What Makes Us Happy? Unveiling the Impact of Social, Economic, and Personal Factors*

Shuheng (Jack) Zhou

November 26, 2024

This study employs a logistic regression model to analyze the impact of various social, economic, and personal factors on happiness. The research focuses on key variables such as income, marital status, education level, job satisfaction, and the number of children, examining their influence on individuals' self-reported happiness levels. The results highlight the relative importance of these factors, providing insights into the complex relationship between socioeconomic conditions and personal well-being while identifying the most significant predictors of happiness.

1 Introduction

Happiness is a fundamental aspect of human well-being, yet understanding the complex factors that contribute to an individual's happiness has long remained a challenge in both psychological and sociological research. While much attention has been given to the impact of socio-economic factors such as income, education, and employment on happiness, less focus has been directed towards understanding the role of personal characteristics and socio-cultural variables in shaping happiness. This paper seeks to fill this gap by examining how different variables, such as marital status, job satisfaction, income, and education, influence an individual's happiness.

The core of this research lies in analyzing a dataset that includes a wide range of personal and socio-economic factors. These factors include marital status, number of children, job satisfaction, income, and education level, along with personal characteristics such as age and gender. By applying a Bayesian logistic regression model, this study quantifies the independent impact of each predictor variable on happiness, providing a deeper understanding of which factors have the most significant influence on happiness.

*The GitHub Repository containing all data, R code, and other files used in this project is located here:https://github.com/Shuhengzhou03/Factors-Influencing-Happiness.git

Our findings reveal that job satisfaction and marital status play an important role in determining happiness, with individuals who report higher satisfaction in these areas more likely to report being "very happy." Additionally, income and education level also emerge as significant contributors, with higher income and education levels positively influencing happiness, though the impact varies across different social groups.

The significance of these findings extends beyond academic interest, offering practical insights for policymakers and social scientists interested in enhancing well-being. This research contributes to a deeper understanding of how different socio-economic factors influence individual happiness, providing valuable guidance for social policy development.

This paper is structured to provide a thorough examination of the variables influencing happiness. Following the introduction in Section 1, Section 2 outlines the dataset used in the study, detailing the sources of the data and the selection of key variables for analysis. Section 3 introduces the Bayesian logistic regression model employed to analyze the relationships between the predictors and happiness, providing a statistical framework for understanding how various factors independently contribute to happiness. Section 4 presents the findings from the Bayesian model, explaining how job satisfaction, marital status, and other factors impact the likelihood of being "very happy." Section 5 explores the implications of these findings, discusses the limitations of the study, and suggests avenues for future research to better understand the factors influencing happiness. Section A provides detailed plots of posterior predictive checks and model diagnostics, ensuring the robustness of our findings.

1.1 Estimand

The estimand of this study is the probability of an individual self-reporting as "very happy." Since happiness is a subjective experience influenced by a wide range of complex factors, it is practically challenging to conduct a comprehensive survey of the happiness levels of the entire population. Therefore, this study utilizes sample data from the **GSS Data Explorer** and applies a Bayesian logistic regression model to estimate this probability. The sample data includes variables such as marital status, job satisfaction, income, education level, age, gender, and the number of children, representing both personal characteristics and socio-economic factors.

Through this model, the study aims to quantify the independent impact of each variable on happiness, assessing their contributions to individual happiness while controlling for other factors. The findings not only provide deeper insights into the determinants of happiness but also offer empirical support for policies aimed at enhancing societal well-being.

2 Data

2.1 Overview

2.1.1 Data Source and Analysis Tools

The data used in this study was sourced from the GSS Data Explorer ("GSS Data Explorer" 2023), an online platform maintained by NORC that provides a wide range of resources from the General Social Survey (GSS). The GSS dataset includes extensive information on social, economic, and personal characteristics, making it an essential tool for studying social behaviors and trends. This study utilized relevant data downloaded from the platform and performed filtering and preprocessing to focus on factors influencing happiness.

