inconsistencies were handled afterwards by support staff). Non-labor costs likely included paying for equipment, phone service, offices, and so forth. Again, although we don't have exact costs, we can conclude that the time and costs associated with conducting the GSS is a clear reason why it is only administered once a year.

Per the report on the 2017 GSS from Statistics Canada, extensive research and testing was conducted when designing the questionnaire. Consequently, a major strength of the questionnaire is that it contains focused questions that comprehensively and extensively capture the subject of interest (the Canadian family). Upon reading through the questionnaire made available by Statistics Canada, the wording of each question is precise and clear, leaving little room for ambiguity. Additionally, another strength of the survey is that a vast majority of questions were objective (dates, events, counts) removing potential response biases that occur with subjective questions. (Not all questions were objective however, in fact the variable of interest we will model in subsequent sections consists of subjective responses.) On the other hand, because of the specificity of the questions, the survey is very long with several dozens sections and several questions per section. Furthermore, as a result of the large scope of the target population, many questions in the survey did not apply to a large majority of respondents (e.g., number of grandchildren, questions about additional marriages, etc.). The data collected is also incomplete because participants were given the option to refuse to answer or answer "I don't know" to each question since participation was voluntary.

Overall, the surveying methodology and distributed questionnaire were carefully designed in the interest of collecting accurate, representative data wherever possible.

Data Characteristics

The full dataset of responses to the 2017 General Social Survey (Family cycle) contains 20,602 observations for over 400 variables relating to the Canadian family. A large reason for our choice to use this dataset is because it is the most recent GSS cycle available for modeling. Other benefits of this dataset have been previously touched upon in the previous section. Namely, the data was checked for consistency in real time by the CATI system (as well as by survey support staff) so there is a certain measure of accuracy that other survey results lack. Additionally, the stratified simple random sampling method used to distribute the survey suggests that the results are representative of the Canadian population to some degree (in the geographical sense at the very least). A major weakness of the data is that it is not complete because of the voluntary nature of the surveying.

In the interest of space, we will only discuss the variables in the dataset that are relevant to our model. The variable we aim to predict is self_rated_mental_health while the factors that we chose to inform this prediction are age, sex, marital_status, and self_rated_health. We chose these factors based on the demographic information mentioned in mental health statistics (age, sex, and health) and based on what we suspected might contribute to mental health in the context of family composition (marital status, has children). More explicitly, here are the chosen variables we used from the original dataset:

- agedc (renamed to age): the exact age of the respondent (in decimals) at the time of the survey
- sex (sex): sex of the respondent, the options being "Male" or "Female"
- marstat (marital_status): marital status of the respondent, the options being "Single, never married", "Married", "Living common-law", "Separated" (but still legally married), "Divorced", or "Widowed"
- totchdc (total_children): total number of children reported by respondent
- srh_110 (self_rated_health): self-rated physical health, the options being "Excellent", "Very good", "Good", "Fair", and "Poor"
- srh_115 (self_rated_mental_health): self-rated mental health, the options being "Excellent", "Very good", "Good", "Fair", and "Poor"

For age, there is a similar variable in the original dataset that uses only natural numbers (agec), however, we chose to use agedc (and renamed it to "age") in the interest of accuracy. There are many variables related to marriage in the original dataset (totunc: total number of marriage and common-law unions, nmarevrc: number of marriages the respondent has had, etc.) but they don't capture the same scope of information as marstat (renamed to "marital_status") does (for example, being divorced or widowed is not reflected in those variables). Consequently, we chose marital_status as opposed to the other available variables related to marriage. We transformed totchdc (renamed to "total_children") into a binary variable has_children (if total children is 0, has_children equals FALSE, if total children is greater than 0, has_children equals TRUE). For the other three variables we use (sex, self_rated_health or srh_110 in the original dataset, and self_rated_mental_health or srh_115 in the original dataset), there are no similar equivalents.

We will refer to the variables by their renamed identifiers for the rest of the paper.

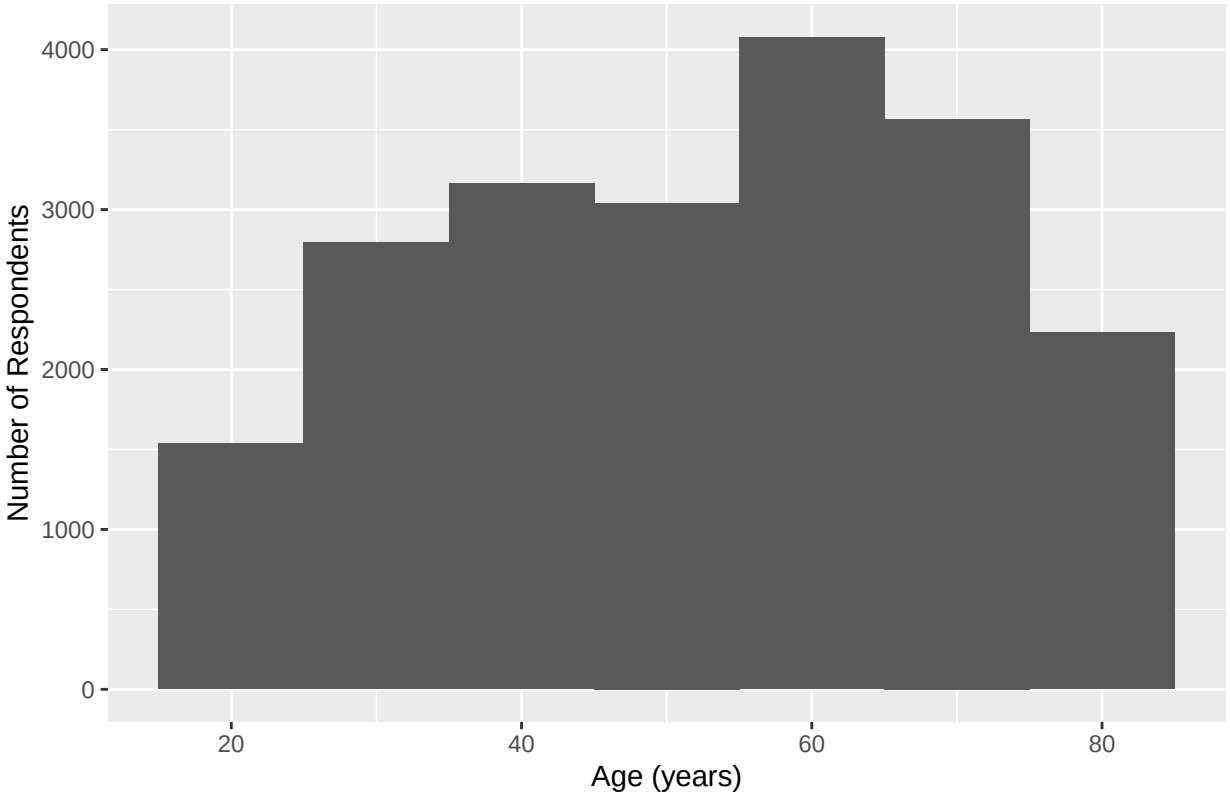
Data Visualization

Specific instructions on how to download the 2017 GSS dataset can be found in the header of the gss_cleaning.R file found here (https://github.com/cindyzhang99/sta304_ps3/blob/main/gss_cleaning/gss_cleaning.R).

We cleaned the original 2017 GSS dataset using gss_cleaning.R, producing gss_cleaned.csv (with variables renamed accordingly).

