Stat 109 Final Project

Paweł Rybacki

May 11, 2020

Predicting Happiness and Life Satisfaction Globally

1. Research Question

What are the statistically significant predictors of 1) the feeling of happiness, and 2) life satisfaction?

2. Motivation

Although not all of us have a well-defined purpose of life, we all can feel different forms and degrees of happiness and of life satisfaction. When growing up, we shape our beliefs and set ourselves goals based on them. However, while education systems help us form our opinions and realize many of these goals, we often end up ignorant about what ultimately matters in life. The purpose of this paper is to establish what factors are associated with the feelings of happiness and life satisfaction. Since establishing casual relationships is beyond the scope of this paper, since one size does not always fit all, and since the project has many limitations, none of the results is a recipe for happiness. However, the surprising little discoveries my analysis uncovers are a good starting point to reflect again on what brings us happiness and life satisfaction.

3. Data

a) Dataset

My dataset comes from the World Values Survey (WVS) and includes 98 countries over the years 1981-2016. The WVS is an international, longitudinal project measuring the values, beliefs, opinions, attitudes, habits together with demographics from people across the globe. This is an excellent source of information for psychologists, sociologists, economists, and other social scientists. The authors boast "the largest non-commercial, cross-national, time series investigation of human beliefs and values ever executed" that helps "analyze such topics as economic development, democratization, religion, gender equality, social capital, and subjective well-being."[1] I have previously used the dataset for my project on the sociopolitical determinants of trust.

The original dataset contains 348,532 observations. Due to listwise deletion of all problematic observations, my final dataset is limited to 221,813 observations.

---

[1] http://www.worldvaluessurvey.org/WVSContents.jsp

b) Variables

From 1,446 variables in the original dataset, I narrowed the list down by selection based on relevance and the number of non-problematic observations. As a result of variable selection and splitting into binaries, I obtained 32 variables in my final dataset. The variables are measures of different aspects of a person's life situation and values, such as financial situation, form of employment, health, religious beliefs, feeling of happiness, feeling of freedom, and other attitudes and habits. I renamed all variables from codes into intuitive names.

For each variable in the dataset, the values of -5, -4, -3, -2, and -1 mean 'Missing; Unknown,' 'Not asked in survey,' 'Not applicable,' 'No answer,' and 'Don´t know.' I replaced these negative values of each variable by a missing value marker (i.e. *NA* in R). Some dummies, such as `sex` are coded as '1' for the positive response and '2' for the negative response. I changed them into '0' meaning the negative response and '1' meaning the positive response: `male` replaced `sex` and is still indicated by '1', while females are my baseline.

Factor variables in the dataset were inconsistent in terms of scale. Some of them took values from 0 to 10, others from 0 to 4, and others from 1 to 5, where a higher value did not always indicate a higher degree. I recoded variables so that the indicators that involved scales were standardized to take values between 0 and 1, where 1 is the highest degree of a given characteristic. This makes it no longer necessary to use logarithmic transformations.

Although for many factor variables it made sense to standardize them (as above) while keeping them incremental, many were better represented when split into one or more binaries. One obvious example is the variable indicating the marital status of the respondent; I divided it into the `married, `living_together_as_married`, `separated`, `divorced`, and `widowed` binaries. Another example is the variable originally indicating 1 for a religious person, 2 for not a religious person, and 3 for a convinced atheist. I assumed "not a religious person" to become my baseline and added two dummies: "religious" and "atheist."

I kept select number of variables with their original values. One of them was `number_of_children`, which corresponds with a tangible quantity.

Overall, my binary variables are chosen in such a way that **my baseline is a non-religious fully employed single female at the age of 13-24 who attends religious services less than once a week.**

a. Preparing the dataset

The libraries used in the analysis are the following:

```
library("rio")
library("pscl")
library("car")
library("lmtest")
library("tidyverse")
library("moderndive")
library("alpaca")
```

```
library("ggpubr")
library("ggplot2")
library("psych")
```