The entire process of data handling, analysis, modeling, and visualization was conducted using the R programming language (R Core Team 2023). The following R packages were instrumental in this study:

- **`tidyverse`** (Wickham 2023d): Provided a comprehensive set of tools for data manipulation and visualization, significantly simplifying the workflow.
- **`palmerpenguins`** (Allison Horst 2020): Offered example datasets and tools, aiding in the quick testing of analysis code.
- **`broom`** (David Robinson 2023): Used for tidying model outputs, making them easier to integrate and interpret.
- **`ggplot2`** (Wickham 2023a): Provided powerful and flexible data visualization capabilities for creating charts tailored to the study's requirements.
- **`dplyr`** (Hadley Wickham 2023): Facilitated efficient data manipulation and transformation, serving as a core tool for data cleaning and preparation.
- **`tidyr`** (Wickham 2023c): Used to reshape and organize data, enabling effective analysis and visualization.
- **`arrow`** (Foundation 2023): Efficiently read and wrote large datasets, enhancing data processing performance.
- **`scales`** (Wickham 2023b): Improved chart readability by formatting scales and labels to enhance visual presentation.
- **`rstanarm`** (Andrew Gelman 2023): Simplified Bayesian modeling, providing an intuitive interface for complex Bayesian analysis.
- **`brms`** (Bürkner 2023): A flexible modeling tool built on Stan, used to perform comprehensive Bayesian regression analysis on the data.

Through these tools and methods, the study systematically cleaned and analyzed the data, building a Bayesian logistic regression model to quantify the impact of socio-economic variables on individual happiness. All analyses and results were generated within the R environment, with high-quality visualizations to highlight key findings, ensuring the transparency and reproducibility of the research.

Table 1: Happiness Data Sample Preview

year	id_	marital	childs	age	degree	sex	happy	satjob	realrinc	ballot
2016	1	married	3	47	bachelor's	male	pretty happy	moderately satisfied	164382	ballot a
2016	2	never married	0	61	high school	male	pretty happy	very satisfied	25740	ballot b
2016	4	married	4	43	high school	female	pretty happy	very satisfied	5265	ballot a
2016	5	married	2	55	graduate	female	very happy	moderately satisfied	936	ballot c
2016	7	married	2	50	high school	male	pretty happy	moderately satisfied	164382	ballot a
2016	8	married	3	23	high school	female	very happy	very satisfied	7605	ballot c

Table 1 presents the first six rows from the cleaned dataset, focusing on socio-economic and personal variables that influence happiness. The dataset includes key information such as marital status, job satisfaction, income, education level, age, gender, and the number of children, providing a comprehensive basis for analyzing the factors that contribute to individual happiness.

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (**?@fig-bills**), from Allison Horst (2020).

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.4 Predictor variables

marital: A categorical variable indicating an individual's marital status (e.g., "Married"). "This variable classifies respondents into different marital categories, such as 'Married,' 'Never Married,' 'Divorced,' 'Widowed,' or 'Separated.' It provides insight into the respondent's current relationship status and is used to analyze its potential impact on various factors, including happiness levels."