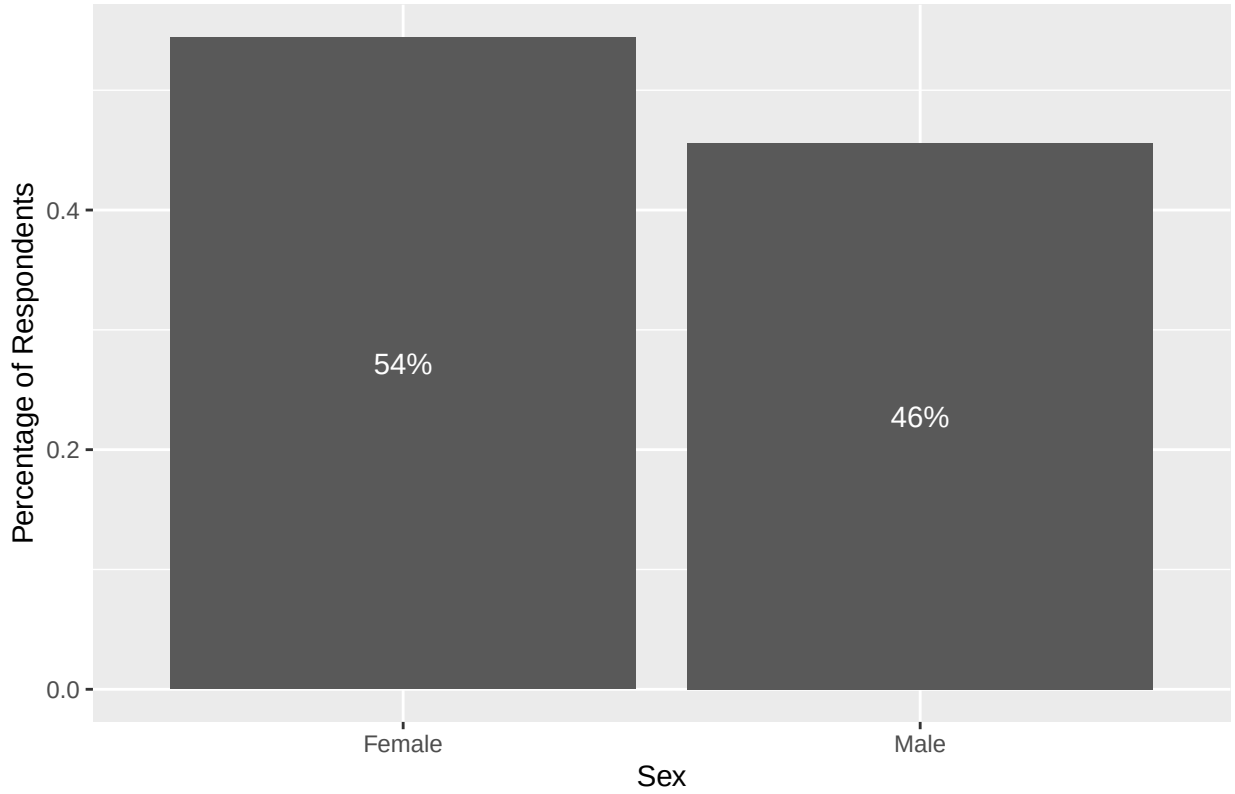
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   37.30   54.20   52.16   66.70   80.00
```

Figure 1: Distribution of the age of respondents.



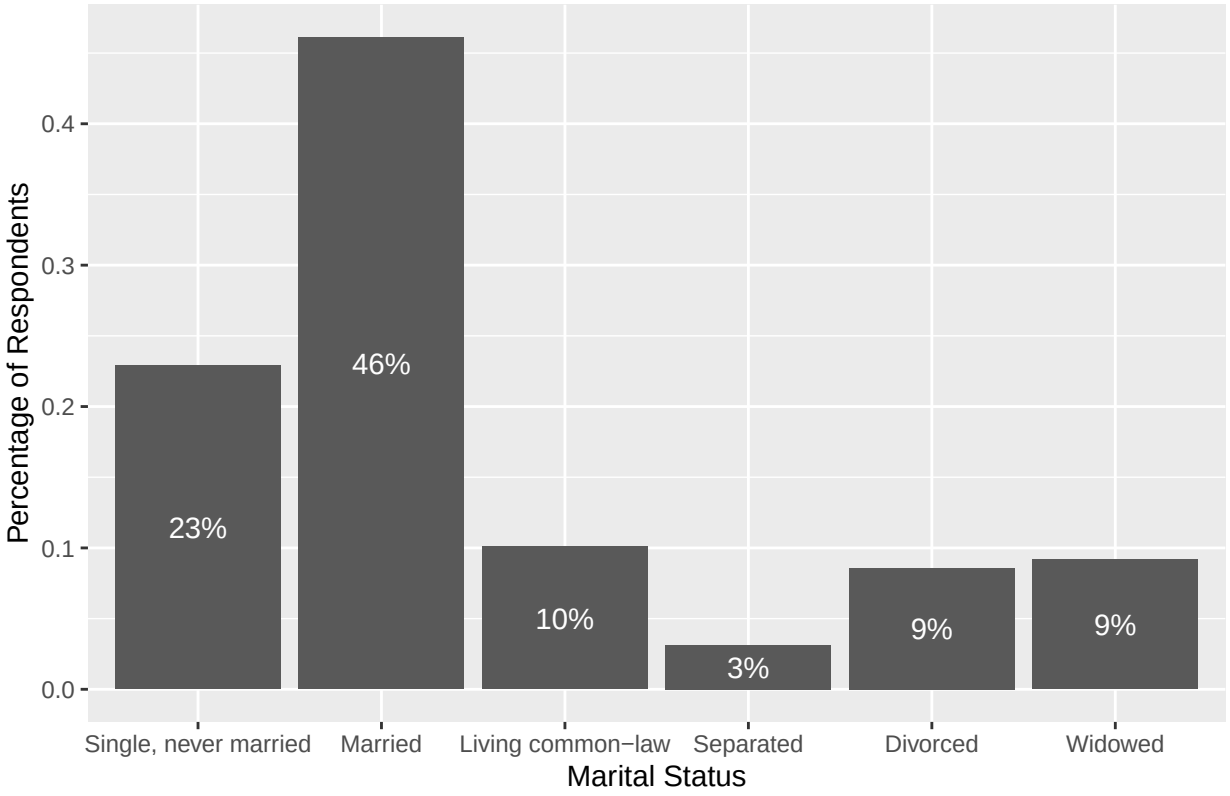

```
##      sex count
## 1 Female 11105
## 2   Male  9303
```

Figure 2: Distribution of the sex of respondents in percentages.



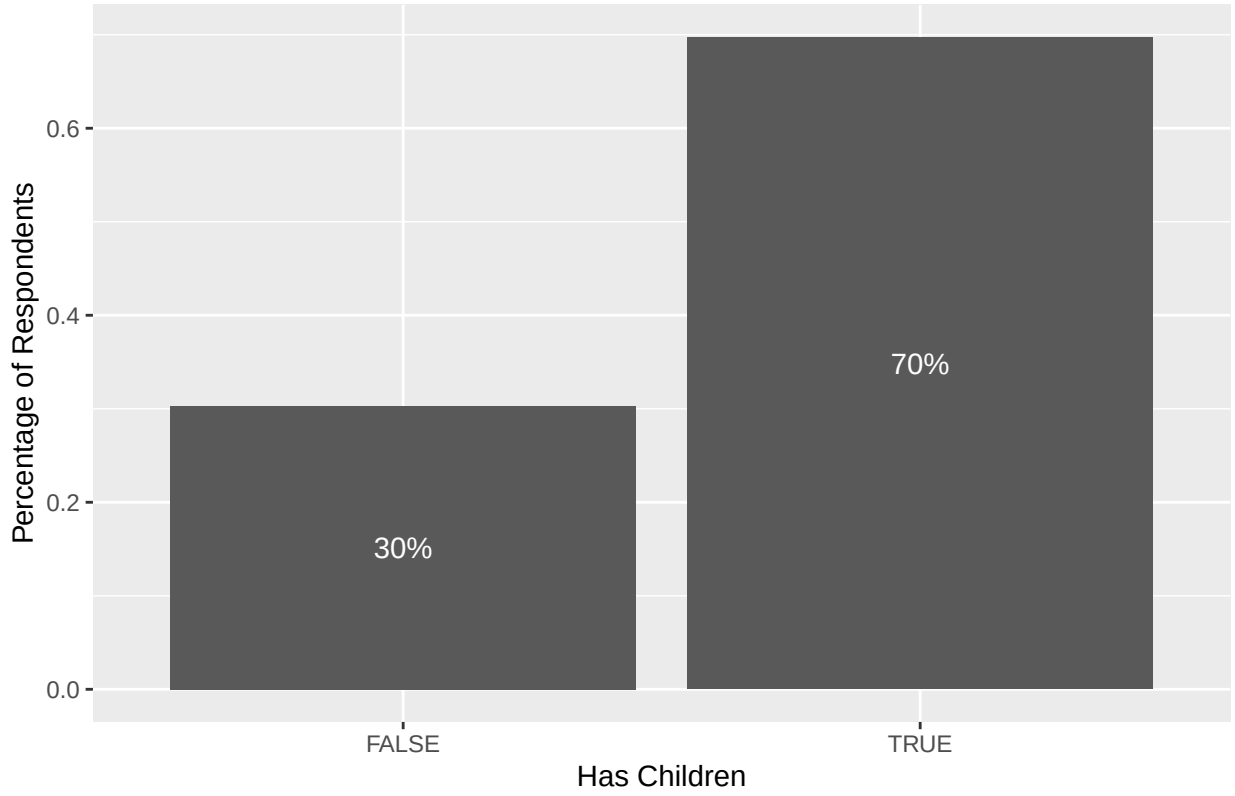

