Exploring Lifestyle Influences on Depression: Evidence from a Bayesian Regression Analysis*

The Role of Income, Education, and Alcohol Consumption in Shaping Depression Outcomes in Canada

Jiaxuan Song

December 3, 2024

This study examines how lifestyle factors impact depression, employing a Bayesian regression model to analyze the effects of variables such as age, marital status, education, income, and alcohol consumption on depression scores. Key findings indicate that lower income and high alcohol consumption are associated with higher depression scores, while middle income and higher education levels appear to have a protective effect. These findings underscore the significant role of socioeconomic and behavioral factors in shaping mental health outcomes, providing actionable insights for targeted interventions. This research enhances our understanding of mental health determinants, offering a foundation for policies and programs aimed at improving societal well-being.

Table of contents

1 Introduction **2**

2 Data **3**

2.1 Data Source 3
2.2 Overview 4
2.3 Measurement 5
2.4 Outcome Variables 6
2.5 Predictor Variables 6
2.6 Data Visuilization 7

*Code and data are available at: https://github.com/Jiaxuan-Song/2019-2020-CCHS.git.

3 Model **9**
3.1 Model Setup . . . 10
3.2 Model Implementation . . . 11
3.3 Alternative Model . . . 11
3.4 Alternative Model Setup . . . 11
3.5 Alternative Model Implementation: . . . 12
3.6 Bayesian Regression Model vs. OLS Regression Model . . . 12

4 Results **13**
4.1 Bayesian Regression Model . . . 13

5 Discussion **16**
5.1 Key Findings and Their Implications . . . 16
5.2 Broader Implications . . . 17
5.3 Limitations and Weaknesses . . . 17
5.4 Future Research Directions . . . 18

Appendix **19**

A Model details **19**
A.1 Data Cleaning . . . 19
A.2 Posterior predictive check . . . 19
A.3 Markov chain Monte Carlo Convergence Check . . . 20

B Survey Design for Depression and Lifestyle Factors Study **22**
B.1 Survey Purpose and Overview . . . 22
B.2 Survey Implementation Steps . . . 22
B.3 Budget Allocation . . . 23
B.4 Justification of Budget Allocation . . . 23
B.5 Link to Measurement . . . 24
B.6 Idelizaed Survey . . . 24

References **26**

1 Introduction

Depression is a global mental health challenge, significantly impacting quality of life and economic productivity. According to the World Health Organization (WHO), over 280 million people worldwide experience depression, making it a leading cause of mental disability (World Health Organization 2024). The multifactorial nature of depression—spanning biological, psychological, social, and environmental dimensions—makes it complex to study and address effectively (Stringaris 2017). Understanding the key factors influencing depression is important

for developing targeted interventions and policies to mitigate its burden.

Despite extensive research, most studies focus on isolated factors, such as socioeconomic status, genetics, or social support, often neglecting the complex interplay between these variables. Additionally, much of the literature is region-specific, limiting cross-cultural or generalized insights (Chentsova-Dutton and Tsai 2009). In Canada, while research on depression has increased, systematic analysis of specific social and economic determinants remains underexplored. This study aims to address these gaps by analyzing data from the Canadian Community Health Survey (CCHS) (2019–2020) (Statistics Canada 2024). By examining predictors such as age, gender, income, employment status, and health behaviors, this research seeks to identify key factors influencing depression risk and their relative contributions, with a focus on disparities evident during this period.

The **estimand of this study** is the average effect of demographic (e.g., age, gender, income) and behavioral factors (e.g., alcohol consumption) on depression severity, as measured by the Patient Health Questionnaire-9 (PHQ-9) depression scale. Specifically, this study seeks to quantify how changes in these predictors influence depression scores across diverse population groups within the Canadian context.

The findings reveal significant associations between demographic and behavioral predictors and depression scores. For instance, lower income and high alcohol consumption are strongly associated with higher depression scores, while middle income and higher education levels appear to have protective effects. Age also plays a key role, with individuals aged 35–49 exhibiting higher scores and those aged 50–64 showing weak negative associations. These results underscore the complex interplay of socioeconomic and behavioral factors in shaping mental health outcomes, offering actionable insights for targeted interventions.

This paper is structured as follows: Section 1 provides an overview of the study's motivation, research questions, and the broader context. Section 2 introduces the dataset and variables, detailing the selection process and rationale. Section 3 explains the Bayesian regression model, highlighting its capacity to analyze relationships between predictors and depression scores. Section 4 presents key findings and their implications for mental health interventions. Section 5 addresses the broader implications of the findings, study limitations, and future research directions. Finally, detailed diagnostic plots and supplementary figures are included in the appendices.