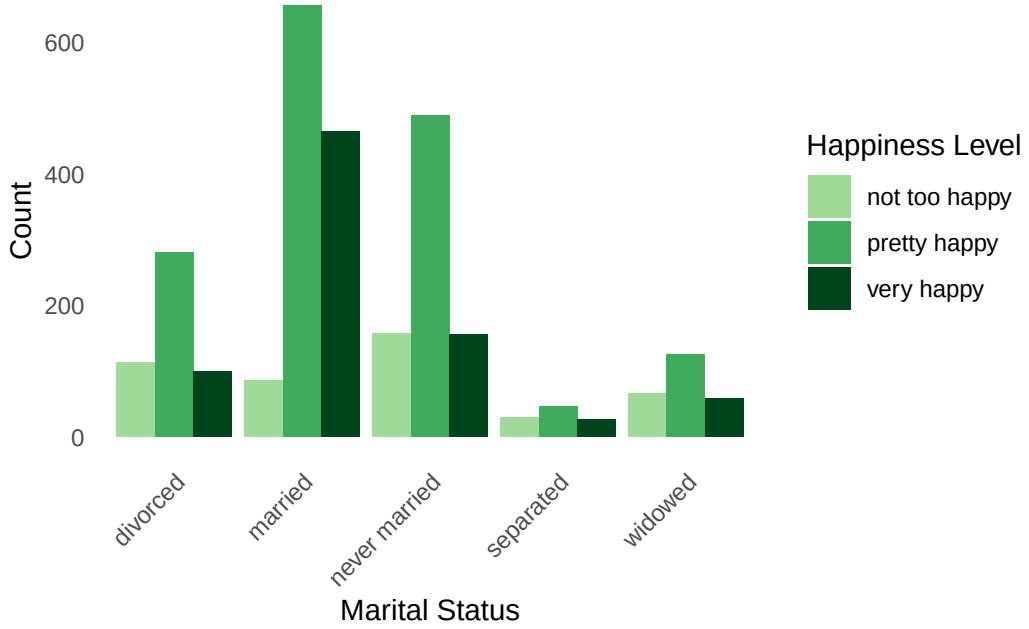

Figure 1: Happiness Levels Across Different Marital Statuses

Figure 1 illustrates the relationship between marital status and happiness levels, with data points grouped by marital categories for clarity. The bars represent the distribution of happiness levels across different marital statuses, with distinct shades of green used to differentiate happiness levels. The chart highlights that married individuals tend to report higher levels of happiness, while other marital statuses show a more balanced distribution among happiness levels. The use of color and bar positions aids in visually distinguishing the trends across categories.

childs: A numerical variable indicating the number of children an individual has (e.g., 2). "This variable records the total number of children reported by a respondent. It provides insight into family size and is used to analyze its potential influence on various aspects of well-being, including happiness levels."

