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

Marital Status and Job Satisfaction as Major Positive Predictors of Happiness: A Bayesian Analysis of Socio-Economic and Personal Factors Influencing Well-Being

Shuheng (Jack) Zhou

December 2, 2024

This study employs a logistic regression model to investigate the influence of diverse social, economic, and personal factors on self-reported happiness. By focusing on variables such as income, marital status, education level, job satisfaction, and the number of children, it examines their respective roles in shaping individual well-being. The findings underscore the significance of these factors, illustrating how socioeconomic conditions and personal attributes interact to influence happiness. This analysis offers a clearer perspective on the primary predictors of well-being, with meaningful implications for enhancing both individual and societal happiness.

Table of contents

*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

1 Introduction

Happiness is a cornerstone of human well-being and a long-standing focus of psychological and sociological inquiry. Despite extensive research, understanding the complex interplay of factors contributing to happiness remains a significant challenge. While prior studies have primarily centered on socio-economic determinants such as income, education, and employment, comparatively less attention has been devoted to the influence of personal characteristics and socio-cultural variables. This paper seeks to address this issue by examining how a combination of socio-economic and personal factors collectively shape individual happiness.

The essential role of socio-economic stability in fostering happiness is well-documented. For example, Layard (Layard 2005) emphasizes income and job security as fundamental determinants of life satisfaction, while Easterlin (Easterlin 1974) demonstrated that rising income does not always translate into increased happiness—a phenomenon widely recognized as the "Easterlin Paradox." More recently, Helliwell et al. (Helliwell et al. 2020) have highlighted the importance of social connections and subjective well-being, suggesting that these dimensions often outweigh purely economic metrics. Despite these meaningful contributions, a notable issue persists in understanding how socio-economic factors interact with personal characteristics—such as marital status, gender, and the number of children—in shaping happiness.

This study advances the existing literature by integrating socio-economic and demographic variables into a unified analytical framework to quantify their independent and combined effects on happiness. Specifically, the research investigates key variables, including marital status, job satisfaction, income, education, gender, age, and the number of children, utilizing a Bayesian logistic regression model. By focusing on the interplay between personal and socio-economic factors, the study provides a more detailed perspective, addressing important gaps in the literature.

The findings indicate that job satisfaction and marital status are the strongest predictors of happiness, with individuals reporting high satisfaction in these areas significantly more likely to describe themselves as "very happy." Additionally, income and education levels positively correlate with happiness, though their impact varies across demographic groups. Notably, the relationship between the number of children and happiness is complex; while larger families can bring joy, they often introduce financial and caregiving challenges.

These results validate earlier findings on the importance of socio-economic stability while offering new findings into how demographic factors influence well-being. The study provides practical recommendations for policymakers and social scientists to design targeted interventions aimed at enhancing happiness across different population segments.

This paper is structured as follows: Section 1 introduces the context and objectives of the study. Section 2 outlines the dataset and describes the selection of key variables. Section 3 presents the Bayesian logistic regression framework employed in the analysis. Section 4 discusses the findings, focusing on the effects of job satisfaction, marital status, and other predictors. Section 5 explores the broader implications, limitations, and potential directions for

future research, while Section B provides detailed diagnostics and validation to ensure the robustness of the results.

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 detailed 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 a better understanding of the determinants of happiness but also offer empirical support for policies aimed at enhancing societal well-being.

2 Data

2.1 Data Source

The data utilized in this study was sourced from the GSS Data Explorer ("GSS Data Explorer" 2023), an online platform maintained by NORC that provides an extensive repository of information collected through the General Social Survey (GSS). The GSS dataset captures a wide range of social, economic, and personal characteristics, making it a useful resource for analyzing societal behaviors and trends. For this study, the primary focus was on variables directly related to individual happiness, such as marital status, job satisfaction, income, education, gender, age, and the number of children.

In addition to the GSS dataset, there are other datasets that could have been used to analyze the factors influencing happiness. For example, the European Social Survey (ESS) provides rich data on life satisfaction and happiness, but it primarily focuses on European countries, making it less suitable for this study's focus on American society. In contrast, the GSS dataset is centered on the U.S. population, offering more relevant data for examining the impact of socio-economic conditions and personal characteristics on happiness in the United States. Similarly, the World Values Survey (WVS) offers global social-economic data, but due to its broad scope, it may not be as suitable for focusing on happiness factors specific to the U.S. population. The American Time Use Survey (ATUS) primarily focuses on time allocation rather than directly measuring life satisfaction or happiness, making it less aligned with the objectives of this study. The GSS dataset was chosen for its focus on U.S. residents

and its detailed records of socio-economic and personal characteristics, which are essential for analyzing the factors influencing happiness among Americans.

2.2 Measurement

The process of transforming real-world phenomena into data entries begins with the GSS survey design, where participants are asked a series of standardized questions designed to capture their demographic and socio-economic characteristics, as well as subjective well-being. For example:

- Happiness was measured using responses to the question, *"Taking all things together, would you say you are very happy, pretty happy, or not too happy?"* Responses were coded as categorical variables in the dataset.
- Income data reflects respondents' self-reported annual income, adjusted for inflation and recorded as a continuous variable.
- Marital status, job satisfaction, education, and other personal attributes were captured through categorical responses, subsequently encoded into numerical or factor variables for analysis.

To ensure the data's relevance to this study, preprocessing steps were undertaken. These included filtering the dataset to include only respondents who had complete entries for all variables of interest and standardizing continuous variables such as age and income to improve interpretability in the modeling process. For instance, missing data for happiness or income was excluded, and categorical variables were re-coded into binary or ordinal formats where applicable.

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 suite 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 Bayesian regression analysis on the data.
- **bayesplot**(Gabry et al. 2024): A visualization package for Bayesian model diagnostics, posterior checks, and MCMC outputs, offering tools for clear and intuitive graphical summaries.
- **testthat** (Wickham, Hester, and Müller 2023): Enabled modular and reproducible testing of data and analysis workflows, ensuring the validity of datasets, functions, and model outputs while enhancing code reliability.

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: Sample Preview of Happiness Data: A snapshot showing key variables

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 detailed basis for analyzing the factors that contribute to individual happiness.