```
##          marital_status count
## 1              Divorced  1749
## 2     Living common-law  2059
## 3               Married  9411
## 4             Separated   639
## 5 Single, never married  4674
## 6               Widowed  1876
```

Figure 3: Distribution of the marital status of respondents in percentages.



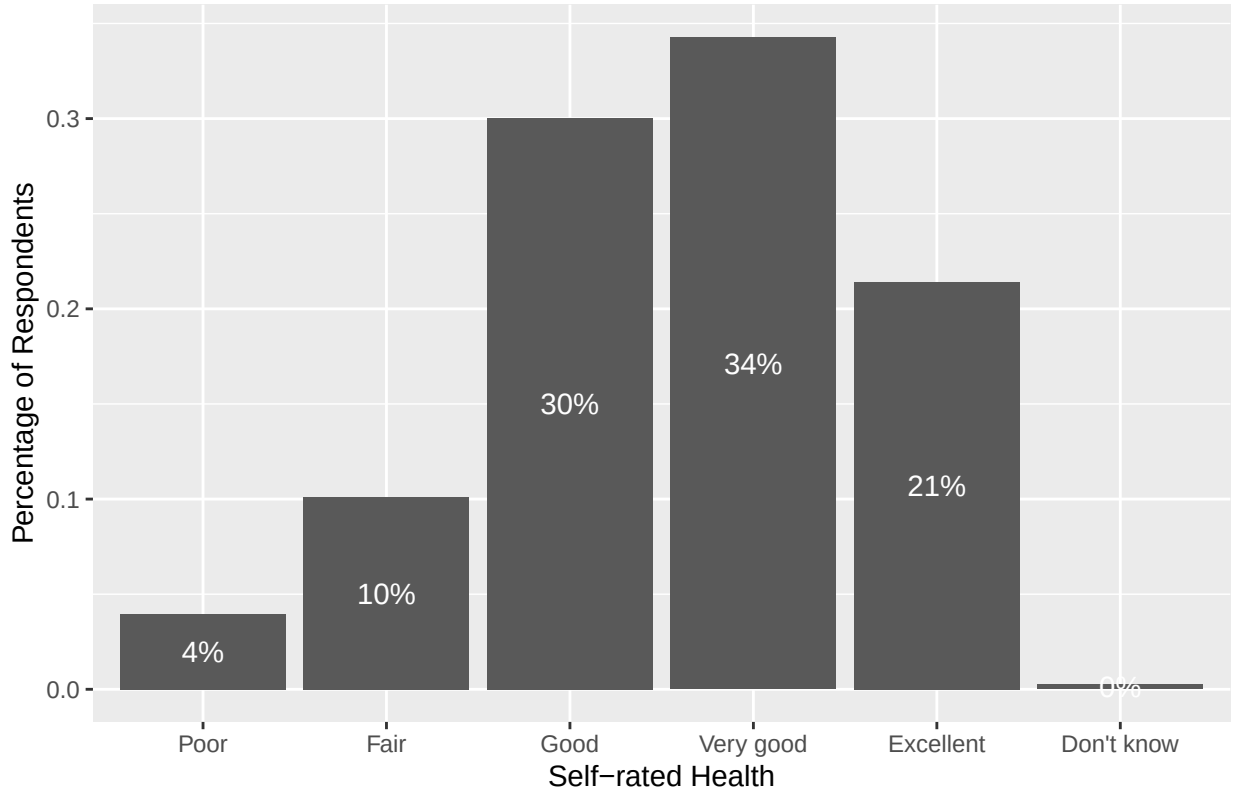

```
##   has_children count
## 1        FALSE  6179
## 2         TRUE 14229
```

Figure 4: Distribution of respondents with and without children.



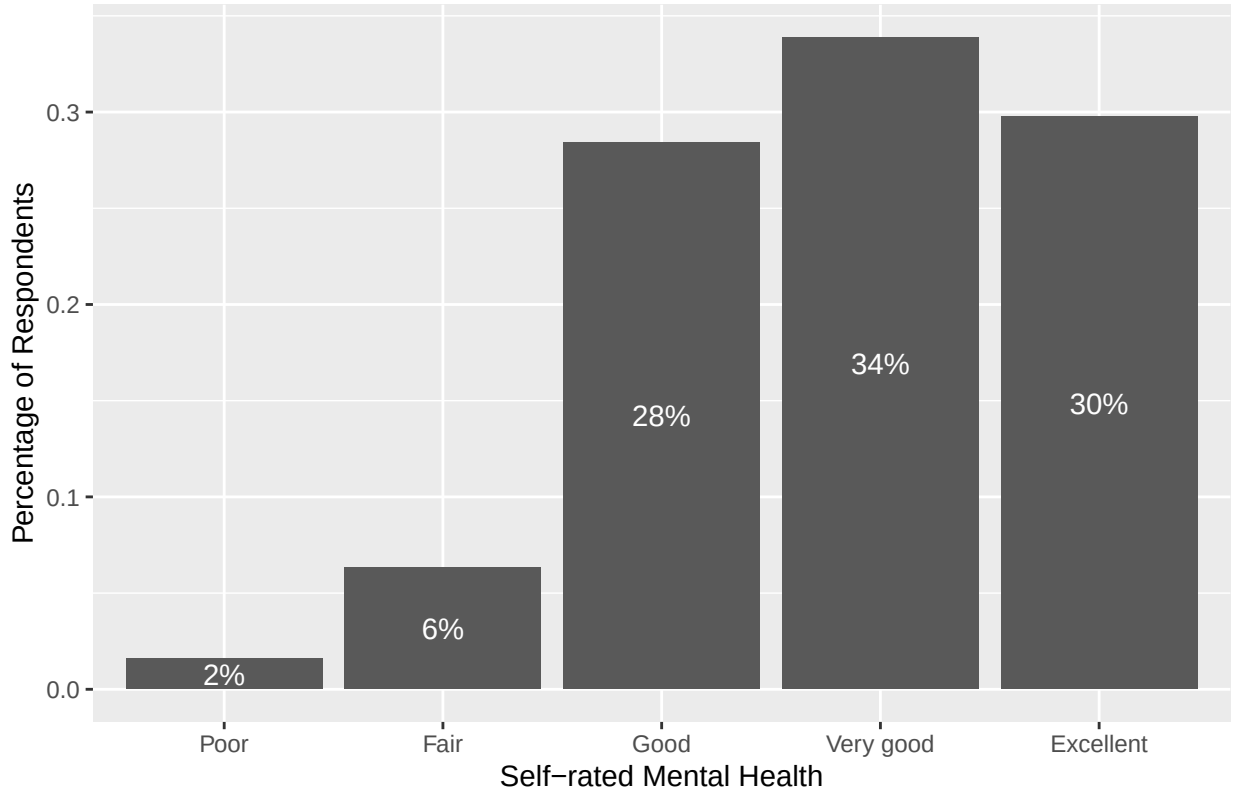

```
##       health count
## 1 Don't know    51
## 2  Excellent  4369
## 3       Fair  2062
## 4       Good  6126
## 5       Poor   806
## 6  Very good  6994
```

Figure 5: Distribution of the self−rated health of respondents in percentages