b. Exploring the dataset

Summary statistics:

```
describe(mydata)
```

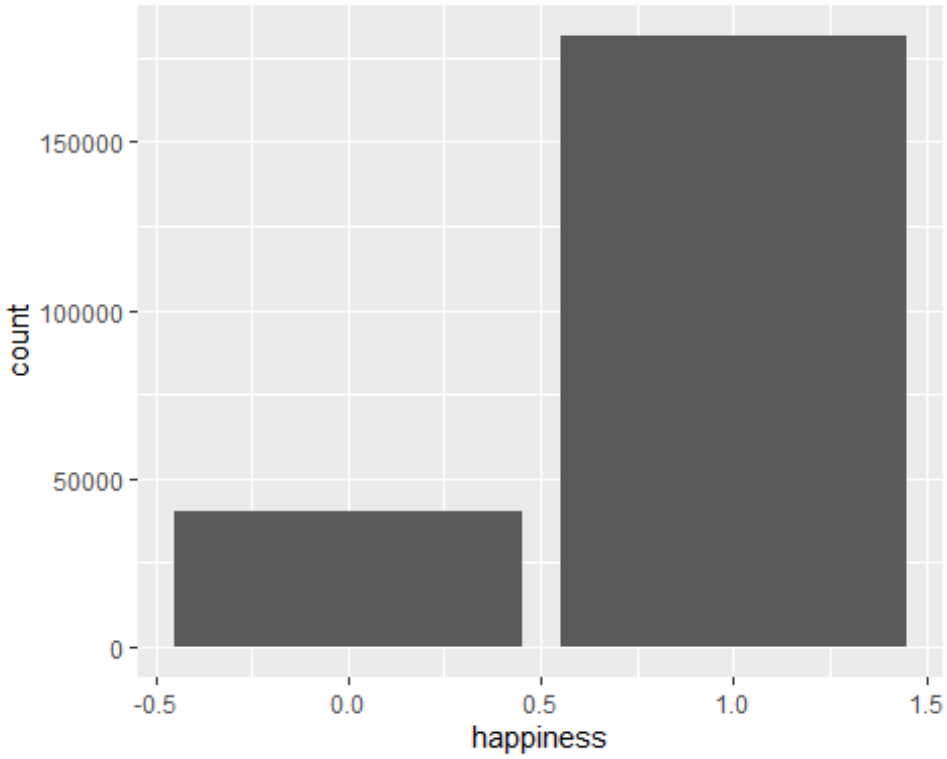
```
##                            vars      n    mean     sd  median trimmed
mad
## year                          1 222047 2004.00   7.29 2006.00 2004.48
8.90
## country_name                  2 222047  463.67 260.63  466.00  469.05
351.38
## male                          3 222047    0.48   0.50    0.00    0.48
0.00
## age_13_24                     4 222047    0.17   0.37    0.00    0.09
0.00
## age_25_40                     5 222047    0.37   0.48    0.00    0.34
0.00
## age_41_60                     6 222047    0.32   0.46    0.00    0.27
0.00
## age_61_80                     7 222047    0.13   0.34    0.00    0.04
0.00
## age_81_more                   8 222047    0.01   0.10    0.00    0.00
0.00
## married                       9 222047    0.58   0.49    1.00    0.60
0.00
## living_together_as_maried    10 222047    0.07   0.25    0.00    0.00
0.00
## divorced                     11 222047    0.03   0.18    0.00    0.00
0.00
## separated                    12 222047    0.02   0.14    0.00    0.00
0.00
## widowed                      13 222047    0.06   0.24    0.00    0.00
0.00
## number_of_children           14 222047    1.93   1.82    2.00    1.68
1.48
## part_time                    15 222047    0.08   0.27    0.00    0.00
0.00
## self_employed                16 222047    0.12   0.33    0.00    0.03
0.00
## unemployed                   17 222047    0.10   0.29    0.00    0.00
0.00
## retired                      18 222047    0.12   0.33    0.00    0.03
0.00
## housewife                    19 222047    0.15   0.36    0.00    0.07
0.00
## student                      20 222047    0.07   0.26    0.00    0.00
```

```
0.00
## financial_satisfaction      21 222047    0.52   0.29    0.56    0.53
0.33
## scale_incomes               22 222047    0.41   0.26    0.44    0.40
0.33
## health                      23 222047    0.70   0.26    0.80    0.72
0.30
## religious                   24 222047    0.72   0.45    1.00    0.77
0.00
## atheist                     25 222047    0.04   0.20    0.00    0.00
0.00
## god_important               26 222047    0.75   0.33    1.00    0.81
0.00
## attend_church_often         27 222047    0.34   0.47    0.00    0.30
0.00
## thinks_about_purpose_life   28 222047    0.67   0.32    0.50    0.70
0.74
## trust                       29 222047    0.26   0.44    0.00    0.20
0.00
## freedom_choice_control      30 222047    0.66   0.26    0.67    0.68
0.33
## happiness                   31 222047    0.82   0.39    1.00    0.90
0.00
## life_satisfaction           32 222047    0.68   0.46    1.00    0.73
0.00
##                            min  max range  skew kurtosis   se
## year                      1981 2016    35 -0.56    -0.21 0.02
## country_name                 8  914   906 -0.13    -1.23 0.55
## male                         0    1     1  0.07    -2.00 0.00
## age_13_24                    0    1     1  1.77     1.14 0.00
## age_25_40                    0    1     1  0.53    -1.72 0.00
## age_41_60                    0    1     1  0.79    -1.37 0.00
## age_61_80                    0    1     1  2.16     2.65 0.00
## age_81_more                  0    1     1  9.76    93.26 0.00
## married                      0    1     1 -0.33    -1.89 0.00
## living_together_as_maried    0    1     1  3.50    10.23 0.00
## divorced                     0    1     1  5.08    23.78 0.00
## separated                    0    1     1  7.10    48.46 0.00
## widowed                      0    1     1  3.74    12.00 0.00
## number_of_children           0    8     8  1.09     1.15 0.00
## part_time                    0    1     1  3.11     7.70 0.00
## self_employed                0    1     1  2.32     3.37 0.00
## unemployed                   0    1     1  2.74     5.49 0.00
## retired                      0    1     1  2.28     3.21 0.00
## housewife                    0    1     1  1.91     1.65 0.00
## student                      0    1     1  3.32     9.05 0.00
## financial_satisfaction       0    1     1 -0.18    -0.80 0.00
## scale_incomes                0    1     1  0.28    -0.61 0.00
## health                       0    1     1 -0.58    -0.96 0.00
## religious                    0    1     1 -0.97    -1.07 0.00
```