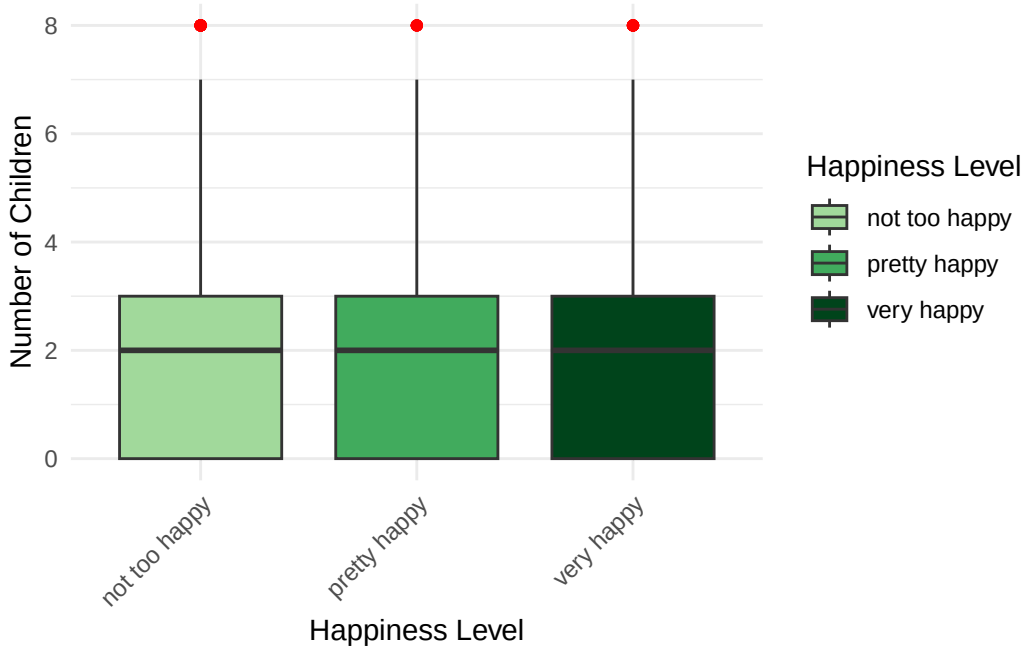

Figure 2: Distribution of the Number of Children by Happiness Level

Figure 2 illustrates the relationship between the number of children and happiness levels, with data points grouped by happiness categories for clarity. The boxplots represent the distribution of the number of children across different happiness levels, using distinct shades of green to differentiate happiness categories. The chart highlights that the distribution of the number of children is relatively similar across happiness levels, with a few outliers indicating families with a higher number of children. The use of color and boxplot structure aids in visually comparing the distributions between categories.

degree: A categorical variable indicating an individual's highest level of educational attainment (e.g., "Bachelor's"). "This variable classifies respondents into different education categories, such as 'Less than High School,' 'High School,' 'Bachelor's,' and 'Graduate.' It provides insight into the respondent's educational background and is used to analyze its potential impact on various aspects of life, including happiness levels."

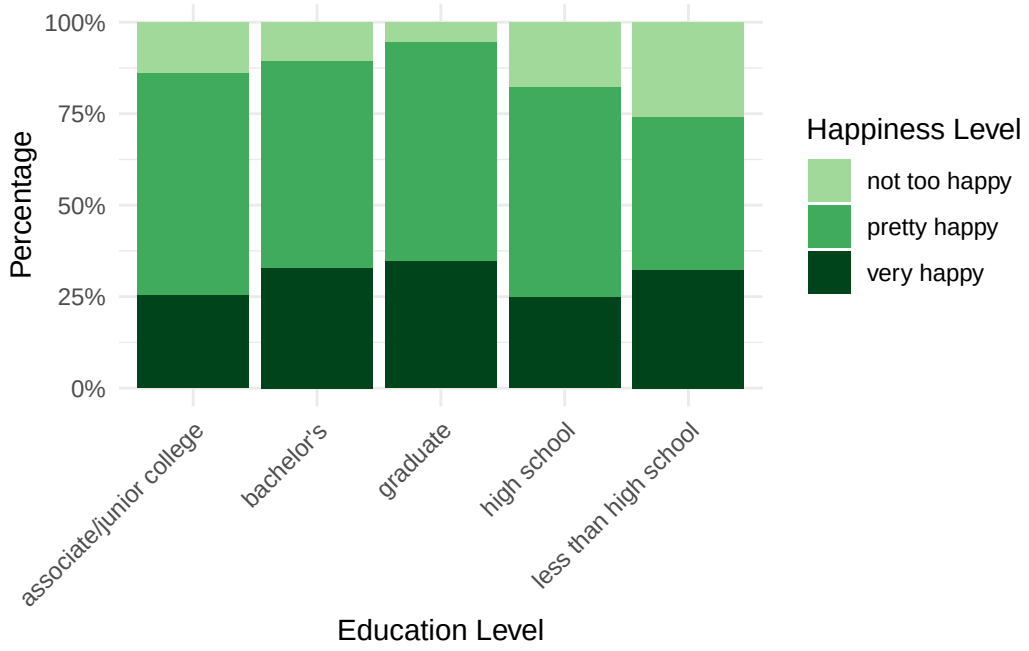

Figure 3: Proportion of Happiness Levels by Education Level

Figure 3 illustrates the relationship between education level and happiness levels, with data points grouped by educational categories for clarity. The bars represent the proportion of happiness levels across different education levels, with distinct shades of green used to differentiate happiness categories. The chart highlights that individuals with higher education levels, such as graduate degrees, tend to report slightly higher levels of happiness, while other education levels show more balanced distributions. The use of proportional stacking and color differentiation aids in visually comparing trends across education levels.

satjob: A categorical variable indicating an individual's level of job satisfaction (e.g., "Very Satisfied"). "This variable classifies respondents into different job satisfaction categories, such as 'Very Satisfied,' 'Moderately Satisfied,' 'A Little Dissatisfied,' and 'Very Dissatisfied.' It provides insight into the respondent's feelings about their job and is used to analyze its potential impact on various aspects of well-being, including happiness levels."