```
##   mental_health count
## 1     Excellent  6073
## 2          Fair  1294
## 3          Good  5801
## 4          Poor   325
## 5     Very good  6915
```

Figure 6: Self−rated mental health of respondents in percentages.



Out of the respondents, approximately 54% were female and 46% were male (Figure 2). Most of the respondents were in their 60's and 70's, while the next most common demographic were respondents in their 40's and 50's (Figure 1). This preliminary look at the dataset is fairly consistent with Canadian demographics according to the 2016 Census, with the female response being approximately 3% higher than expected and the average age being approximately 11 years older than expected (the average Canadian age is 41 while the average respondent age was 52). This older demographic makes sense as individuals less than 15 years of age were not eligible to respond to the survey and are therefore not represented here.

Of the 20,602 responses, 194 rows were dropped if the value for a variable was not available. For our purposes of attempting to model mental health, we consequently removed these individuals from the dataset we used in generating our model. Furthermore, according to Figure 5, the responses are heavily skewed towards positive responses, with 30% of respondents replying with 'Excellent' and 34% replying 'Very good'. 28% rated their mental health as 'Good' with the remaining 8% split 6 to 2 with regards to 'Fair' and 'Poor', respectively. These results overwhelmingly indicate that a large proportion of the sampled population feel that their mental is very strong. However, we proceed with modeling in the next section of this paper to better understand the contribution of the chosen demographic and family factors on self-rated mental. Is there a pattern of traits that separate "Excellent", "Very good", and "Good" ratings? What are the biggest distinctions between an individual with good mental health and poor mental health? These are some of the motivating questions we strive to answer with our model.

Model

The purpose of the model is to predict a person's self-rated mental health based our selected factors of age, sex, marital status, and self-rated physical health. Since self-rated mental health is a categorical data type in this dataset, the task at its core is a classification problem.

Model Selection

Some models that we considered were linear regression, naive Bayes, binary logistic regression, and multinomial logistic regression. To begin, linear regression is not suitable because it is often difficult to find an accurate linear relationship between predictors and categories, not to mention the fact that linear regression is more suitable when the dependent variable is continuous. Naive Bayes is a viable option since it is able to handle classification of more than two categories using joint probability and Bayes' Theorem. However, naive Bayes only works well under the assumption that the explanatory variables are independent, but this is often not the case. Given the context of the data, it is highly anticipated that the characteristics of a person are correlated in some way or another (e.g., age showing a correlation with self-rated physical health).

In addition, generative models (e.g., Naive Bayes) have a higher asymptotic error than discriminative models (e.g., logistic regression), but they approach the asymptotic error faster. In other words, discriminative models tend to perform better given a large enough dataset while generative models will perform better on small dataset as they learn faster. Since the dataset is large, choosing a discriminative model would be more appropriate for this task.

It is also important to note that the dependent variable has more than two categories. One way to handle this is to group multiple categories together so that there are only two categories. Then we would be able to use binary logistic regression to model the relationship. While this simplifies the complexity of implementing the model itself, there will be a loss in information from merging classes, impacting the strength of the conclusions we can draw. Another option to handle multi-classification is to use multinomial logistic regression which is an extension of binary logistic regression. The basic idea of this approach is to create a binary logistic regression model for each class. Each binary logistic regression model will do a one-versus-rest prediction for the corresponding class. The class with the highest probability will be the output prediction of the multinomial model. One of the caveats of this approach is that even more data is required to provide enough information for all binary logistic regression models for each class; otherwise, this method is prone to overfitting. This is less of a concern in this case because the size of the dataset is large enough to predict 5 classes. Thus, we chose multinomial logistic regression as our model for this task.

As described in the Data Characteristics section of this report, we chose four features (age, sex, marital_status, and self_rated_health) based on their perceived impact on self-rated mental health (self_rated_mental_health). We chose age as a continuous factor as opposed to the discrete version (where age is only given in whole numbers) in the interest of accuracy and being able to provide the model with the most fine-grained ata. Sex and marital status are categorical variables due to the nature of the information they provide (you're either single, married, in a common-law partnership, separated, divorced, or widowed, there is no status that exists in-between any of these options). Although self-rated health and self-rated mental health exists in real life on a continuous scale, due to the nature of responses gathered by the 2017 GSS, self_rated_health and self_rated_mental are categorical variables with possible values such as Poor, Fair, Good, Very good, and Excellent. There is no way for us to transform the categorical responses into a continuous variable that remains true to the data collected and there isn't any need to do so because we adjusted the selection of our model accordingly.

Mathematical Model

We fully derive the mathetical representation of the multinomial logistic regression model in the Appendix. In the interest of space, here we will just provide the equations we derived for the probability of each outcome:

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta_1} \cdot \boldsymbol{X_i}}}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol{\beta_2} \cdot \boldsymbol{X_i}}}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 3) = \frac{e^{\boldsymbol{\beta_3} \cdot \boldsymbol{X_i}}}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 4) = \frac{e^{\boldsymbol{\beta_4} \cdot \boldsymbol{X_i}}}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 5) = \frac{1}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

We will explain each variable that appears in the equations above (again, please see the Appendix for a more natural definition of each variable as we derive these equations). Y_i represents the outcome of the response variable for the ith observation. Since we have a total of 5 possible outcomes (Poor, Fair, Good, Very Good, and Excellent), we represent each of them as 1, 2, 3, 4, and 5 in the interest of space.

$\boldsymbol{\beta_k}$ is the row vector with elements that are the coefficients for the explanatory variables for the kth outcome. More explicitly, $\boldsymbol{\beta_k} = [\beta_{0,k}, \beta_{1,k}, \beta_{2,k}, \beta_{3,k}, \beta_{4,k}, \beta_{5,k}]$ where $\beta_{0,k}$ is the intercept for the kth outcome, $\beta_{1,k}$ is the coefficient for the first explanatory variable (age) for the kth outcome, $\beta_{2,k}$ is the coefficient for the second explanatory variable (sex) for the kth outcome, $\beta_{3,k}$ is the coefficient for the third explanatory variable (marital_status) for the kth outcome, $\beta_{4,k}$ is the coefficient for the fourth explanatory variable (has_children) for the kth outcome, and $\beta_{5,k}$ is the coefficient for the fifth explanatory variable (self_rated_health) for the kth outcome.

$\boldsymbol{x_i}$ is the row vector of explanatory variables for the ith observation. Specifically, $\boldsymbol{x_i} = [1, x_{1,i}, x_{2,i}, x_{3,i}, x_{4,i}, x_{5,i}]$ where $x_{1,i}$ is the value of the first explanatory variable (age) for the ith observation, $x_{2,i}$ is the value of the second explanatory variable (sex) for the ith observation, and so on.

Therefore, the first equation represents the probability that the self_rated_mental_health of the ith observation is Poor given values for age, sex, marital_status, has_children, and self_rated_health. The other 4 equations can be interpreted in a similar manner.

Running the Model

To estimate the model, we use the multinom function from the library nnet written in the programming language R. Our script is run using the software RStudio.

```
# install.packages("nnet")
library(nnet)
# install.packages("Metrics")
library(Metrics)

# shuffle rows of data tibble for cross-validation
# for reproducibility
set.seed(42)
# generate randomized indices for rows
shuffled_indices <- sample(total_count)
# shuffle data tibble accordingly
data <- data[shuffled_indices,]
# set boundary between training and testing
boundary = as.integer(total_count * 0.8)
# take subset of dataset to form training dataset
training_dataset = data[0:boundary,]

# run model where self_rated_mental_health is response variable and age, sex,
# marital status, has children, and self rated health are explanatory variables
model <- nnet::multinom(self_rated_mental_health ~ age + sex + marital_status + has_children +
                          self_rated_health, data = training_dataset)
```

```
## # weights:  75 (56 variable)
## initial  value 26275.683358
## iter  10 value 19588.463000
## iter  20 value 19283.841204
## iter  30 value 18586.312225
## iter  40 value 18332.871105
## iter  50 value 18246.392191
## iter  60 value 18182.094276
## final  value 18179.462642
## converged
```

Our model successfully converged after 60 iterations on the training dataset to a final negative log-likelihood value of 18179. Although this a really large number, this is not too bad considering we have over 16300 observations in the training dataset and the negative log-likelihood is equal to the likelihood we would observe the specific dataset (after taking the log and multiplying by -1).

We did not run into any diagnostic issues when running our model.

As we hinted at with the phrase "training dataset", we conducted cross-validation to check our model. The procedure and the accuracy of our model is covered in the next section (Results).

Results

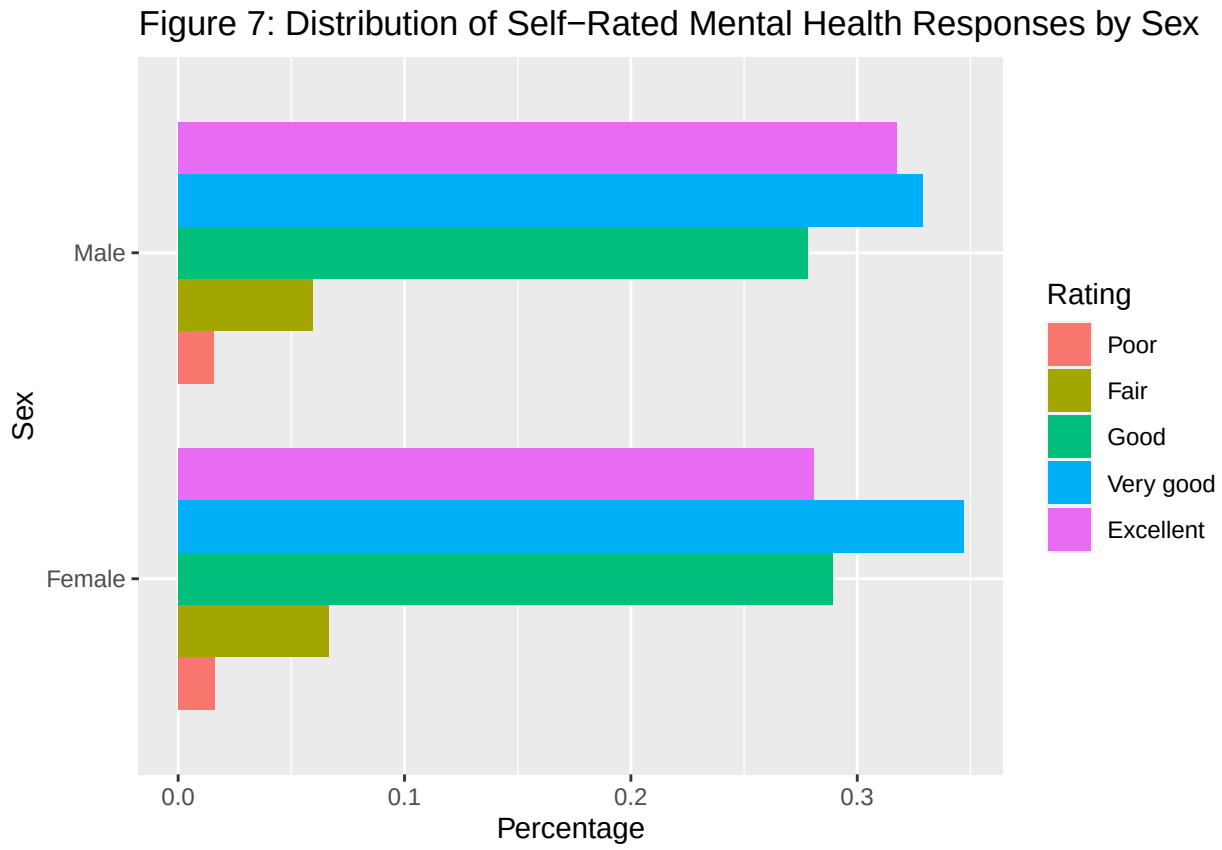
We will graph the relationship between mental health and each of the explanatory variables as well as discuss the results of our model.