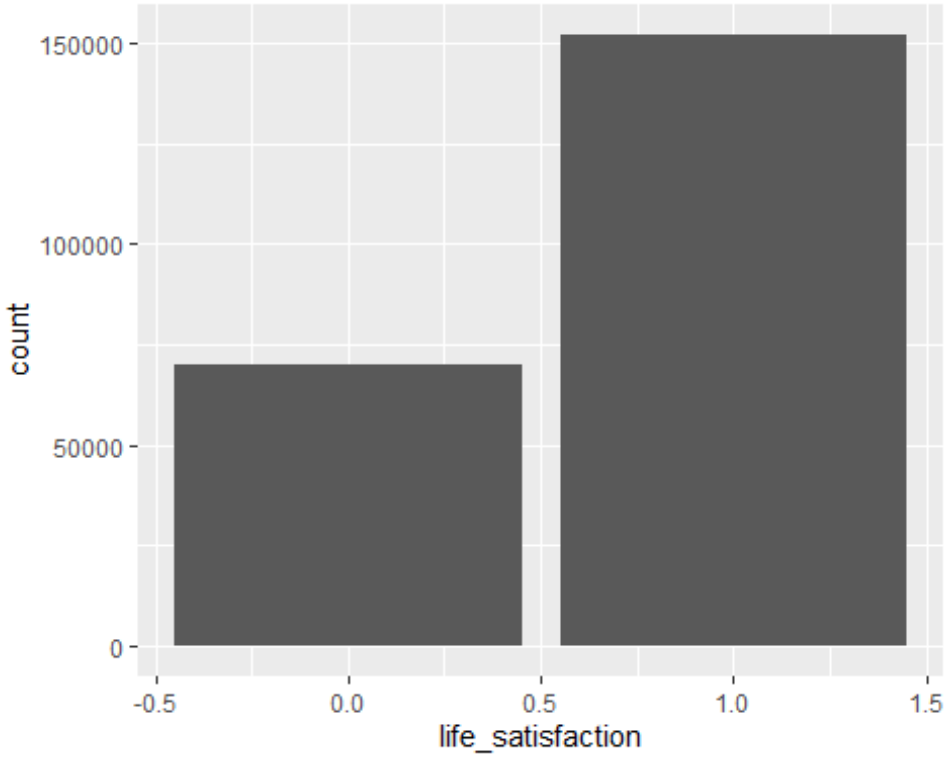
```
## atheist                      0    1     1  4.53    18.54 0.00
## god_important                0    1     1 -1.13    -0.06 0.00
## attend_church_often          0    1     1  0.66    -1.56 0.00
## thinks_about_purpose_life    0    1     1 -0.28    -1.25 0.00
## trust                        0    1     1  1.09    -0.82 0.00
## freedom_choice_control       0    1     1 -0.59    -0.28 0.00
## happiness                    0    1     1 -1.65     0.71 0.00
## life_satisfaction            0    1     1 -0.79    -1.37 0.00
```

How many people are happy, and how many are satisfied with their lives?

```
ggplot(mydata, aes(x = happiness)) +
  geom_bar()
```




```
ggplot(mydata, aes(x = life_satisfaction)) +
  geom_bar()
```



The data look good. There are a bit more women, although the gender ratio is close to 1:1. Half of the people are married and have two children on average. People in the sample feel healthy and are moderately satisfied with their financial situation. They are religious and believe that God is important, but a minority attend religious services often. It is comforting to see that people feel generally happy and satisfied with their lives, although they seem overall happier than satisfied.

c. Expanding the dataset

Create non-linear terms based on the variables whose scale was 1-10:

```
mydata[, "scale_incomes_sq"] = (mydata$scale_incomes)^2
mydata[, "god_important_sq"] = (mydata$god_important)^2
mydata[, "thinks_about_purpose_life_sq"] =
(mydata$thinks_about_purpose_life)^2
```

4) Hypothesis

I hypothesize that there is a strong correlation between life satisfaction and the feeling of happiness, although these two may be influenced by particular factors differently. For example, it could be the case that being religious makes you less happy due to the additional rules you need to obey but brings you a higher level of life satisfaction instead.

Among the factors I expect to be strongly associated with higher happiness is being higher in the scale of income, being employed or retired, and being married. I also expect higher age, the feeling of freedom, the number of children, and trusting people to be weakly

correlated with higher happiness. My expectations are similar with regards to life satisfaction, although perhaps the number of children and religiosity play a greater role.

As far as factors correlated with lower happiness are concerned, I would expect them to include being divorced, separated, and widowed, being unemployed. I would guess that there is also a weak association between lower happiness and thinking about the purpose of life frequently (as melancholic people appear sadder on average), and being a woman (due to various adverse forms of behavior experience from men). I expect similar outcomes for life satisfaction, although perhaps thinking about the purpose of life ultimately leads people to the life of higher satisfaction.

5) Modelling and Empirical Exploration

a. The Correlation between Happiness and Life Satisfaction

```
cor(happiness, life_satisfaction)

## [1] 0.3681492
```

The correlation between `happiness` and `life_satisfaction` is weak. This is a very surprising outcome. It would seem that these two ideas are similar.