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

2.3 Variables

2.3.1 Outcome Variables

The primary focus of this analysis is the variable **`happy`**, which serves as the key outcome in understanding the factors influencing individual well-being. This variable captures self-reported happiness levels and is categorized into three distinct groups:

- **Very happy**: Represents individuals who report a high level of happiness, often indicative of positive life circumstances and strong emotional well-being.
- **Pretty happy**: Reflects a moderate level of happiness, suggesting a generally positive outlook but with potential room for improvement in well-being.
- **Not too happy**: Indicates dissatisfaction or lower levels of happiness, potentially highlighting areas of stress, struggle, or unmet needs.

2.3.2 Predictor Variables

age: Represents the respondent's age, measured in years.

sex: Denotes the biological sex of the respondent, categorized as: male and female.

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 shows family size and is used to analyze its potential influence on various aspects of well-being, including happiness levels."

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 finding into the respondent's current relationship status and is used to analyze its potential impact on various factors, including happiness levels."

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.

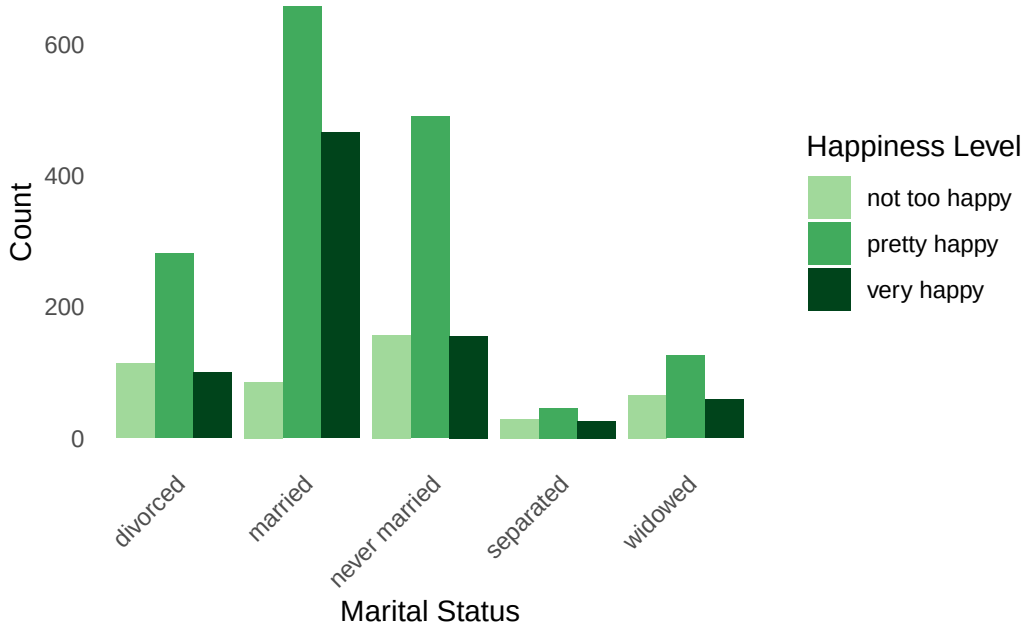


Figure 1: Happiness Levels Across Different Marital Statuses: Depicting the Distribution of Self-Reported Happiness Levels ('Very Happy', 'Pretty Happy', 'Not Too Happy') by Marital Status Categories

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,' 'associate/junior college,' and 'Graduate.' It examines the respondent's educational background and is used to analyze its potential impact on various aspects of life, including happiness levels."

Figure 2 illustrates the relationship between education level and happiness levels, with data points grouped by educational categories for clarity. The bars display the proportion of happiness levels by education categories, using distinct shades of green to represent happiness levels. The chart indicates that individuals with higher education, such as graduate degrees, are more likely to report higher happiness levels, while other education categories exhibit more uniform distributions. Proportional stacking and color differentiation help compare trends effectively across education levels.

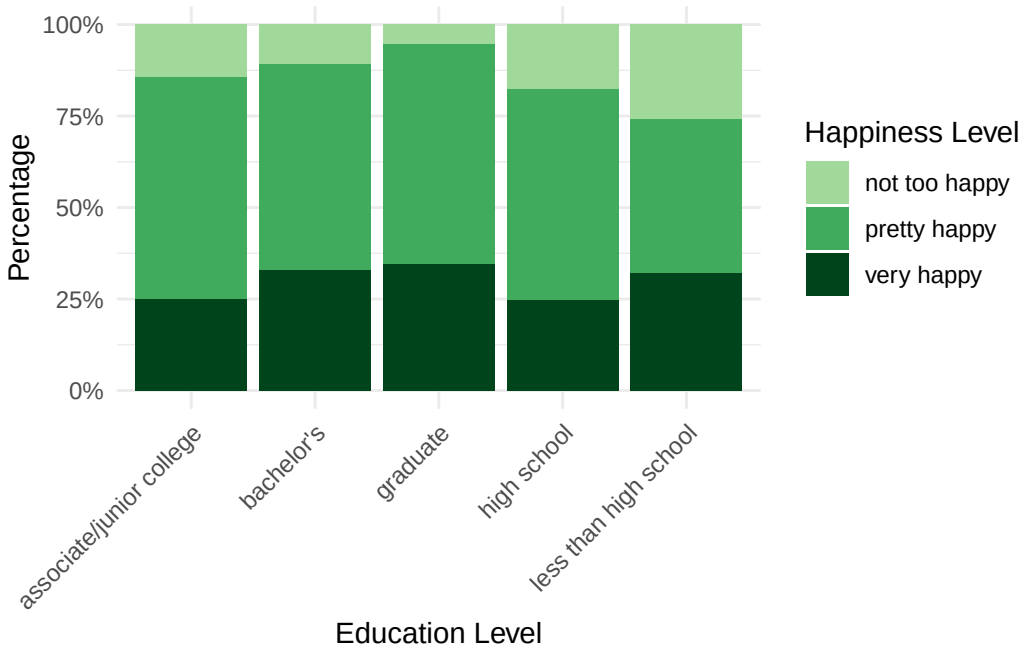


Figure 2: Proportion of Happiness Levels by Education Level: Representing the Distribution of Happiness Categories Across Different Levels of Educational Attainment

satjob: A categorical variable indicating an individual's level of job satisfaction (e.g., "Very Satisfied"). "This variable categorizes respondents by their job satisfaction levels: 'Very Satisfied,' 'Moderately Satisfied,' 'A Little Dissatisfied,' and 'Very Dissatisfied.' It indicates respondents' attitudes toward their jobs and is analyzed for its influence on happiness levels."