```
grouped_bar_chart <- function(data, xlab, ylab, title){
  ggplot(data,
       aes(y=grouped, x=value, fill=Rating, label=Rating))+
  geom_bar(stat="identity",position="dodge", width=0.8) +
  scale_x_continuous(ylab) +
  scale_y_discrete(xlab) +
  ggtitle(title)
}
# graph distribution of self-rated mental health for each gender

# data parsing into format I want for chart
grouped <- data %>%
  count(sex, self_rated_mental_health) %>%
  rename(
    count=n
  )
#total number of votes
total_vote <- aggregate(grouped$count,
                        by=list(sex=grouped$sex), FUN=sum)

#process data into desired graphing format
grouped<-left_join(grouped, total_vote, by="sex") %>%
  rename(total=x) %>%
  mutate(value=count/total) %>%
  rename(grouped=sex, Rating=self_rated_mental_health)
# manually order bars in bar graph
grouped$Rating <- factor(grouped$Rating,
```

```
                                levels = c("Poor", "Fair", "Good", "Very good",
                                           "Excellent", "Don't know"))
grouped
```

```
## # A tibble: 10 x 5
##    grouped Rating    count total  value
##    <chr>   <fct>     <int> <int>  <dbl>
##  1 Female  Excellent  3120 11105 0.281
##  2 Female  Fair        740 11105 0.0666
##  3 Female  Good       3212 11105 0.289
##  4 Female  Poor        179 11105 0.0161
##  5 Female  Very good  3854 11105 0.347
##  6 Male    Excellent  2953  9303 0.317
##  7 Male    Fair        554  9303 0.0596
##  8 Male    Good       2589  9303 0.278
##  9 Male    Poor        146  9303 0.0157
## 10 Male    Very good  3061  9303 0.329
```

```
grouped_bar_chart(grouped, "Sex", "Percentage",
                  "Figure 7: Distribution of Self-Rated Mental Health Responses by Sex")
```



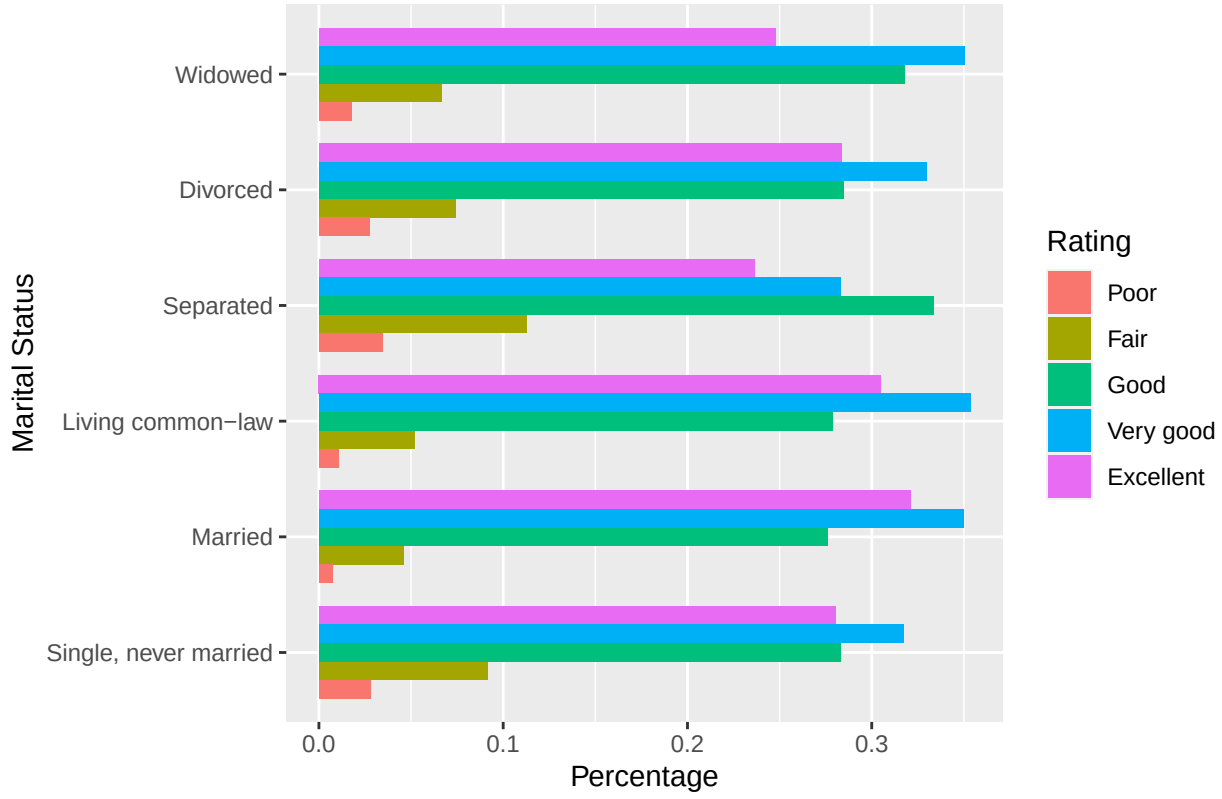

```
# graph marital status vs distribution of self rated mental health
#data parsing into format I want for chart
grouped <- data %>%
```

```
  count(marital_status, self_rated_mental_health) %>%
  rename(
    count=n
  )

#total number of votes
total_vote <- aggregate(grouped$count,
                        by=list(marital_status=grouped$marital_status), FUN=sum)

#process data into desired graphing format
grouped<-left_join(grouped, total_vote, by="marital_status") %>%
  rename(total=x) %>%
  mutate(value=count/total) %>%
  rename(grouped=marital_status, Rating=self_rated_mental_health)
# manually order bars in bar graph
grouped$Rating <- factor(grouped$Rating,
                                levels = c("Poor", "Fair", "Good", "Very good",
                                           "Excellent", "Don't know"))
# manually categories on y-axis
grouped$grouped <- factor(grouped$grouped,
                                levels = c("Single, never married",
                                                        "Married", "Living common-law",
                                                        "Separated", "Divorced",
                                                        "Widowed"))
grouped
```