Perhaps we would like to investigate the wording of the WVS questions. The question about happiness was as follows: “Taking all things together, would you say you are:” and its four possible answers, “Very happy,” “Quite happy,” “Not very happy,” and “Not at all happy.” I assigned the value of 1 to the first two answers, and the value of 0 to the other two answers. The life satisfaction question was the following: “All things considered, how satisfied are you with your life as a whole these days? Please use this card to help with your answer.” The possible answers were numbers on the scale from 1 to 10, where 10 means “satisfied,” and 1 means “dissatisfied.” I assigned the value of 1 to answers between 6 and 10, and the value of 0 to the answers between 1 and 5.

There are different interpretations possible. Perhaps my transformation into a binary value did not capture the possibility that people are biased toward giving higher values on a 10-point scale so that people who are dissatisfied with their lives and are unhappy would answer “Not very happy” or “Not at all happy” in the first question but would give a number like 6 in the second question. However, this would be inconsistent with the bar plots.

Another interpretation, which I assume in this analysis, is that happiness and life satisfaction are two distinct things. Perhaps happiness refers mostly to emotions while satisfaction to an unemotional outlook at one’s situation. Happiness could be more influenced by things that matter more subjectively and subconsciously, while satisfaction could be more a result of a more rational and objective reflection. This would be an interesting topic for a future investigation.

b. The Logit Model

Logit Model Assumptions

Logistic regression does not assume a linear relationship between response and explanatory variables, does not assume a normal distribution of error terms, and does not assume homoskedasticity.

However, binary logistic regression assumes:

a. A binary response variable. This is the case with both `happiness` and `life_satisfaction`.

b. Observations independent of each other. This should be true for the World Values Survey within a country in a given year, although may be violated when combining data from all countries and years. I will address this issue in my analysis.

c. No perfect multicollinearity among the independent variables. I will eliminate the variables causing this problem.

d. Linearity of independent variables and log odds. This means that I should not have continuous independent variables. The only variable that potentially could have caused a problem was the variable `age`, which I divided into `age_13_24`, `age_25_40`, `age_41_60`, `age_61_80`, and `age_81_more`. I chose the age between 13 and 24 (inclusive) as my baseline.

e. A large sample size. With hundreds of thousands of observations, this is not an issue in my study.

Proposing the Logit Model

By including all the explanatory variables from my dataset, I propose the following model for `happiness`:

```
h_model_full_formula <- happiness ~ male + age_25_40 + age_41_60 + age_61_80
+ age_81_more +
  married + living_together_as_maried + divorced + separated +
  widowed + number_of_children + financial_satisfaction + scale_incomes +
scale_incomes_sq +
  part_time + self_employed + retired + housewife + student +
  unemployed + trust + freedom_choice_control + health +
thinks_about_purpose_life +
  thinks_about_purpose_life_sq + god_important + god_important_sq +
  religious + atheist + attend_church_often
```

and the following model for `life satisfaction`:

```
s_model_full_formula <- life_satisfaction ~ male + age_25_40 + age_41_60 +
age_61_80 + age_81_more +
  married + living_together_as_maried + divorced + separated +
  widowed + number_of_children + financial_satisfaction + scale_incomes +
scale_incomes_sq +
  part_time + self_employed + retired + housewife + student +
```

```
  unemployed + trust + freedom_choice_control + health +
thinks_about_purpose_life +
  thinks_about_purpose_life_sq + god_important + god_important_sq +
  religious + atheist + attend_church_often
```

I fit both, along with their null versions, into a general linear model function in R:

```
h_model_full = glm(formula <- h_model_full_formula, family = binomial(link =
"logit"))
h_model_null = glm(formula = happiness ~ 1, family = binomial(link =
"logit"))

s_model_full = glm(formula <- s_model_full_formula, family = binomial(link =
"logit"))
s_model_null = glm(formula = life_satisfaction ~ 1, family = binomial(link =
"logit"))
```

Refining the Logit Model

a. Stepwise Regression

Now, by looking at regression output summaries, I could remove each statistically insignificant variable in my model, but I prefer to use the backward stepwise selection algorithm to do this for me.

I use the following code:

```
h_model_backward <- step(h_model_full, scope = list(lower = h_model_null),
direction = "backward", trace= F)
h_model_backward_formula <- formula(h_model_backward)

s_model_backward <- step(s_model_full, scope = list(lower = s_model_null),
direction = "backward", trace= F)
s_model_backward_formula <- formula(s_model_backward)
```

And I obtain the following results:

```
## happiness ~ male + age_25_40 + age_41_60 + age_61_80 + age_81_more +
##     married + living_together_as_maried + divorced + separated +
##     widowed + financial_satisfaction + scale_incomes + scale_incomes_sq +
##     part_time + self_employed + retired + housewife + student +
##     unemployed + trust + freedom_choice_control + health +
thinks_about_purpose_life +
##     thinks_about_purpose_life_sq + god_important + god_important_sq +
##     religious + attend_church_often

## life_satisfaction ~ male + age_25_40 + age_41_60 + age_61_80 +
##     age_81_more + married + living_together_as_maried + divorced +
##     separated + widowed + number_of_children + financial_satisfaction +
##     scale_incomes_sq + self_employed + housewife + unemployed +
##     trust + freedom_choice_control + health + thinks_about_purpose_life_sq
```

```
+
##     god_important + god_important_sq + religious + attend_church_often
```

The algorithm excluded `number of children` and `atheist` in the case of `happiness`, and `scale_incomes`, `retired`, `student`, `thinks_about_purpose_life`, and `atheist` in the case of life satisfaction.