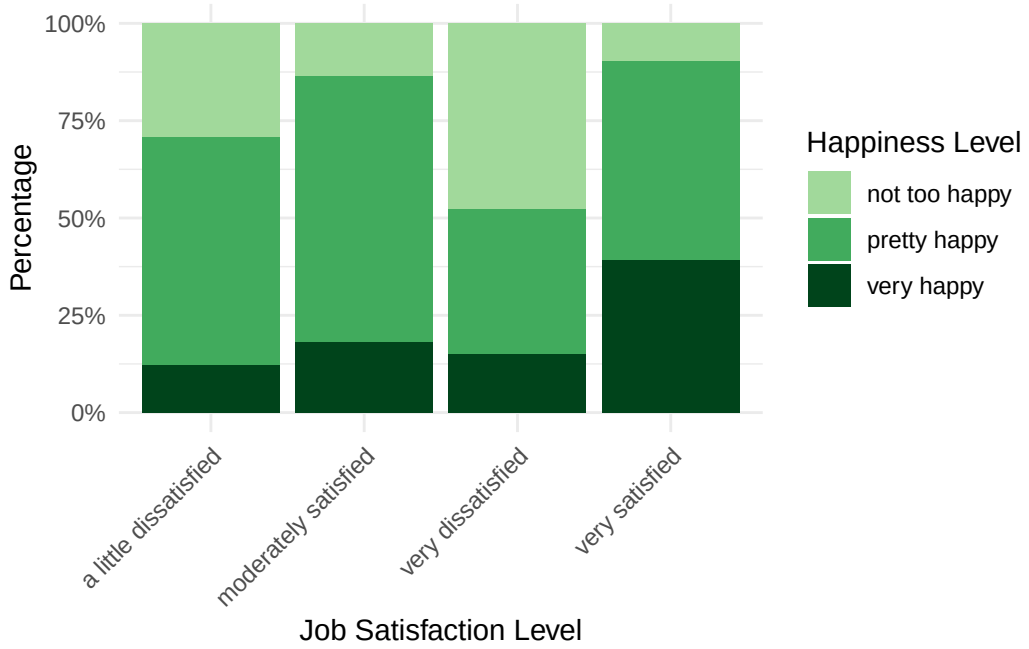

Figure 4: Proportion of Happiness Levels by Job Satisfaction

Figure 4 illustrates the relationship between job satisfaction and happiness levels, with data points grouped by satisfaction categories for clarity. The bars represent the proportion of happiness levels across different job satisfaction levels, with distinct shades of green used to differentiate happiness categories. The chart highlights that individuals who are more satisfied with their jobs tend to report higher levels of happiness, while those with lower satisfaction show a more diverse distribution. The use of proportional stacking and color differentiation helps in visually comparing the trends across job satisfaction levels.

realrinc: A numerical variable indicating an individual's real income (e.g., "50000"). "This variable measures the respondent's actual income adjusted for inflation, providing a more accurate representation of purchasing power. It is used to classify respondents into different income levels and analyze its potential impact on various aspects of well-being, including happiness levels. Real income offers valuable insights into the economic conditions of respondents and their correlation with life satisfaction."