```
## # A tibble: 30 x 5
##    grouped           Rating    count total  value
##    <fct>             <fct>     <int> <int>  <dbl>
##  1 Divorced          Excellent   496  1749 0.284
##  2 Divorced          Fair        130  1749 0.0743
##  3 Divorced          Good        498  1749 0.285
##  4 Divorced          Poor         48  1749 0.0274
##  5 Divorced          Very good   577  1749 0.330
##  6 Living common-law Excellent   628  2059 0.305
##  7 Living common-law Fair        107  2059 0.0520
##  8 Living common-law Good        574  2059 0.279
##  9 Living common-law Poor         22  2059 0.0107
## 10 Living common-law Very good   728  2059 0.354
## # ... with 20 more rows
```

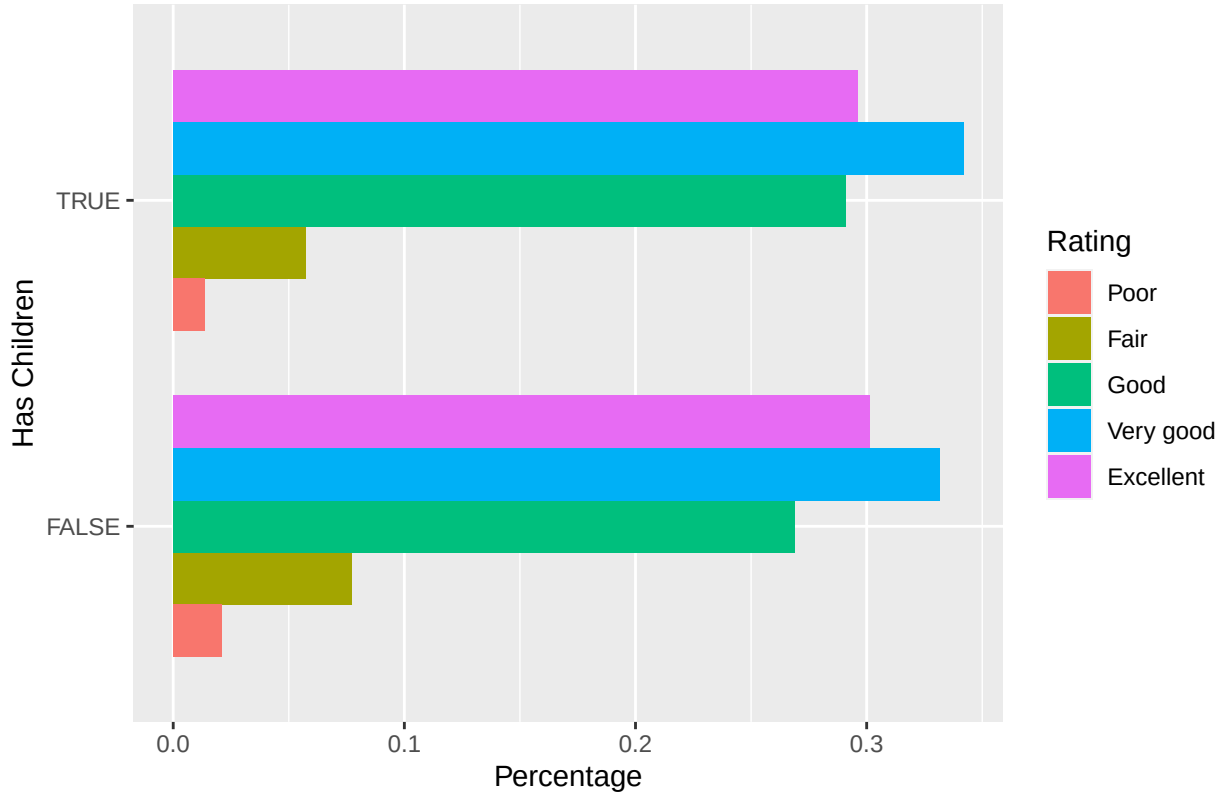
```
grouped_bar_chart(grouped, "Marital Status",
                  "Percentage",
                  "Figure 8: Self-Rated Mental Health Against Marital Status")
```

Figure 8: Self−Rated Mental Health Against Marital Status



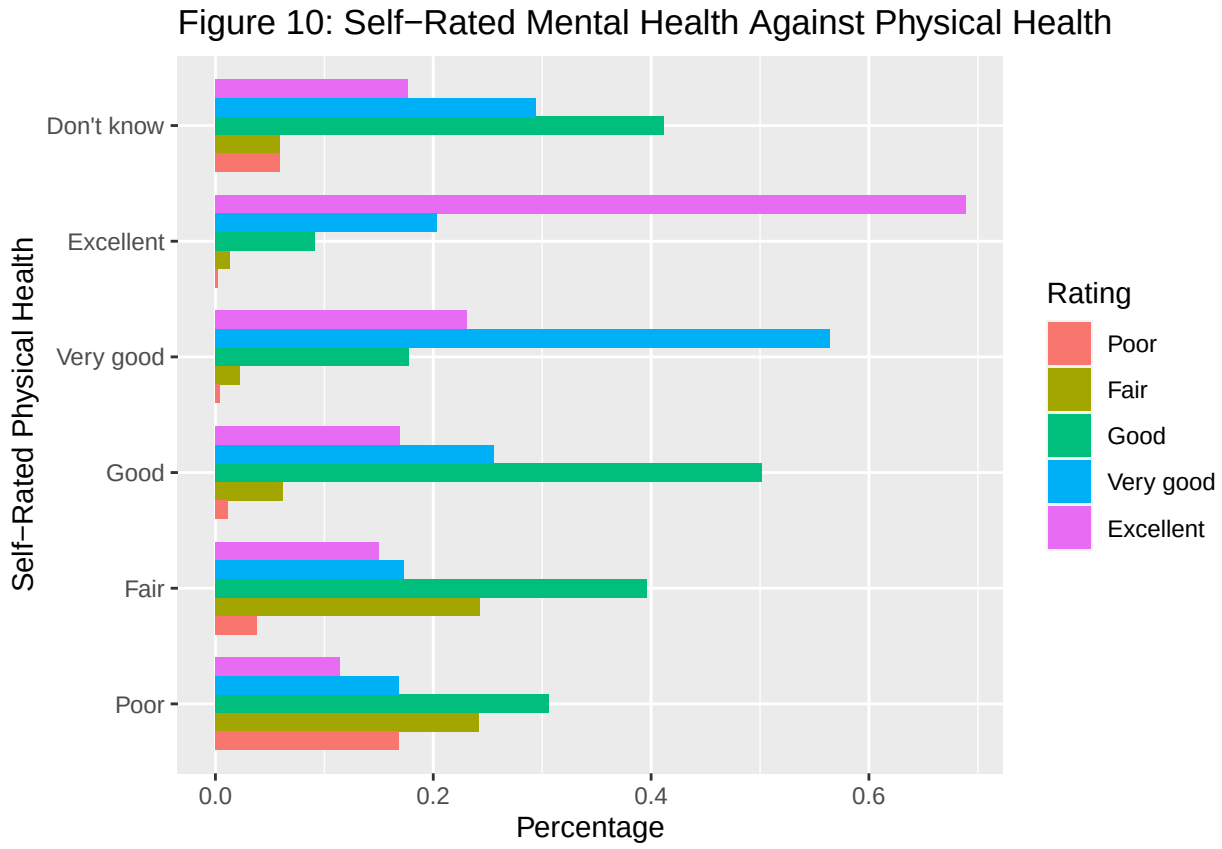

```
# graph distribution of self-rated mental health for each gender

# data parsing into format I want for chart
grouped <- data %>%
  count(has_children, self_rated_mental_health) %>%
  rename(
    count=n
  )
#total number of votes
total_vote <- aggregate(grouped$count,
                        by=list(has_children=grouped$has_children), FUN=sum)

#process data into desired graphing format
grouped<-left_join(grouped, total_vote, by="has_children") %>%
  rename(total=x) %>%
  mutate(value=count/total) %>%
  rename(grouped=has_children, Rating=self_rated_mental_health)
# manually order bars in bar graph
grouped$Rating <- factor(grouped$Rating,
                                levels = c("Poor", "Fair", "Good", "Very good",
                                           "Excellent", "Don't know"))
grouped_bar_chart(grouped, "Has Children", "Percentage",
                  "Figure 9: Self-Rated Mental Health With and Without Children")
```

Figure 9: Self−Rated Mental Health With and Without Children




```
# graph distribution of self-rated mental health for each gender

# data parsing into format I want for chart
grouped <- data %>%
  count(self_rated_health, self_rated_mental_health) %>%
  rename(
    count=n
  )
#total number of votes
total_vote <- aggregate(grouped$count,
                        by=list(self_rated_health=grouped$self_rated_health), FUN=sum)

#process data into desired graphing format
grouped<-left_join(grouped, total_vote, by="self_rated_health") %>%
  rename(total=x) %>%
  mutate(value=count/total) %>%
  rename(grouped=self_rated_health, Rating=self_rated_mental_health)
# manually order bars in bar graph
grouped$Rating <- factor(grouped$Rating,
                                levels = c("Poor", "Fair", "Good", "Very good",
                                           "Excellent", "Don't know"))
# manually categories on y-axis
grouped$grouped <- factor(grouped$grouped,
                                levels = c("Poor", "Fair", "Good", "Very good",
                                           "Excellent", "Don't know"))
grouped_bar_chart(grouped, "Self-Rated Physical Health", "Percentage",
```

```
  "Figure 10: Self-Rated Mental Health Against Physical Health")
```




```
summary(model)
```

```
## Call:
## nnet::multinom(formula = self_rated_mental_health ~ age + sex +
##     marital_status + has_children + self_rated_health, data = training_dataset)
##
## Coefficients:
##           (Intercept)          age    sexMale marital_statusLiving common-law
## Fair        0.9337995 -0.030530728 -0.2999169                      -0.2103499
## Good        1.2018423 -0.011147383 -0.2693399                       0.1470838
## Poor        1.9127518 -0.044465991 -0.2423522                      -0.8218660
## Very good   0.4184083 -0.004902573 -0.2110663                       0.1134176
##           marital_statusMarried marital_statusSeparated
## Fair                -0.42164405              0.5830375
## Good                -0.02828272              0.3532270
## Poor                -0.97325625              0.6550756
## Very good           -0.03537081              0.0680468
##           marital_statusSingle, never married marital_statusWidowed
## Fair                                0.2548927            0.11257388
## Good                                0.2748687            0.19844122
## Poor                               -0.0154718            0.06705713
## Very good                           0.0891638            0.17794433
##           has_childrenTRUE self_rated_healthExcellent self_rated_healthFair
```