Figure 3 illustrates the relationship between job satisfaction and happiness levels, with data points grouped by satisfaction categories for clarity. The bars display the proportion of happiness levels by job satisfaction categories, using distinct shades of green to represent happiness levels. The chart indicates that individuals with higher job satisfaction are more likely to report greater happiness, while those with lower satisfaction exhibit a broader distribution. Proportional stacking and color differentiation facilitate comparisons across job satisfaction levels.

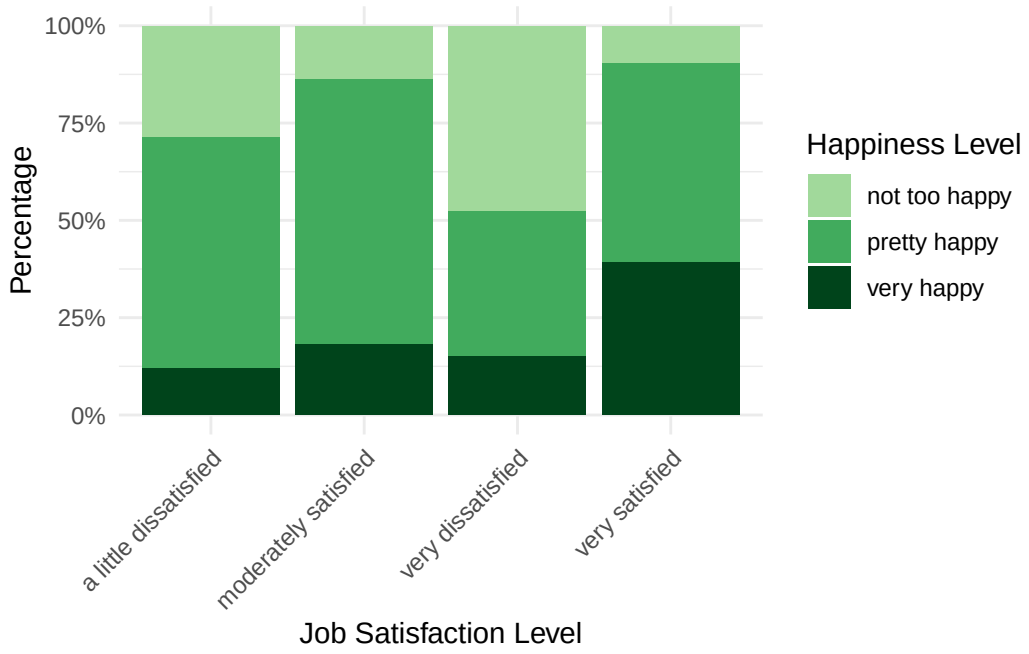


Figure 3: Proportion of Happiness Levels by Job Satisfaction: Depicting the Distribution of Self-Reported Happiness Categories Across Different Job Satisfaction Levels

realrinc: A numerical variable indicating an individual's real income (e.g., "50000"). "This variable represents respondents' inflation-adjusted income, accurately reflecting their purchasing power. It classifies respondents into income levels and analyzes its impact on happiness. Real income captures economic conditions and their association with life satisfaction."

Figure 4 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.

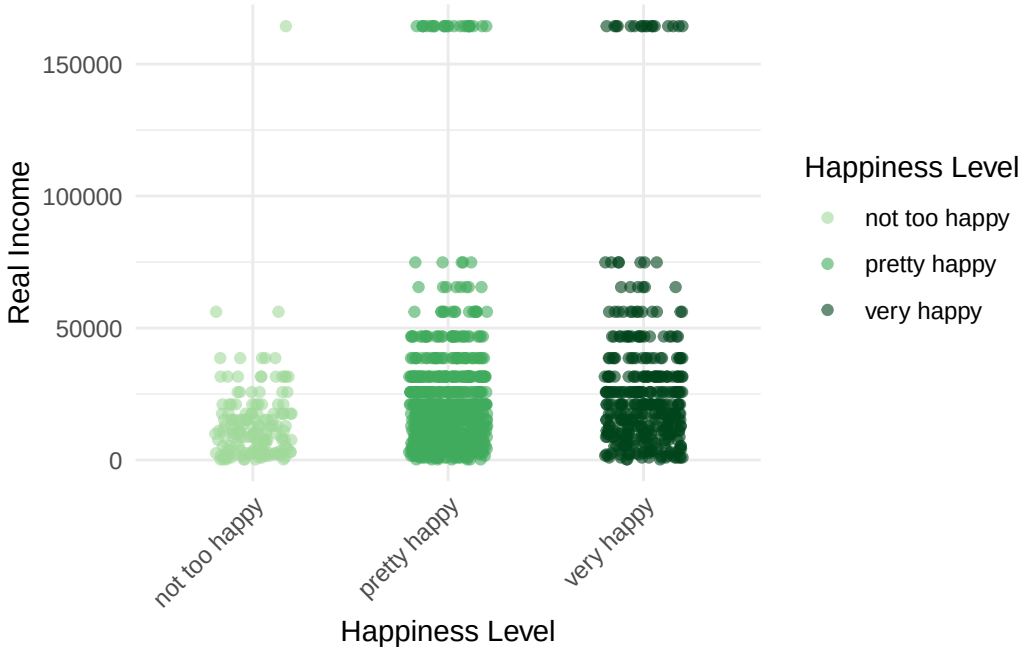


Figure 4: Real Income and Happiness Level: This analysis explores the distribution of real income across different happiness categories, using jitter points to illustrate the variability of individual data within each category

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 B.

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.