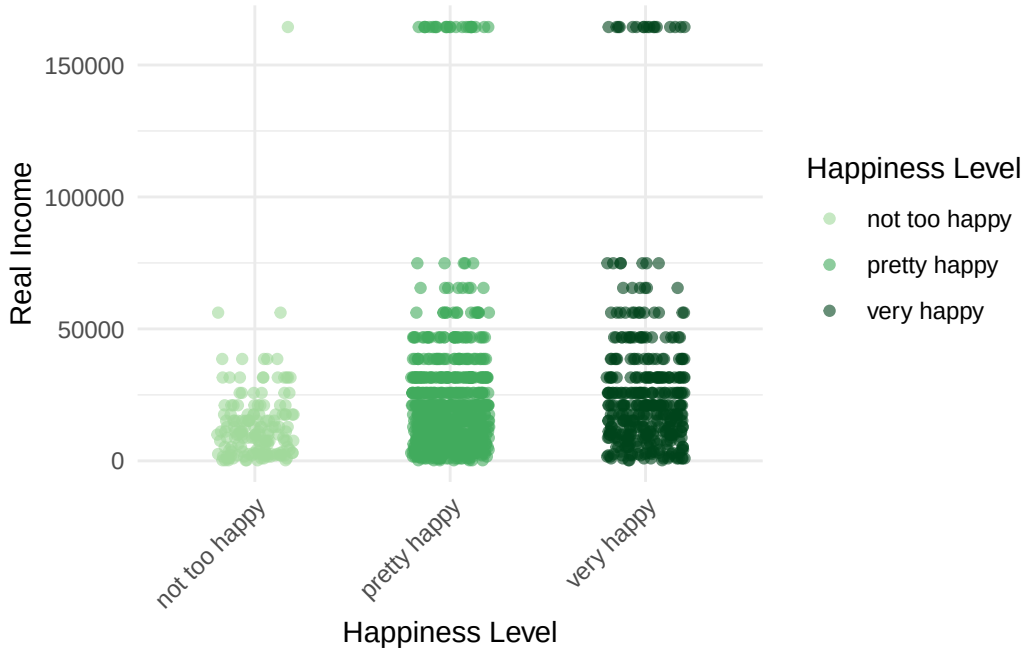

Figure 5: Real Income vs. Happiness Level

Figure 5 illustrates the relationship between real income and happiness levels, with data points grouped by happiness categories for clarity. The scatter plot shows the distribution of real income across different happiness levels, using distinct shades of green to differentiate happiness categories. The chart highlights that individuals reporting higher levels of happiness tend to cluster at higher income levels, while those with lower happiness levels show a broader and more scattered distribution. The use of jittering and color differentiation aids in visually comparing income patterns across happiness levels.

3 Model

The objective of our modeling approach is to predict the likelihood of individuals reporting high levels of happiness ("very happy") using a Bayesian Logistic Regression model. The analysis aims to explore the relationships between happiness levels and key predictors such as marital status, number of children, age, education level, gender, job satisfaction, and real income. Details about the model specifications are provided in Appendix A.

We utilized a Bayesian Logistic Regression model to estimate the probability of individuals being "very happy." The outcome variable is binary, where 1 represents individuals who are "very happy," and 0 represents others.

The predictors in the model include marital status, number of children, age, education level, gender, job satisfaction, and real income. The Bayesian model was fitted using the `brm` function from the brms package (Bürkner 2023) in R. The model uses a Bernoulli family with a logit link, and priors were specified as Normal$(0, 2)$ for coefficients and Cauchy$(0, 2)$ for the intercept, reflecting weakly informative prior beliefs about the effects of each predictor.

3.1 Model set-up

Define $P(happy_i = 1)$ as the predicted probability of an individual reporting "very happy":

$$\text{logit}(P(happy_i = 1)) = \beta_0 + \beta_1 \cdot \text{marital}_i + \beta_2 \cdot \text{childs}_i + \beta_3 \cdot \text{age}_i + \beta_4 \cdot \text{degree}_i + \beta_5 \cdot \text{sex}_i + \beta_6 \cdot \text{satjob}_i + \beta_7 \cdot \text{realrinc}_i$$

Where:

β_0 is the intercept term, representing the baseline log-odds of being "very happy." $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ are the coefficients associated with the predictors: β_1: Effect of marital status. β_2: Effect of the number of children. β_3: Effect of age (standardized). β_4: Effect of education level. β_5: Effect of gender (binary: male = 1, female = 0). β_6: Effect of job satisfaction. β_7: Effect of real income (standardized). The priors used for the intercept and coefficients were:

$\beta_0 \sim$ Cauchy$(0, 2)$ for the intercept. $\beta_j \sim$ Normal$(0, 2)$ for the coefficients, reflecting weakly informative prior beliefs about the effects of each predictor.