```
## Fair            0.10194759                  -3.389274             1.5483003
## Good            0.19639066                  -2.790659             0.3892906
## Poor            0.04146269                  -5.040113            -0.2449083
## Very good       0.09421069                  -1.383704             0.0524162
##           self_rated_healthGood self_rated_healthPoor
## Fair                -0.2345330             1.8856345
## Good                 0.3978945             0.4278481
## Poor                -2.0682804             1.6473043
## Very good            0.2858592             0.3824895
##           self_rated_healthVery good
## Fair                        -1.691814
## Good                        -1.015674
## Poor                        -3.661171
## Very good                    0.724645
##
## Std. Errors:
##           (Intercept)         age    sexMale marital_statusLiving common-law
## Fair        0.7223167 0.002799945 0.07755772                        0.1817139
## Good        0.4625500 0.001671573 0.04734890                        0.1120796
## Poor        0.7953331 0.005293736 0.14031256                        0.3398460
## Very good   0.5041773 0.001530634 0.04348912                        0.1031619
##           marital_statusMarried marital_statusSeparated
## Fair                 0.13709454              0.2232300
## Good                 0.08769625              0.1591539
## Poor                 0.23639191              0.3249925
## Very good            0.08156606              0.1537470
##           marital_statusSingle, never married marital_statusWidowed
## Fair                                0.15973973             0.1745110
## Good                                0.10688746             0.1135165
## Poor                                0.25572329             0.2847674
## Very good                           0.09986703             0.1063236
##           has_childrenTRUE self_rated_healthExcellent self_rated_healthFair
## Fair             0.10551905                  0.7050123             0.6927843
## Good             0.06595897                  0.4448638             0.4464983
## Poor             0.18496512                  0.7721172             0.7148895
## Very good        0.05998372                  0.4890369             0.4944139
##           self_rated_healthGood self_rated_healthPoor
## Fair                  0.6915177             0.7034397
## Good                  0.4421292             0.4622504
## Poor                  0.7167538             0.7183514
## Very good             0.4888389             0.5103979
##           self_rated_healthVery good
## Fair                        0.6950549
## Good                        0.4425959
## Poor                        0.7400007
## Very good                   0.4880967
##
## Residual Deviance: 36358.93
## AIC: 36470.93
```

Here, we have output a summary of the coefficients of our model. Recall the mathematical representation of our multinomial logistic regression model. The equations used to compute probabilities for each outcome only need coefficients for non-pivots which is why the outcome Excellent (which serves as the pivot in our model) doesn't appear in the table. Another interesting thing to note is that because the variables sex,

marital_status, has_children, and self_rated_health are categorical, they are represented using dummy variables in our model.

```
z <- summary(model)$coefficients/summary(model)$standard.errors
(1 - pnorm(abs(z), 0, 1)) * 2
```

```
##           (Intercept)          age      sexMale marital_statusLiving common-law
## Fair      0.196085729 0.000000e+00 1.101752e-04                      0.24703213
## Good      0.009368742 2.579070e-11 1.282291e-08                      0.18941347
## Poor      0.016173819 0.000000e+00 8.412618e-02                      0.01559115
## Very good 0.406604740 1.360185e-03 1.214164e-06                      0.27158761
##           marital_statusMarried marital_statusSeparated
## Fair               2.100997e-03             0.009006069
## Good               7.470681e-01             0.026459143
## Poor               3.836196e-05             0.043835174
## Very good          6.645454e-01             0.658062723
##           marital_statusSingle, never married marital_statusWidowed
## Fair                               0.11056143            0.51887422
## Good                               0.01012383            0.08044201
## Poor                               0.95175572            0.81383595
## Very good                          0.37195081            0.09420699
##           has_childrenTRUE self_rated_healthExcellent self_rated_healthFair
## Fair            0.33396742               1.529080e-06            0.02542422
## Good            0.00290641               3.540126e-10            0.38327667
## Poor            0.82262898               6.680811e-11            0.73191290
## Very good       0.11627462               4.662839e-03            0.91556899
##           self_rated_healthGood self_rated_healthPoor
## Fair                 0.734491581           0.007349217
## Good                 0.368146475           0.354666020
## Poor                 0.003906413           0.021838016
## Very good            0.558701201           0.453619347
##           self_rated_healthVery good
## Fair                     1.493000e-02
## Good                     2.174390e-02
## Poor                     7.516354e-07
## Very good                1.376408e-01
```

Here we conduct a two-tailed z-test on the coefficients. To set up our analysis in the discussion section, we will define the null and alternative hypothesis here. The null hypothesis for each coefficient is that the coefficient is equal to 0, i.e., there is no difference in likelihood between one of the four values of the response variable ("Fair", "Good", "Poor", or "Very good") compared to the value "Excellent" for the given explanatory variable. The alternative hypothesis is therefore that the coefficient is not equal to zero, i.e., that there is a difference between the odds of the value of the response variable being a non-pivot outcome ("Fair", "Good", "Poor", or "Very good") and the odds of the response variable being "Excellent". Setting the significance level at 0.05, if the p-value for the coefficient of a certain explanatory variable is less than 0.05, then we can reject the null hypothesis and conclude there is statistically significant evidence that there is a difference in odds between the response variable and "Excellent" given the explanatory variable.

```
# take subset of dataset to form testing dataset
testing_dataset = data[boundary:total_count,]
testing_dataset = testing_dataset[complete.cases(testing_dataset), ]
# predict what is the likelihood of each outcome for each observation in the testing dataset
testing_probabilities <- predict(model, newdata = testing_dataset, "probs")
```

```r
# identify the predicted outcome based on which probability is the largest
# find the column index with the largest probability
column_index <- max.col(testing_probabilities, tie="random")
# form a list of outcomes with column names for testing dataset
testing_predictions <- colnames(testing_probabilities)[column_index]
testing_ground_truth <- testing_dataset %>% pull(self_rated_mental_health)
accuracy(testing_ground_truth, testing_predictions)
```

```
## [1] 0.5486162
```

Additionally, we checked the predictions of our model by conducting cross validation. We split the shuffled dataset into training and testing subsets (an 80-20 split, respectively). Then, we trained the model using only the training data and made predictions on the testing data. Finally, we compared the model's predictions to the ground truth values for self-rated mental health for the test set. Our predictions for the test set had an accuracy of 54.9%.

Discussion

As depicted in Figure 7, the main difference between male and female self-rated mental health is the gap between the percentages of "Excellent" and "Very good" responses. Namely, 28% of female respondents reported their mental health was "Excellent" while 32% of male respondents reported the same answer. Furthermore, 35% of female respondents reported their mental health was "Very good" while only 33% of male respondents reported the same answer. Overall, a larger percentage of males reported that their mental health is eitiher "Very good" or "Excellent" and there isn't as much of a difference in the distribution of those responses as was observed in female responses. This suggests that females are less likely to consider their mental health as being "Excellent" compared to males.

A smaller percentage of respondents rated their mental health as poor if they were married (0.7%) or living in a common-law relationship (1.1%) (Figure 8). In order of the percentage of "Poor" responses received, it goes Widowed (1.8%), Divorced (2.7%), Single Never Married (2.8%), and Separated (3.4%). Interestingly, individuals who were widowed were much less likely to rate their mental health as poor compared to individuals who were either divorced, never married, or separated. Furthermore, married and living in a common-law relationship all received the highest response rates for "Excellent" at 32.1% and 30.1%, respectively. These results seem to strongly indicate that the bond between partners have an impact on mental health and lacking that bond may be detrimental to one's mental health.

Surprisingly, there is not a significant difference between self-rated mental health with and without children (Figure 9).

Lastly, there seems to be the largest variation in responses for self-rated mental health with respect to self-rated physical health (Figure 10). In fact, there appears to be a direct correlation between self-ratings of mental health and self-ratings of physical health. The largest percentage of "Excellent" ratings for mental health occurred in the cohort with "Excellent" physical health, the largest rating of "Very good" mental health occurred in the cohort with "Very good" physical health, and so forth. The opposite trend also exists where smallest rating of "Poor" mental health occurs in the cohort with "Excellent" physical health, and so on. This is evidence that there is a strong correlation between physical health and mental health despite being fundamentally different aspects of one's life. That is, an individual with self-evaluated "Excellent" physical health likely has a corresponding "Excellent" self-rating of mental health. This could be due to a variety of factors, for example having higher self-confidence in their physical fitness, experiencing the benefits of regular exercise on their mental health, and so forth.

By analyzing the coefficients of our model, we can determine what kind of effect certain each variable has on an individual's self-rating of their mental health. For example, being married makes individuals less likely

to report "Poor" mental health compared to "Excellent" mental health by -0.97 logarithmic units. However, being married has half the effect on the odds of having a "Fair" mental health self-rating compared to "Excellent" and close to no affect on differentiating between "Good" and "Excellent" and "Very good" and "Excellent." Noticeably, a strongest correlation is found between self-rated physical health and self-rated mental health as mentioned previously. An individual marking themselves as having excellent self-rated physical health decreases their logarithmic chances of having poor self-rated mental health by 5 compared to their chances of having excellent self-rated mental. Variables such as age and number of kids have a limited effect on a person's mental health rating.

The two-tailed z-test conducted on the coefficients indicates unsurprisingly that the most statistically significant coefficients are found in the self-rated mental health explanatory variables, with all coefficients being statistically significant for the explanatory variable excellent self-rated physical health and 3/4 coefficients being statistically significant for the explanatory variable very good self-rated physical health. Although some coeffcients such as those for age and sex are also statistically significant, the values of the coefficients themselves are not very large (close to 0 for age and within 0.3 logarithmic units of 0 for sex). Therefore, the tests appear to reflect that there is little difference between the odds for non-pivot outcomes and the "Excellent" outcome given the explanatory variable age and sex.

Lastly, we will discuss the accuracy of our model. Compared to random predictions which would have an accuracy of only 20% (since there are five possible outcomes for the response variable), our cross validation results indicate that our model performs at least twice as well. Furthermore, this accuracy is calculated based on exact matches between predicted and ground truth values. If we considered predictions of Very good to be close enough to ground truth values of Excellent, the accuracy of our model will be much higher.