I fit both into a general linear model function:

```
h_model_backward <- glm(formula <- h_model_backward_formula, family =
binomial(link = "logit"))
s_model_backward <- glm(formula <- s_model_backward_formula, family =
binomial(link = "logit"))
```

b. Goodsness of Fit Selection

I now compare the AIC scores for full models and for stepwise-selected models:

```
AIC(h_model_full)

## [1] 172602.6

AIC(h_model_backward)

## [1] 172599.9

AIC(s_model_full)

## [1] 206144.7

AIC(s_model_backward)

## [1] 206134.9
```

I conclude that the stepwise model is a better fit in both cases. The differences are not large, suggesting that the full models were already good, and a large majority of the exploratory variables I selected are relevant for the study.

c. Eliminating Perfect Multicollinearity

Now, I check for multicollinearity for both models by using the `vif()` command:

```
vif(h_model_backward)

##                          male                      age_25_40
##                      1.271091                       2.977894
##                     age_41_60                      age_61_80
##                      3.425506                       3.365530
##                   age_81_more                        married
##                      1.257367                       2.312030
##    living_together_as_maried                       divorced
##                      1.270917                       1.302432
##                     separated                        widowed
##                      1.144094                       1.837524
```

```
##       financial_satisfaction                  scale_incomes
##                     1.154231                       9.188255
##             scale_incomes_sq                     part_time
##                     8.902867                       1.152243
##                self_employed                        retired
##                     1.250293                       2.163266
##                    housewife                         student
##                     1.577364                       1.449678
##                   unemployed                            trust
##                     1.311349                       1.026373
##       freedom_choice_control                          health
##                     1.065055                       1.158263
##    thinks_about_purpose_life thinks_about_purpose_life_sq
##                    23.004220                      23.089869
##                god_important              god_important_sq
##                    20.344898                      19.951084
##                    religious           attend_church_often
##                     1.473267                       1.218507

vif(s_model_backward)

##                         male                      age_25_40
##                     1.262700                       2.610550
##                    age_41_60                      age_61_80
##                     3.093168                       2.405256
##                  age_81_more                        married
##                     1.147180                       2.497054
##    living_together_as_maried                       divorced
##                     1.301939                       1.261774
##                    separated                        widowed
##                     1.129624                       1.715778
##           number_of_children       financial_satisfaction
##                     1.587235                       1.059772
##             scale_incomes_sq                 self_employed
##                     1.087085                       1.109485
##                    housewife                    unemployed
##                     1.386791                       1.102200
##                        trust       freedom_choice_control
##                     1.027988                       1.018773
##                       health thinks_about_purpose_life_sq
##                     1.136768                       1.065389
##                god_important              god_important_sq
##                    20.102007                      19.708239
##                    religious           attend_church_often
##                     1.470332                       1.231114
```

High VIF's are generated because of the inclusion of the non-linear terms. I address this issue for each model separately, by eliminating one variable with the highest VIF score at a time.

The formula updating process results in the elimination of `thinks_about_purpose_life`, `god_important`, and `scale_incomes` (the latter was not higher than 10 exactly, but close to it), after which I obtain the following `happiness` model without evidence for perfect multicollinearity:

```
h_model_backward_formula
```

```
## happiness ~ male + age_25_40 + age_41_60 + age_61_80 + age_81_more +
##     married + living_together_as_maried + divorced + separated +
##     widowed + financial_satisfaction + scale_incomes_sq + part_time +
##     self_employed + retired + housewife + student + unemployed +
##     trust + freedom_choice_control + health + thinks_about_purpose_life_sq
+
##     god_important_sq + religious + attend_church_often
```

```
vif(h_model_backward)
```

```
##                          male                    age_25_40
##                      1.269809                     2.980105
##                    age_41_60                    age_61_80
##                      3.428477                     3.368512
##                   age_81_more                      married
##                      1.257534                     2.312328
##    living_together_as_maried                     divorced
##                      1.270937                     1.302614
##                     separated                      widowed
##                      1.143684                     1.837811
##       financial_satisfaction             scale_incomes_sq
##                      1.132771                     1.112635
##                     part_time                self_employed
##                      1.151612                     1.249730
##                       retired                    housewife
##                      2.162747                     1.573794
##                       student                   unemployed
##                      1.450058                     1.305750
##                         trust       freedom_choice_control
##                      1.026298                     1.064647
##                        health thinks_about_purpose_life_sq
##                      1.157355                     1.059423
##             god_important_sq                     religious
##                      1.546030                     1.386549
##          attend_church_often
##                      1.209675
```