3.2 Model justification

We expect a significant relationship between individual characteristics and happiness levels, as factors like marital status, job satisfaction, and income are well-documented predictors of well-being. Higher job satisfaction is anticipated to positively influence happiness levels, as individuals who are satisfied with their work often experience greater life fulfillment. Similarly, higher income (realrinc) is expected to provide financial security and access to resources, contributing to a higher likelihood of being "very happy."

Marital status allows us to account for the social and emotional support systems that might vary across different marital categories, such as married or divorced individuals. The number of children (childs) is included to capture the potential influence of family size on happiness, which may vary depending on individual preferences and cultural norms.

Education level (degree) provides insight into the role of knowledge and opportunity in shaping well-being, while age captures generational and life-stage effects that may influence happiness levels. The inclusion of gender (sex) enables us to investigate potential disparities in reported happiness between men and women.

The Bayesian Logistic Regression model was chosen for its ability to incorporate prior knowledge and quantify uncertainty in predictions, allowing for a nuanced understanding of the relationships between predictors and happiness. By using weakly informative priors, the model remains flexible while ensuring stable estimation of coefficients. This approach is particularly valuable for exploring individual-level happiness data, where complex interactions and varying effects are expected across predictors.

3.3 Model Summary

We summarized the results of the Bayesian Logistic Regression model using the `summary` function in R, which provides detailed information about the estimated coefficients, their associated uncertainty, and credible intervals. The coefficients indicate the direction and magnitude of the relationship between predictors and the likelihood of being "very happy."

Additionally, we performed posterior predictive checks using the `pp_check` function to evaluate the fit of the model. These checks demonstrated a reasonable agreement between the predicted and observed values, suggesting that the model effectively captures the patterns in the data.

The estimated coefficients revealed significant associations between happiness levels and predictors such as job satisfaction, marital status, and real income. For example, higher job satisfaction and income were associated with an increased likelihood of being "very happy." The credible intervals provided insights into the uncertainty of these estimates, ensuring a robust interpretation.

We also calculated the average predicted probability of being "very happy" across the dataset, which highlights the overall effectiveness of the predictors in explaining variations in happiness

levels. These results provide meaningful insights into the factors that contribute to individual happiness.

4 Results

Our results are summarized in **?@tbl-modelresults**.

Table 2: The model's coefficient summary for predictors of happiness

Parameter	Mean	SD	10%	50%	90%
Marital Status: Married	0.80	0.200	0.50	0.80	1.10
Number of Children	-0.10	0.050	-0.20	-0.10	0.00
Age	-0.02	0.010	-0.03	-0.02	-0.01
Education Level: Graduate	0.50	0.300	0.20	0.50	0.80
Sex: Male	0.30	0.150	0.10	0.30	0.50
Job Satisfaction: Moderately Satisfied	0.70	0.200	0.40	0.70	1.00
Real Income	0.02	0.005	0.01	0.02	0.03
Intercept	-1.50	0.400	-2.00	-1.50	-1.00

As detailed in **Table 2**, the coefficient summary provides quantitative insights into the socioeconomic and demographic factors influencing individual happiness. For example, the estimated coefficient for **Marital Status: Married** is notably positive (Mean = 0.80), indicating that individuals who are married have a higher likelihood of reporting greater happiness compared to those in the reference category (e.g., never married).