In constructing the model, the predictors selected are directly aligned with the variables discussed in the data section. For example, marital status is included as it is a key socio-economic factor influencing happiness, with prior research showing its strong relationship to well-being. Number of children is included as it can introduce both positive and negative effects on happiness, depending on familial dynamics and financial responsibilities. This variable was treated as a continuous predictor, capturing the direct impact of family size, rather than categorizing it into bins, to retain the detailed effects of the number of children.

Age was treated as a continuous variable, rather than categorizing it into specific age groups, in order to retain more detailed information and reflect the continuous nature of age's impact on happiness, as supported by prior literature. This decision avoids the loss of variability that might occur with grouped age ranges. Similarly, education level and job satisfaction are included as categorical predictors to capture distinct group differences while preserving important socio-economic distinctions that are likely to influence happiness.

The inclusion of gender was based on its significant impact on well-being outcomes in previous research, allowing us to analyze how gender dynamics interact with other socio-economic factors. Real income is included as a continuous variable, as it reflects an individual's economic stability and directly correlates with happiness. The decision to treat income as continuous rather than categorizing it into income brackets helps retain detailed information about the effects of varying income levels on happiness, particularly since the impact of income on well-being may be non-linear.

We employed a Bayesian Logistic Regression model using the `brm` function from the `brms` package (Bürkner 2023) in R, which is particularly useful for fitting complex regression models with a probabilistic framework. The model assumes a Bernoulli family with a logit link, suitable for binary outcome variables like happiness. Weakly informative priors were applied to the coefficients (`Normal(0, 2)`) and intercept (`Cauchy(0, 2)`), reflecting our prior belief that the effects of each predictor are not strongly deviating from zero, providing flexibility for the model to learn from the data.

In summary, the predictors were selected based on their relevance to understanding individual happiness, with each variable being included to reflect real-world influences as thoroughly as

possible. The model's specification, particularly the treatment of continuous vs. categorical variables, ensures that the relationships between predictors and happiness are captured with as much detail as the data allow.

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:

$\beta_0 \sim \text{Cauchy}(0, 2)$ for the intercept. $\beta_j \sim \text{Normal}(0, 2)$ 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 finding 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 detailed 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 meaningful 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 indicated a reasonable agreement between the predicted and observed values, suggesting that the model effectively captures the patterns in the data.

The estimated coefficients indicated 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 findings about 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 findings about the factors that contribute to individual happiness.

3.4 Model Assumptions and Limitations

The Bayesian logistic regression model is built upon several key assumptions that underpin its structure and interpretation. First, the model assumes a linear relationship between the predictors and the log-odds of happiness, meaning that the effects of predictors remain constant across their range. This assumption simplifies the modeling process but may fail to capture complex non-linear patterns or interactions between variables, such as potential non-linear relationships between age and happiness. The model also assumes independence between predictors, which may not hold in real-world data where multicollinearity or interaction effects are common. Additionally, the model relies on weakly informative priors (e.g., Normal(0, 2) for coefficients and Cauchy(0, 2) for the intercept) to provide flexibility while avoiding overfitting. While these priors stabilize the estimation process, they may still introduce bias or increase uncertainty in cases with small samples or sparse categories. Lastly, the model

assumes homoscedasticity and Gaussian-distributed residuals, which may not fully account for variability or skewness in happiness data.

Despite its strengths, the model has certain limitations that should be noted. First, the linearity assumption between predictors and log-odds may oversimplify relationships, leading to potential biases in effect estimation. For instance, the relationship between income and happiness might exhibit diminishing returns, which a linear model may struggle to capture. Second, the imbalanced distribution of the target variable—where "very happy" is a minority category—presents a challenge, as the model may prioritize predicting majority categories ("not too happy" and "pretty happy") while underperforming in classifying the minority category. Additionally, while weakly informative priors are generally robust, they may lead to unstable coefficient estimates in cases of sparse data or rare categories, increasing uncertainty in the results. The assumption of homoscedasticity may also be violated in situations where variability in happiness levels depends on specific predictors, such as income or job satisfaction.

3.5 Alternative Models Considered

The Bayesian logistic regression model was chosen due to its unique ability to incorporate prior knowledge, quantify uncertainty, and provide robust parameter estimates in the analysis of complex individual-level happiness data. Unlike traditional logistic regression, the Bayesian framework allows for the inclusion of weakly informative priors, which help stabilize estimates in the presence of limited or imbalanced data (e.g., the "very happy" category being underrepresented) while avoiding overfitting. This is particularly advantageous in handling imbalanced target variable distributions, where the Bayesian model performs better in classifying minority categories. Additionally, the Bayesian model computes posterior distributions for each predictor, offering a detailed understanding of the uncertainty and variability in the estimated effects. This probabilistic framework not only enhances interpretability but also supports better decision-making by providing credible intervals instead of single-point estimates.

In contrast, other models such as decision trees may perform poorly when analyzing complex social data. While decision trees offer advantages in interpretability and handling non-linear relationships, they are highly sensitive to changes in data distribution and noise. In cases of class imbalance, decision trees tend to overfit the majority class, reducing their predictive accuracy for underrepresented categories. Moreover, decision trees lack the ability to provide posterior distributions or quantify uncertainty, making it challenging to robustly evaluate the effects of predictors. These limitations render decision trees less reliable for analyzing complex problems like happiness data.

For these reasons, the Bayesian logistic regression model stands out for its flexibility and robustness, capturing complex non-linear relationships and potential interactions while quantifying uncertainty. It offers more accurate and detailed findings for predicting happiness compared to alternative models, making it a superior choice for this analysis.

4 Results

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

4.1 Coefficient Summary

As detailed in Table 2, the coefficient summary offers a quantitative perspective on the socio-economic and demographic determinants of individual happiness. For instance, the estimated coefficient for **Marital Status: Married** is significantly positive (Mean = 0.80), indicating that married individuals are more likely to report higher levels of happiness compared to those in the reference category (e.g., never married). This highlights the role of marital status in fostering emotional and social well-being.

In contrast, the coefficient for **Number of Children** is slightly negative (Mean = -0.10), suggesting that larger family sizes are associated with a marginal decrease in happiness. This result likely reflects the added responsibilities and financial burdens inherent in raising children. Similarly, the coefficient for **Age** (Mean = -0.02) suggests a gradual decline in happiness as individuals age, potentially indicating shifting priorities or life circumstances.