I repeat the process for `life_satisfaction` and remove only `god_important`. Note that the other redundant variables resulting from the inclusion of non-linear terms were already removed with the AIC selection process. I obtain the following formula and VIF scores:

```
s_model_backward_formula
```

```
## life_satisfaction ~ male + age_25_40 + age_41_60 + age_61_80 +
##     age_81_more + married + living_together_as_maried + divorced +
##     separated + widowed + number_of_children + financial_satisfaction +
##     scale_incomes_sq + self_employed + housewife + unemployed +
##     trust + freedom_choice_control + health + thinks_about_purpose_life_sq
+
##     god_important_sq + religious + attend_church_often

vif(s_model_backward)

##                          male                     age_25_40
##                      1.261852                      2.610649
##                    age_41_60                     age_61_80
##                      3.093284                      2.405290
##                   age_81_more                       married
##                      1.147213                      2.496726
##    living_together_as_maried                      divorced
##                      1.301777                      1.261781
##                     separated                       widowed
##                      1.129594                      1.715779
##           number_of_children       financial_satisfaction
##                      1.586718                      1.058073
##             scale_incomes_sq                 self_employed
##                      1.086744                      1.109331
##                     housewife                    unemployed
##                      1.386181                      1.102162
##                         trust       freedom_choice_control
##                      1.028048                      1.018162
##                        health thinks_about_purpose_life_sq
##                      1.136781                      1.062621
##             god_important_sq                     religious
##                      1.564800                      1.382178
##          attend_church_often
##                      1.225075
```

d. Assessing the goodness of fit:

I look at the goodness of fit using pseudo-R squared:

```
pR2(h_model_backward)

## fitting null model for pseudo-r2

##           llh       llhNull            G2      McFadden          r2ML
## -8.636760e+04 -1.053937e+05  3.805212e+04  1.805238e-01  1.574899e-01
##          r2CU
##  2.569230e-01

pR2(s_model_backward)

## fitting null model for pseudo-r2
```

```
##           llh       llhNull            G2      McFadden          r2ML
## -1.030449e+05 -1.384094e+05  7.072897e+04  2.555064e-01  2.727838e-01
##          r2CU
##  3.828337e-01
```

The McFadden pseudo-R squared scores of 0.18 for the happiness model and 0.25 for the life satisfaction model are very low. However, it is important to remember that there are potentially thousands of factors that influence people's feelings of happiness and of life satisfaction, and the variables included in the dataset are just some of them, even if plausibly the most important ones.

e. Controlling for fixed effects

There is a relatively new library called `alpaca`, which was released in 2016. It includes `feglm()` and `biasCorr()` for logit models with time- and entity-specific fixed effects. This is a good choice for my international longitudinal dataset.

I fit it use it as follows:

```
# Fit the happiness model
h_model_fe <- feglm(happiness ~ male + age_25_40 + age_41_60 + age_61_80 +
age_81_more +
    married + living_together_as_maried + divorced + separated +
    widowed + financial_satisfaction + scale_incomes_sq + part_time +
    self_employed + retired + housewife + student + unemployed +
    trust + freedom_choice_control + health + thinks_about_purpose_life_sq +
    god_important_sq + religious + attend_church_often | country_name + year,
mydata, binomial("logit"))
# Bias correction routine
h_model_fe <- biasCorr(h_model_fe)

# Fit the life_satisfaction model
s_model_fe <- feglm(life_satisfaction ~ male + age_25_40 + age_41_60 +
age_61_80 +
    age_81_more + married + living_together_as_maried + divorced +
    separated + widowed + number_of_children + financial_satisfaction +
    scale_incomes_sq + self_employed + housewife + unemployed +
    trust + freedom_choice_control + health + thinks_about_purpose_life_sq +
    god_important_sq + religious + attend_church_often |
                      country_name + year, mydata, binomial("logit"))
# Bias correction routine
s_model_fe <- biasCorr(s_model_fe)
```