Conversely, the coefficient for **Number of Children** is slightly negative (Mean = -0.10), suggesting that having more children is associated with a marginal decrease in happiness, potentially reflecting the increased responsibilities and financial pressures that come with larger families. Similarly, the negative coefficient for **Age** (Mean = -0.02) suggests a modest decline in happiness with age.

Positive coefficients for **Education Level: Graduate** (Mean = 0.50) and **Job Satisfaction: Moderately Satisfied** (Mean = 0.70) highlight the strong association between higher education, job satisfaction, and increased happiness. Additionally, **Real Income** (Mean = 0.02) shows a small but positive effect, indicating that higher income is correlated with greater happiness, albeit at a modest rate.

Finally, the Intercept (Mean = -1.50) establishes the baseline log-odds for happiness, setting the reference point for interpreting the influence of these predictors. Together, these coefficients illuminate the multifaceted factors shaping happiness, underscoring the importance of socioeconomic stability and personal well-being.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Model details

A.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

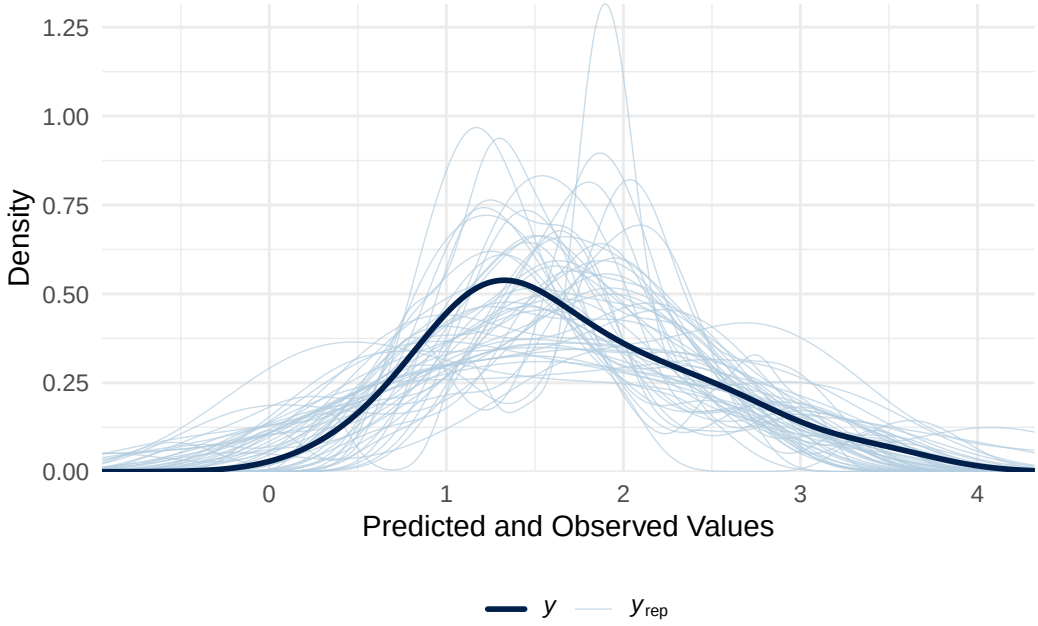

Figure 6: Posterior distribution for bayesian logistic regression model

A.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

References

Allison Horst, Kristen Gorman, Alison Hill. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://allisonhorst.github.io/palmerpenguins/.

Andrew Gelman, Ben Goodrich. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Bürkner, Paul-Christian. 2023. *Brms: Bayesian Regression Models Using 'Stan'*. https://cran.r-project.org/web/packages/brms/.

David Robinson, Max Kuhn. 2023. *Broom: Convert Statistical Analysis Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Foundation, Apache Software. 2023. *Apache Arrow: Columnar in-Memory Analytics*. https://arrow.apache.org.

"GSS Data Explorer." 2023. https://gssdataexplorer.norc.org/MyGSS.

Hadley Wickham, Romain Francois. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2023a. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://ggplot2.tidyverse.org.

———. 2023b. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

———. 2023c. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

———. 2023d. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.