Positive coefficients for **Education Level: Graduate** (Mean = 0.50) and **Job Satisfaction: Moderately Satisfied** (Mean = 0.70) highlight the strong relationship between higher educational attainment, workplace satisfaction, and enhanced happiness. Furthermore, the coefficient for **Real Income** (Mean = 0.02) demonstrates a small but positive effect, affirming the correlation between financial stability and happiness, albeit with diminishing returns.

The Intercept (Mean = -1.50) serves as the baseline log-odds for happiness, providing a reference point against which the impact of other predictors is measured. Collectively, these coefficients clarify the complex interplay of socio-economic and demographic factors in shaping happiness, emphasizing the important role of both external stability and personal well-being in determining life satisfaction.

As illustrated in Figure 5, the Bayesian logistic regression model provides a clear visualization of the influence of predictor variables on happiness. Each point represents the posterior mean

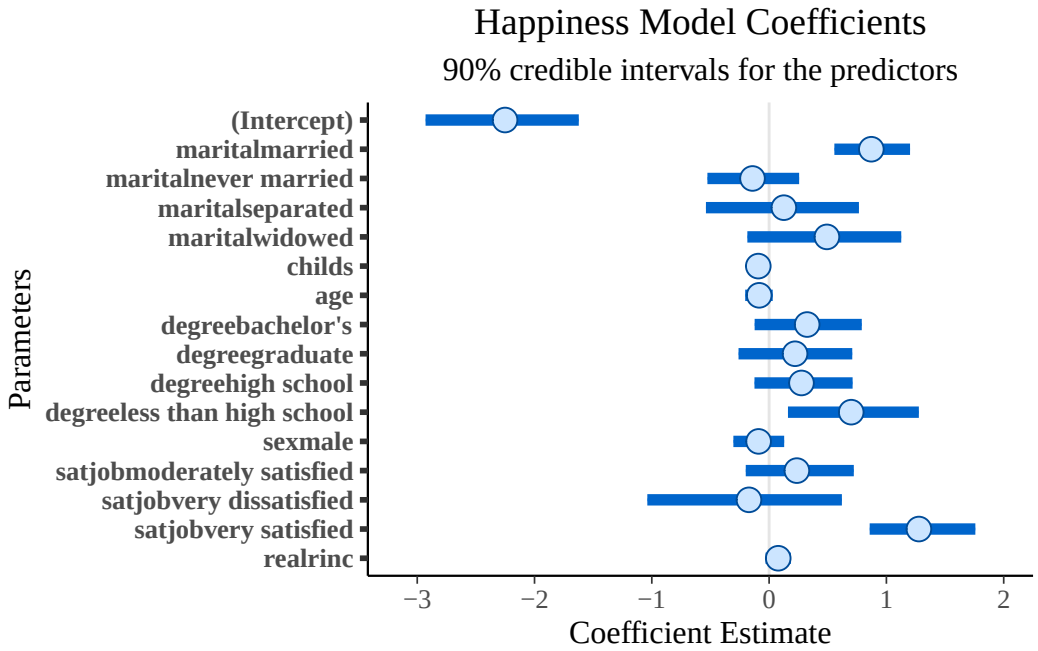


Figure 5: The 90% credible intervals for all model coefficients in the happiness model

estimate, with horizontal lines denoting the 90% credible intervals, offering a reliable measure of uncertainty. Key findings include:

- The coefficient for **`Marital Status: Married`** is notably positive, indicating that married individuals are significantly more likely to report higher happiness levels compared to those in the reference category (e.g., never married). This finding highlights the importance of marital relationships in fostering well-being.
- **`Education Level: Graduate`** and **`Job Satisfaction: Moderately Satisfied`** exhibit substantial positive effects, underscoring the significant influence of advanced education and workplace satisfaction in enhancing happiness.
- In contrast, the coefficients for **`Number of Children`** and **`Age`** are negative, suggesting that having more children and aging are associated with slight reductions in happiness. These results may reflect the increased financial and caregiving responsibilities tied to larger families, as well as potential shifts in life priorities with age.
- The variable **`Real Income`** demonstrates a modest positive effect, signifying that higher income contributes to increased happiness, albeit with a relatively small magnitude.

Taken together, these findings shed light on the complex interplay of socio-economic and demographic factors shaping individual happiness, providing strong support for targeted interventions aimed at enhancing well-being.

5 Discussion

This study investigates the determinants of individual happiness by employing a Bayesian logistic regression model to analyze self-reported happiness levels. Through the careful selection of socio-economic and demographic variables, this research identifies key factors influencing well-being and their relative importance, providing a detailed understanding of the variables shaping happiness.

5.1 Detailed Exploration of Factors Influencing Happiness

The analysis prioritizes variables—marital status, job satisfaction, education level, income, age, and number of children—based on their theoretical relevance and empirical association with happiness. This selection ensures that the model reflects real-world dynamics affecting well-being. Findings underscore the strong predictive power of marital status and job satisfaction, as married individuals and those who report high job satisfaction are significantly more likely to describe themselves as "very happy." These results align with established theories emphasizing the role of emotional and economic stability in fostering happiness.

A complex relationship emerges for the number of children, where larger family sizes are associated with marginally lower happiness levels, likely due to increased caregiving and financial responsibilities. Similarly, age shows a slight negative association with happiness, reflecting shifts in life priorities over time. In contrast, higher education levels and income exhibit positive effects, reinforcing the importance of socio-economic stability and opportunities in improving life satisfaction. These findings provide a detailed perspective on the relationships among socio-economic and demographic factors influencing happiness.

5.2 Strategic Implications of Variable Selection

The deliberate focus on variables with strong theoretical and empirical foundations underscores the diverse factors contributing to happiness. Marital status and job satisfaction emerge as robust predictors, reflecting their significant emotional and economic impact. The inclusion of education and income highlights the role of socio-economic stability in fostering well-being, while the effects of age and family size indicate deeper complexities in how life circumstances shape happiness.