I obtain the following regression summaries:

```
# with the "sandwitch" mode, we obtain heteroskdasticity-robust standard
errors:
summary(h_model_fe, "sandwich")

## binomial - logit link
##
```

```
## happiness ~ male + age_25_40 + age_41_60 + age_61_80 + age_81_more +
##     married + living_together_as_maried + divorced + separated +
##     widowed + financial_satisfaction + scale_incomes_sq + part_time +
##     self_employed + retired + housewife + student + unemployed +
##     trust + freedom_choice_control + health + thinks_about_purpose_life_sq
+
##     god_important_sq + religious + attend_church_often | country_name +
##     year
##
## Estimates:
##                               Estimate Std. error z value Pr(> |z|)
## male                          -0.14193    0.01453  -9.768   < 2e-16 ***
## age_25_40                     -0.19584    0.02298  -8.521   < 2e-16 ***
## age_41_60                     -0.24040    0.02567  -9.366   < 2e-16 ***
## age_61_80                     -0.13135    0.03372  -3.895  9.82e-05 ***
## age_81_more                   -0.07238    0.07025  -1.030   0.30283
## married                        0.53380    0.01994  26.766   < 2e-16 ***
## living_together_as_maried      0.27013    0.03156   8.558   < 2e-16 ***
## divorced                      -0.23184    0.03580  -6.475  9.47e-11 ***
## separated                     -0.32102    0.04435  -7.238  4.56e-13 ***
## widowed                       -0.17009    0.03194  -5.326  1.00e-07 ***
## financial_satisfaction         1.76080    0.02751  64.005   < 2e-16 ***
## scale_incomes_sq               0.67035    0.03526  19.012   < 2e-16 ***
## part_time                     -0.05865    0.02626  -2.234   0.02550 *
## self_employed                 -0.10678    0.02277  -4.691  2.73e-06 ***
## retired                       -0.02159    0.02688  -0.803   0.42200
## housewife                      0.07057    0.02336   3.021   0.00252 **
## student                        0.15248    0.03253   4.687  2.77e-06 ***
## unemployed                    -0.29926    0.02245 -13.328   < 2e-16 ***
## trust                          0.25114    0.01626  15.448   < 2e-16 ***
## freedom_choice_control         0.98499    0.02504  39.332   < 2e-16 ***
## health                         2.50713    0.02737  91.611   < 2e-16 ***
## thinks_about_purpose_life_sq -0.04948    0.01617  -3.059   0.00222 **
## god_important_sq               0.14231    0.02424   5.871  4.34e-09 ***
## religious                      0.14290    0.01737   8.225   < 2e-16 ***
## attend_church_often            0.07496    0.01640   4.571  4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## residual deviance= 161628.83,
## null deviance= 210787.32,
## n= 222047, l= [97, 26]
##
## Number of Fisher Scoring Iterations: 6

summary(s_model_fe, "sandwich")

## binomial - logit link
##
## life_satisfaction ~ male + age_25_40 + age_41_60 + age_61_80 +
```

```
##     age_81_more + married + living_together_as_maried + divorced +
##     separated + widowed + number_of_children + financial_satisfaction +
##     scale_incomes_sq + self_employed + housewife + unemployed +
##     trust + freedom_choice_control + health + thinks_about_purpose_life_sq
+
##     god_important_sq + religious + attend_church_often | country_name +
##     year
##
## Estimates:
##                              Estimate Std. error z value Pr(> |z|)
## male                         -0.09385    0.01283  -7.314  2.60e-13 ***
## age_25_40                    -0.16210    0.01909  -8.493   < 2e-16 ***
## age_41_60                    -0.21286    0.02207  -9.643   < 2e-16 ***
## age_61_80                    -0.11149    0.02750  -4.054  5.04e-05 ***
## age_81_more                  -0.21532    0.06439  -3.344  0.000826 ***
## married                       0.21356    0.01876  11.382   < 2e-16 ***
## living_together_as_maried     0.15431    0.02896   5.328  9.93e-08 ***
## divorced                     -0.13606    0.03495  -3.893  9.91e-05 ***
## separated                    -0.17613    0.04486  -3.927  8.62e-05 ***
## widowed                      -0.04045    0.03157  -1.281  0.200184
## number_of_children            0.01348    0.00401   3.361  0.000776 ***
## financial_satisfaction        3.26157    0.02565 127.178   < 2e-16 ***
## scale_incomes_sq              0.76331    0.03024  25.246   < 2e-16 ***
## self_employed                -0.10987    0.01863  -5.898  3.67e-09 ***
## housewife                     0.02234    0.01896   1.178  0.238834
## unemployed                   -0.22962    0.01951 -11.769   < 2e-16 ***
## trust                         0.18742    0.01394  13.444   < 2e-16 ***
## freedom_choice_control        1.77940    0.02366  75.214   < 2e-16 ***
## health                        1.35979    0.02439  55.745   < 2e-16 ***
## thinks_about_purpose_life_sq -0.12687    0.01444  -8.787   < 2e-16 ***
## god_important_sq              0.25116    0.02187  11.484   < 2e-16 ***
## religious                     0.06133    0.01559   3.934  8.35e-05 ***
## attend_church_often           0.03976    0.01447   2.747  0.006017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## residual deviance= 195026.64,
## null deviance= 276818.77,
## n= 222047, l= [97, 26]
##
## Number of Fisher Scoring Iterations: 5
```

6) Limitations and Challenges

There are important limitations and challenges to my results.

A serious problem with my dataset is that I used listwise deletion for problematic observations (such as missing observations or answers like “I don’t know”), as this method is known for producing biased estimates. A multiple imputation procedure for binary logistic regression would have been a significantly better alternative.

I prioritized the number of non-problematic observations and selected questions with the most frequently available answers. I did this by going through the WVS codebook. However, if I had more time, I could have relaxed my expectations regarding the number of observations, and I could have left it to the AIC/BIC algorithm to choose the most relevant variables.

The order in which I was refining my model was debatable. I introduced fixed effects of years and countries as a final step rather doing that in my initial model. However, this is due to the limited availability of beginner-friendly tools for fixed effects in logit. In fact, the models obtained with the `glmfe()` functions are not yet testable with tests such as the `AIC(), pR2()` or `bptest().`

Moreover, my model was not tested for its predictive power. Perhaps there is much work yet to be done with improving the specification. While I added three squared variables to my initial model, I could have tried different transformations of many more variables, expanding the list with interaction terms and higher-degree polynomials. Therefore, my conclusions mostly pertain to the significance of variables and relative magnitudes of their effects.

Although happiness and life satisfaction entered my regressions as dependent variables and I describe my results mostly within this framework, my analysis does not suggest any direction of causality between two variables.