```
map <- data.frame(find=c('Excellent', 'Very good', 'Good', 'Fair', 'Poor'),replace=c('Excellent', 'Excel
testing_ground_truth_simplified <- as.character(map[match(testing_ground_truth, map$find), "replace"])
testing_predictions_simplified <- as.character(map[match(testing_predictions, map$find), "replace"])
accuracy(testing_ground_truth_simplified, testing_predictions_simplified)
```

```
## [1] 0.671565
```

We gain a sense of the relative accuracy of our model by mapping "Very good" values to "Excellent" values and "Fair" values to "Poor" values. This has a 67.2% accuracy which again reinforces the idea that the model has identified non-trivial relationships between the explanatory and response variables.

In summary, although we have found that our model is able to make non-trivial predictions about mental health self-ratings given demographic and family traits for an individual, we also determined that not all of our chosen explanatory variables are equally important in making that prediction.

Weaknesses and future work

Our survey had many weaknesses, such as the fact that we chose to analyze a variable that is very hard to predict, prone to bias and relative to the respondents experience towards mental health. Another large weakness was the data excluded youths aged 15 and younger. This effects our results significantly because youths mental illness is a prominent issue as half of all mental health issues begin by the age of 15 and the lack of this data does not allow us to represent mental health thoroughly. This was a weakness we were aware of and knew it would affect our outcome. However, we were still curious to see how the responses would go and determine if the results can lead to insights. We also were not able to spend time to include all the variables we wanted. Variables such as income, detailed family situations and living situations were things we planned on adding to our model but due to time and covariance concerns we were not able to integrate them into our model. This would be part of our future plans in order to build a more encompassing model. Furthermore, we hope to gain less bias information on an individuals mental health by utilizing mental health surveys dedicated to measuring an individuals mental health with less bias.

Appendix

Derivation of the mathematical representation of the multinomial logistic regression model

Recall that the multinomial logistic regression consists of several binary logistic regression models. Like binary logistic regression, multinomial logistic regression predicts the probability that the ith observation has outcome k using the following function:

$$f(k,i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + ... + \beta_{M,k}x_{M,i}$$

where $\beta_{m,k}$ is the coefficient for the mth explanatory variable and the kth outcome while $x_{m,i}$ is the value of the mth explanatory variable for the ith observation. In our case, we have $M = 5$ (age, sex, marital_status, has_children self_rated_health) explanatory variables so the function as applicable to our model is:

$$f(k,i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \beta_{3,k}x_{3,i} + \beta_{4,k}x_{4,i} + \beta_{5,k}x_{5,i}$$

Note that we can represent $\beta_{0,k}, \beta_{1,k}, \beta_{2,k}, \beta_{3,k}, \beta_{4,k}, \beta_{5,k}$ and $1, x_{1,i}, x_{2,i}, x_{3,i}, x_{4,i}, x_{5,i}$ as row vectors $\boldsymbol{\beta_k}$ and $\boldsymbol{x_i}$, respectively. Then, the function can be simplified as follows:

$$f(k,i) = \boldsymbol{\beta_k} \cdot \boldsymbol{x_i}$$

where we take the dot product of the two row vectors we just defined.

As previously mentioned, the multinomial logistic regression model is a series of binary logistic regressions where the probability of each outcome of the response variable (self-rated mental health) is regressed against a chosen pivot outcome. Let Y_i represent the outcome of the response variable for the ith observation. We have a total of 5 possible outcomes (Poor, Fair, Good, Very Good, and Excellent represented as 1, 2, 3, 4, and 5, respectively). Let's choose the last outcome (Excellent or 5) as the pivot. In mathematical notation, this is:

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = 5)} = \boldsymbol{\beta_1} \cdot \boldsymbol{X_i}$$

$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = 5)} = \boldsymbol{\beta_2} \cdot \boldsymbol{X_i}$$

$$\ln \frac{\Pr(Y_i = 3)}{\Pr(Y_i = 5)} = \boldsymbol{\beta_3} \cdot \boldsymbol{X_i}$$

$$\ln \frac{\Pr(Y_i = 4)}{\Pr(Y_i = 5)} = \boldsymbol{\beta_4} \cdot \boldsymbol{X_i}$$

Then, we solve for the probabilities by exponentiating both sides:

$$\Pr(Y_i = 1) = \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_1} \cdot \boldsymbol{X_i}}$$

$$\Pr(Y_i = 2) = \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_2} \cdot \boldsymbol{X_i}}$$

$$\Pr(Y_i = 3) = \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_3} \cdot \boldsymbol{X_i}}$$

$$\Pr(Y_i = 4) = \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_4} \cdot \boldsymbol{X_i}}$$

The probability of the pivot outcome can be calculated because we know that the probability of all outcomes must sum to 1:

$$\Pr(Y_i = 5) = 1 - \left(\Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_1} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_2} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_3} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_4} \cdot \boldsymbol{X_i}}\right)$$

$$1 = \Pr(Y_i = 5) + \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_1} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_2} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_3} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 5) \cdot e^{\boldsymbol{\beta_4} \cdot \boldsymbol{X_i}}$$

$$1 = \Pr(Y_i = 5)\left(1 + e^{\boldsymbol{\beta_1}\cdot\boldsymbol{X_i}} + e^{\boldsymbol{\beta_2}\cdot\boldsymbol{X_i}} + e^{\boldsymbol{\beta_3}\cdot\boldsymbol{X_i}} + e^{\boldsymbol{\beta_4}\cdot\boldsymbol{X_i}}\right)$$

$$1 = \Pr(Y_i = 5)\left(1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k}\cdot\boldsymbol{X_i}}\right)$$

$$\Pr(Y_i = 5) = \frac{1}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k}\cdot\boldsymbol{X_i}}}$$

Having the expression for $\Pr(Y_i = 4)$, we can represent the probabilities of the other outcomes as follows:

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta_1}\cdot\boldsymbol{X_i}}}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k}\cdot\boldsymbol{X_i}}}$$

$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol{\beta_2}\cdot\boldsymbol{X_i}}}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k}\cdot\boldsymbol{X_i}}}$$

$$\Pr(Y_i = 3) = \frac{e^{\boldsymbol{\beta_3}\cdot\boldsymbol{X_i}}}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k}\cdot\boldsymbol{X_i}}}$$

$$\Pr(Y_i = 4) = \frac{e^{\boldsymbol{\beta_4}\cdot\boldsymbol{X_i}}}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k}\cdot\boldsymbol{X_i}}}$$

$$\Pr(Y_i = 5) = \frac{1}{1 + \sum_{k=1}^{4} e^{\boldsymbol{\beta_k}\cdot\boldsymbol{X_i}}}$$

References

Interview method/survey size: https://www.statcan.gc.ca/eng/survey/household/4501 Detailed information about GSS 2017: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816 Questionnaire: https://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=assembleInstr&lang=en&Item_Id=335815#qb345205 Mental health statistics: https://www.camh.ca/en/Driving-Change/The-Crisis-is-Real/Mental-Health-Statistics Discriminative vs Generative Classifiers (Naive Bayes vs lostic regression): https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf Age and sex https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/as/Table.cfm?Lang=E&T=21