Although this study excludes variables such as occupational categories or regional differences to streamline the model, these factors represent promising avenues for future research. For example, exploring the influence of specific professions or geographic contexts could indicate cultural and environmental interactions with socio-economic predictors. Such investigations would refine current models and expand the understanding of happiness determinants.

5.3 Weaknesses and Future Research Directions

Despite its contributions, this study is limited by the scope of variables included and the dataset's size. Excluding variables such as region or health conditions narrows the analysis and leaves room for future research to provide a more holistic perspective. Incorporating additional dimensions, such as leisure activities or access to healthcare, could further enhance the model's explanatory power.

Expanding the dataset to include broader samples across diverse regions or employing longitudinal data could improve generalizability and offer findings into how happiness evolves over time. Moreover, investigating the impact of external socio-economic factors, such as public policy changes or economic crises, could indicate macro-level influences on well-being.

Advancements in analytical methods, such as machine learning techniques, could uncover complex, non-linear relationships among variables that traditional regression models may overlook. For instance, examining interactions between age, income, and job satisfaction could illuminate essential periods and conditions for optimal well-being.

These directions not only address the limitations of this study but also pave the way for a more detailed understanding of happiness in varied contexts, advancing both academic and policy-oriented research.

5.4 Envisioning the Future of Happiness Studies

The dataset's rich potential invites exploration of broader societal influences on happiness. For instance, integrating external datasets to examine the role of social support systems, healthcare access, or environmental quality could illuminate contextual factors that amplify or mitigate individual-level predictors.

Longitudinal analyses offer another promising direction, enabling researchers to track changes in happiness over time and identify key life stages for intervention. Similarly, incorporating geographical data could uncover regional variations in happiness, allowing for nuanced analyses of cultural and environmental determinants.

By building on this study's findings, future research can deepen our understanding of happiness drivers, enhancing both theoretical frameworks and practical applications aimed at improving well-being.

5.5 The Value of Findings into Happiness Drivers

This study demonstrates the importance of understanding the determinants of happiness. Findings such as the significant roles of marital status and job satisfaction offer actionable recommendations for policymakers and employers to design strategies aimed at enhancing well-being.

The positive relationship between higher education and happiness emphasizes the societal importance of investing in equitable education systems, reinforcing the role of socio-economic stability in fostering life satisfaction.

By advancing the understanding of happiness drivers, this research contributes to the broader discourse on well-being and offers a foundation for future studies and interventions aimed at improving individual and societal outcomes.

Appendix

A Data cleaning process

To prepare the dataset for analysis, I conducted a thorough cleaning process to address missing values, inconsistent formatting, and invalid entries, ensuring the data's accuracy and usability. The cleaning steps were designed to align with the study's focus on socio-economic and personal factors influencing happiness.

A.1 Handling Missing and Invalid Values

- In the `realrinc` (real income) column, entries coded as `-100` were replaced with `NA` to signify missing data, as `-100` was an invalid placeholder value.
- For categorical variables like `marital`, `degree`, `satjob`, and `happy`, entries such as `.n: no answer` and `.d: do not know/cannot choose` were converted to `NA` to exclude non-informative responses. Additionally, in the `satjob` column, responses marked as `.i: inapplicable` were also replaced with `NA` to ensure only relevant data remained.

A.2 Standardizing Categorical Data

- Categorical variables, including `marital`, `degree`, `sex`, `happy`, `satjob`, and `ballot`, were cleaned by removing extra spaces and converting all text to lowercase to maintain consistency and facilitate accurate analysis.

A.3 Adjusting the `childs` Column

- Entries labeled as `"8 or more"` in the `childs` (number of children) column were replaced with `8`, as this was the maximum numeric representation available. This transformation allowed the variable to be converted into a numeric format for quantitative analysis.

A.4 Converting Data Types

- Variables like `age` and `childs` were converted to numeric data types to enable proper statistical analysis and modeling.

A.5 Filtering Irrelevant Responses

- Rows where the `happy` variable contained responses like `.d: do not know/cannot choose` or `.n: no answer` were removed to ensure the dataset only included valid observations relevant to the study's goal of understanding happiness.

A.6 Summary

These cleaning steps were essential for maintaining data integrity and ensuring that the analysis would provide meaningful and reliable results. By standardizing and filtering the data, the study could focus on valid and interpretable responses, minimizing noise and maximizing the quality of findings derived from the dataset.

B Model details

B.1 Posterior predictive check

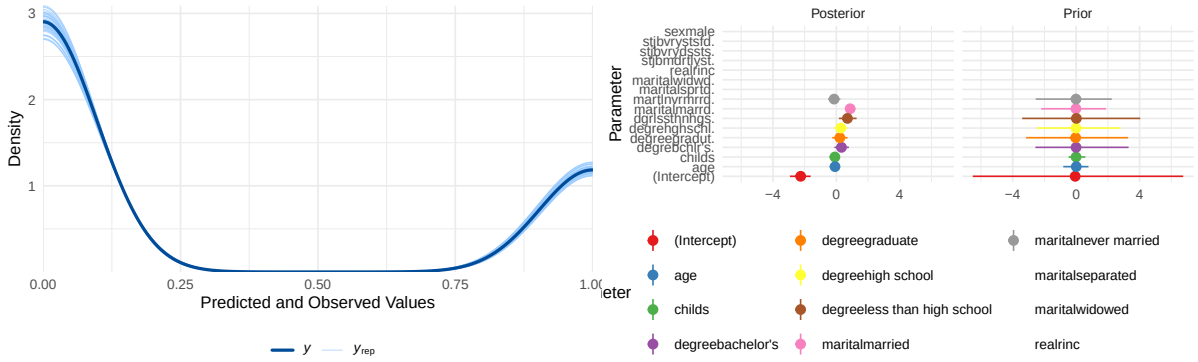
In Figure 6a, we conduct a posterior predictive check to assess the fit of the Bayesian logistic regression model. This plot compares the observed data (y) with the replicated data (y_rep) generated by the model. The overlaid density curves represent multiple posterior predictive distributions, visually indicating how well the model aligns with the observed data. The close alignment of the observed and predicted distributions suggests that the model provides a reasonable fit, accurately reflecting both central tendencies and variability in the data.