The listed potential issues are the problems I knew about but could not address to due time constrains. However, as I do not have much experience in working with categorical data, there are potentially many more errors that could have entered my analysis. For example, the process of cleaning data and recoding variables could have been imperfect.

7) Discussion

Perhaps the largest surprise of my results is that there is only a weak correlation between happiness and life satisfaction. As I discussed in the results section, it could be because these two concepts are truly distinct and perhaps represent a “heart-reason” divergence. Regardless of the weak correlation, there is no estimate that would change its sign from model to model. The only differences between the outcomes pertain to the magnitudes or, in few cases, the significance levels of the results.

Another interesting result is that men are significantly less happy and less satisfied with their lives than women, even after controlling for many other factors. This is contrary to my hypothesis. It may have to do either with genes and metabolism or with cultural norms. Perhaps men’s well-known lower risk aversion leads them to doing things that have seriously negative consequences, and the individuals who face these consequences drive the coefficients down. Alternatively, male hormones make men feel less happy and optimistic. This would be a fascinating topic for further investigation.

Metabolism and homeostasis impact our spirits, not just our bodies. Health is the largest predictor of happiness and the third largest predictor of life satisfaction (yielding to financial satisfaction and the feeling of freedom). A part of this effect may be a comorbidity of physical health deficiency with mental health diseases. However, it is also important to

remember that this is self-assessed health, so general pessimism may influence both answers. Regardless, a takeaway is that there are few things that can compensate health.

The relationship between the dependent variables of interest and age is negative. Interestingly, in regressions without fixed effects (not reported), it seemed that age was positively correlated with happiness and life satisfaction. Perhaps that effect was driven by richer and older societies. The final outcome could be interpreted as a reflection of the possibility that people care a lot about aspects of being young, such as appearance or having more energy.

Married people are the happiest and the most satisfied with their lives, holding the number of children constant. The effect is particularly high for happiness. Those who "live as married" are not far behind in both categories. As expected, being divorced, separated, and widowed makes you less happy or satisfied with your life even compared to being single. Interestingly, being separated, rather than divorced, makes you the saddest. Overall, with the strongest effect of all, marriage seems a worthwhile risk. When it comes to children, they do bring satisfaction, but surprisingly little of it. Although many moms and dads claim to be the happiest people on Earth, the AIC algorithm challenged their declarations to some extent by deleting the variable.

If you do not work, you can as happy as a student, to whom college brings as many positive feelings as religion to the religious people; as sad as an unemployed, for whom their situation is even worse than for those separated; or somewhere in-between, as the retired who are not statistically significantly different than the fully employed. Housewives (or househusbands) are generally happier and more satisfied with their lives than the fully employed, although without enough statistical evidence for the latter feeling.

When looking at financial satisfaction and the scale of incomes, it seems that money is everything. The former beats every other variable, except for health, by a lot in both models. Or, people who generally feel better about everything have more energy to earn more. Since the scale of incomes is represented by a squared term, I could infer that its positive coefficient tells us that being average is the saddest and least satisfying position in society. If you have not made it to the top, you may be better off by giving away everything like St. Francis.

Speaking of religion, being religious and going to church often is associated with higher happiness and life satisfaction, although the effects are not enormous. Reflecting on the sense of your life makes sense when you do it in moderation; too little or too much contemplation seems to make you less happy and less satisfied with your life. Similarly, the importance of God has a quadratic relationship with happiness and life satisfaction, which means that it is worth being decided in the spiritual realm. This reminds me of the biblical quote: "So, because you are lukewarm, neither hot nor cold, I will spit you out of my mouth." (Revelation 3, 16).

Trusting in God makes you feel good, and so does trusting in other people. Or, feeling good makes you trust more in both. In any case, the association is moderately strong for happiness and life satisfaction, with emphasis on the former.

Last but not least, the feeling of freedom of choice and control over your life is one of the most important things for happiness and life satisfaction. According to my model, freedom brings more satisfaction than marriage, children, health, or religion. As I am writing this from home abroad, this makes me miss the Land of Freedom even more.

8) Conclusion

In conclusion, the recipe for happiness and life satisfaction is hard to obtain using scientific methods. However, the more careful the statistical analysis, the better chance that we understand what makes people across the globe feel better. My project is just the first step in this process, and there are still more things I did not consider than those I took into account. From data collection, through data cleaning, to data analysis, my outcomes rely and numerous assumptions. However, I hope that the imperfections of my study and the preliminary results can serve as an inspiration for future investigation. As of now, the WVS data suggest that the *average* key to happiness and life satisfaction is being a healthy married young woman with a few children, who is a pious and religious student, trusts other people, and enjoys has lots of freedom as well as good economic conditions. Certainly, there is at least one omitted variable: `took_stat_109`.

REFERENCES

WORLD VALUES SURVEY 1981-2014 LONGITUDINAL AGGREGATE V.20150418. WORLD VALUES SURVEY ASSOCIATION (WWW.WORLDVALUESSURVEY.ORG). AGGREGATE FILE PRODUCER: JDSYSTEMS, MADRID SPAIN.

https://www.statisticssolutions.com/assumptions-of-logistic-regression/

https://www.statisticssolutions.com/wp-content/uploads/wp-post-to-pdf-enhanced-cache/1/assumptions-of-logistic-regression.pdf