In Figure 6b, we compare the posterior distributions of the model parameters with their corresponding priors. This plot illustrates the influence of the data on parameter estimates. The divergence between the posterior and prior distributions highlights the strength of the evidence provided by the data, showing how the Bayesian model refines prior assumptions to produce updated parameter estimates. This comparison emphasizes the role of the data in shaping final inferences and validating the robustness of the model.

B.2 Diagnostics

Figure 7a is a trace plot, depicting the sampled values of each parameter over iterations. The plot shows that the chains for each parameter exhibit patterns resembling a 'hairy caterpillar,' indicating well-mixed chains that effectively explore the posterior distribution. This evidence suggests that the Markov chains have likely converged, ensuring the reliability of posterior estimates.

This plot serves as a key diagnostic tool, verifying that the sampling process is efficient and that the model's posterior distributions are credible.



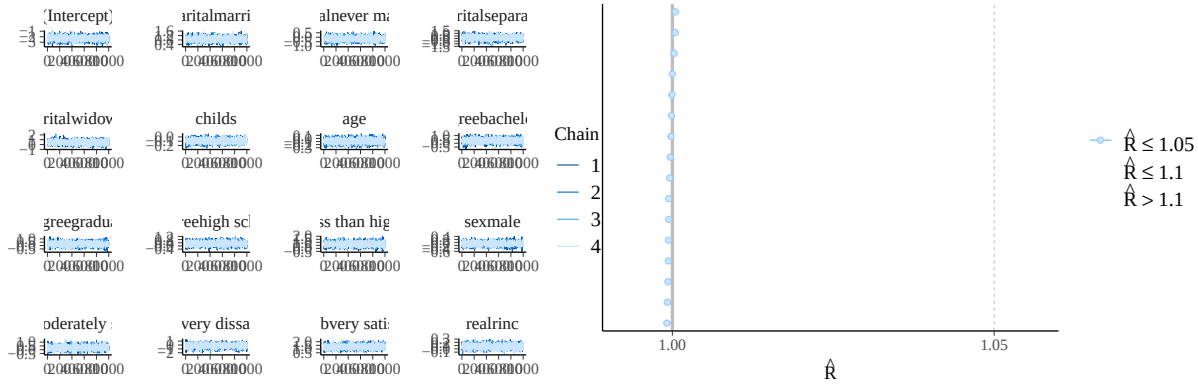
(a) Posterior prediction check

(b) Comparing the posterior with the prior

Figure 6: Examining how the happiness model fits the data and how the posterior compares to the prior

As shown in Figure 7b, the Rhat values are close to 1, further confirming that the chains have likely converged and the model results are reliable.

Together, these diagnostic plots validate the Bayesian model's posterior distributions, supporting robust inference based on the sampled data.



(a) Trace plot

(b) Rhat plot

Figure 7: Checking the convergence of the MCMC algorithm for the happiness model

As shown in Figure 8, the graph visualizes the posterior distributions of the parameters from the Bayesian logistic regression model analyzing happiness predictors. Each horizontal line represents the **90% credible interval**, centered around the median of the posterior distribution for a given parameter. The shaded areas depict the density of the posterior distribution, illustrating the uncertainty and variability of the parameter estimates.

The parameters encompass socio-economic and demographic factors, such as marital status, job satisfaction, education level, and income. Notably, parameters like **Marital Status: Married**

and **Job Satisfaction: Very Satisfied** exhibit positive median values, indicating a positive association with happiness. In contrast, variables such as **Number of Children** and **Age** show slightly negative median values, reflecting a modest decline in happiness with increasing age and family size. This visualization emphasizes the complex relationships between these predictors and individual happiness.

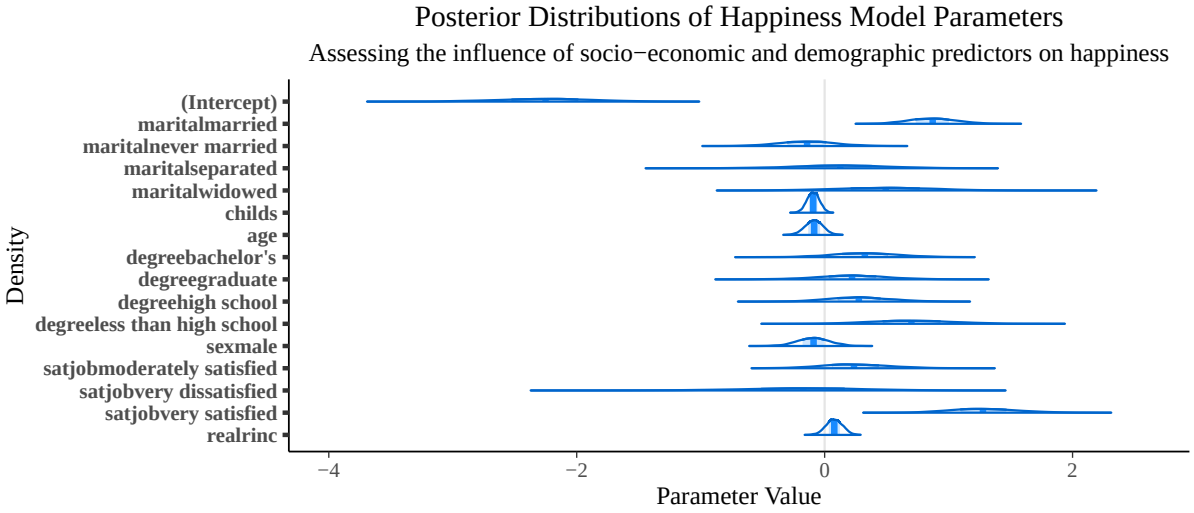


Figure 8: The posterior distributions for all the parameters of the happiness model

C Idealized Happiness Study Methodology and Survey

C.1 Objective

Our goal is to provide a robust analysis of the factors influencing individual happiness by designing a high-quality survey that captures data on key socio-economic and demographic variables. By addressing potential biases and ensuring representative sampling, this methodology aims to produce actionable findings within a manageable budget.

C.1.1 Sampling Approach

We aim for a sample size of approximately 2,500 respondents, selected through **stratified random sampling** to ensure representativeness across demographic groups.

Definition and Justification of Stratified Sampling:

Stratified sampling divides the population into subgroups, or "strata" (e.g., by age, gender, income, education, and marital status). Participants are randomly sampled within each stratum to ensure proportional representation. This method enhances the precision of subgroup

estimates and is well-suited for studies exploring the interplay of socio-economic variables and subjective well-being.

Strengths and Weaknesses of Stratified Sampling:

Stratified sampling reduces sampling error, especially in smaller or more variable subgroups such as single parents or older adults. However, it requires accurate population data to establish strata sizes, and survey costs can increase when oversampling underrepresented groups.

Sampling Simulation:

A simulation can be conducted to compare stratified and simple random sampling methods using previously collected data. This would evaluate the reduction in sampling error and the improved accuracy of subgroup estimates, validating the use of stratified sampling for happiness research.

Stratification and Quotas:

Demographic quotas will be weighted based on recent census data to reflect the population. The primary demographic strata include:

- **Age**: 18-29, 30-44, 45-64, 65+
- **Gender**: Male, Female
- **Marital Status**: Married, Never married, Divorced, Separated, Widowed
- **Education Level**: Less than high school, High school, Associate/Junior college, Bachelor's, Graduate
- **Income Range**: Self-reported continuous values, with adjustments for missing responses.

C.1.2 Recruitment Strategy

We will collaborate with survey panel providers to access verified respondents while supplementing with targeted recruitment via social media ads to reach underrepresented groups. Each participant will receive a $3 incentive to encourage engagement.

Panel Recruitment and Potential Biases:

While panel providers offer efficiency, panel bias may arise due to frequent survey participation by the same individuals. To address this, additional recruitment via social media will focus on underrepresented demographics, ensuring a diverse and balanced sample.

C.1.3 Survey Design and Implementation

The survey will be hosted on Google Forms, featuring a professional introduction, clear instructions, and contact details for queries. It will remain open for 2-3 weeks to allow adequate time for responses.

Survey Structure:

1. **Introduction and Consent**: A message explaining the purpose of the survey and ensuring anonymity.

2. **Screening Questions**: Basic eligibility checks (e.g., age and citizenship).

3. **Core Questions**: Questions on socio-economic and demographic variables, happiness levels, and contributing factors.

4. **Closing Message**: A thank-you note and contact information for follow-up.

Questions will be neutrally worded and logically ordered, with an estimated completion time of 5–7 minutes.

C.1.4 Data Validation

To ensure high-quality responses, the following measures will be implemented:

- **Attention Checks**: Include questions designed to identify inattentive respondents.

- **Duplicate Prevention**: Settings to restrict multiple submissions per respondent.

- **Response Time Analysis**: Flag responses completed significantly faster than the average time for review.

- **Post-Survey Weighting**: Adjust responses to align with demographic quotas.

C.1.5 Analytical Framework

To analyze the collected data, a Bayesian logistic regression model will be applied to quantify the influence of variables like marital status, job satisfaction, and income on happiness levels. The model will incorporate:

- **Priors Based on Previous Research**: Using established studies as a baseline for parameters.

- **Hierarchical Structure**: Accounting for variation across demographic groups.
- **Validation Metrics**: Evaluating model performance using posterior predictive checks and accuracy measures.

This detailed methodology ensures reliable findings into the complex factors shaping individual happiness while laying the groundwork for future studies on well-being determinants.

C.2 Idealized Survey

Welcome to the "Factors Influencing Happiness Survey." This survey seeks to explore the relationship between socio-economic factors, personal characteristics, and individual well-being. Your responses will remain anonymous and contribute to research aimed at understanding the determinants of happiness. Completing the survey will take approximately 5–7 minutes. Thank you for your helpful participation!

If you have any questions or concerns, please contact:

- **Survey Coordinator**: Shuheng (Jack) Zhou
- **Email**: shuheng.zhou@mail.utoronto.ca

1. **What is your marital status?**
    - Married
    - Never married
    - Divorced
    - Separated
    - Widowed
2. **How many children do you have?**
    - 0
    - 1
    - 2
    - 3
    - 4
    - 5
    - 6
    - 7

- 8 or more

3. **What is your age?**
   - Please enter your age in years: _____
4. **What is your gender?**
   - Male
   - Female
5. **What is the highest level of education you have completed?**
   - Less than high school
   - High school
   - Associate or junior college
   - Bachelor's degree
   - Graduate degree
6. **What is your current job satisfaction level?**
   - Very satisfied
   - Moderately satisfied
   - A little dissatisfied
   - Very dissatisfied
   - .i: Inapplicable
   - .d: Do not know/Cannot choose
7. **What is your annual income (in USD)?**
   - Please enter your income: _____

   (If you do not have an income, please enter -100.)
8. **How would you rate your happiness level?**
   - Very happy
   - Pretty happy
   - Not too happy
9. **To ensure data quality, please select "Pretty happy" for this question.**

- Very happy
- Pretty happy
- Not too happy

10. **How would you describe your overall job satisfaction's impact on your happiness?**

    - Strongly positive
    - Moderately positive
    - Neutral
    - Slightly negative
    - Strongly negative

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.

Easterlin, Richard A. 1974. "Does Economic Growth Improve the Human Lot? Some Empirical Evidence." In *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, edited by Paul A. David and Melvin W. Reder, 89–125. New York: Academic Press.

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

Gabry, Jonah et al. 2024. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot/.

"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.

Helliwell, John F., Richard Layard, Jeffrey Sachs, and Jan-Emmanuel De Neve. 2020. *World Happiness Report 2020*. New York: Sustainable Development Solutions Network. https://worldhappiness.report/ed/2020/.

Layard, Richard. 2005. "Happiness: Lessons from a New Science." *Foreign Affairs* 84 (January). https://doi.org/10.2307/20031793.

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.

Wickham, Hadley, Jim Hester, and Kirill Müller. 2023. *Testthat: Unit Testing for r*. https://CRAN.R-project.org/package=testthat.