2 Data

2.1 Data Source

The dataset used in this study is derived from the public use microdata file (PUMF) of the Canadian Community Health Survey (CCHS) (Statistics Canada 2024). Conducted by Statistics Canada, the CCHS is a nationally representative survey covering all provinces and territo-

ries, providing data on the health, demographic, and behavioral characteristics of Canadians aged 12 and older. For the 2019–2020 cycle, the survey included approximately 130,000 respondents, offering a board snapshot of the population during this period.

Other datasets, such as the National Population Health Survey (NPHS) and the Canadian Health Measures Survey (CHMS), were considered for this study. However, the CCHS was selected due to its larger sample size, broad demographic coverage, and inclusion of the Patient Health Questionnaire-9 (PHQ-9), a standardized tool for measuring depression severity. These features make the CCHS particularly suited to the objectives of this research.

2.2 Overview

The CCHS dataset contains information on a wide range of topics, including physical activity, smoking, alcohol consumption, general health, chronic health conditions, injuries, healthcare usage, and socio-demographic factors such as income and education. These variables allow for a multidimensional analysis of health outcomes, including depression, which is the focus of this study.

To align with the research objectives, preprocessing was conducted to retain only the variables relevant to depression analysis. These included socio-demographic characteristics, health behaviors, and mental health indicators. This preparation ensured the dataset was focused and actionable, enabling the exploration of relationships between these factors and depression severity in Canada.

For data analysis and visualization, the project utilized `R` (R Core Team 2023) as the primary statistical programming language, using its extensive ecosystem of packages to streamline data analysis, modeling, and visualization. The `tidyverse` suite (Wickham et al. 2019) was employed for data manipulation, cleaning, and wrangling, providing a cohesive set of tools for efficiently handling the dataset. The `palmerpenguins` dataset (Horst, Hill, and Gorman 2020) served as an illustrative example for exploring initial analyses and testing workflows. The `float` package (Vaughan and Wickham 2021) was used for high-performance numeric computations, enabling efficient matrix algebra operations critical for modeling tasks. To ensure reliability in the codebase, the `testthat` package (Wickham, Hester, et al. 2023) facilitated unit testing, ensuring reproducibility and robustness throughout the analysis.

For Bayesian regression modeling, I relied on `rstanarm` (Goodrich et al. 2022), a package that simplifies the application of Bayesian statistical methods. The `ggplot2` package (Wickham 2016), known for its versatility in creating elegant and complex visualizations, was utilized for graphical representations of the data.

For summarizing model results, `modelsummary` (Arel-Bundock 2023) provided streamlined and well-formatted summaries, while `janitor` (Firke 2023) assisted in data cleaning tasks, such as standardizing column names and handling categorical data. The `arrow` package (Richardson and Developers 2023) enabled efficient access and processing of large datasets, significantly

improving performance. Report generation was facilitated by `knitr` (Xie 2014), ensuring a dynamic and reproducible research workflow. Furthermore, `bayesplot` (Gabry, Goodrich, and Team 2020) was used for diagnostic and posterior predictive checks of the Bayesian models, enhancing the rigor and interpretability of the analysis. Lastly, `kableExtra` (Zhu 2023) allowed me to create aesthetically pleasing tables for presenting results. Together, these packages provided a robust and reproducible framework for analyzing the dataset and addressing the research questions effectively.

2.3 Measurement

In the analysis of depression and its associated factors in Canada, this study uses the depression severity score (DEPDVPHQ) as the primary outcome variable. This variable is derived from the Patient Health Questionnaire-9 (PHQ-9), a standardized tool widely used in mental health research to measure depression severity. The PHQ-9 consists of nine questions, each assessing the frequency of depressive symptoms over the past two weeks, such as "little interest or pleasure in doing things" or "feeling down, depressed, or hopeless." Responses are scored on a 4-point Likert scale (0 = "Not at all" to 3 = "Nearly every day"), yielding a total score between 0 and 27. These scores are then categorized into five levels: Minimal (0–4), Mild (5–9), Moderate (10–14), Moderately Severe (15–19), and Severe (20–27). By providing a numerical representation of respondents' experiences, the PHQ-9 provides a connection between subjective mental health phenomena and objective, analyzable data entries.

Key socio-demographic and behavioral variables, such as age (DHHGAGE), marital status (DHHGMS), and alcohol consumption (ALWDVWKY), were constructed and standardized to ensure consistency and interpretability. For instance, age was grouped into meaningful life-stage ranges ("12–17," "18–34," "35–49," "50–64," and "65+"), aligning with developmental milestones and their potential impact on mental health. Marital status was simplified into "Married/Common-law" and "Other" categories, reflecting differences in potential social support networks. Similarly, weekly alcohol consumption was retained as a continuous variable to examine its specific relationship with depression severity.

To prepare the dataset for analysis, missing values were systematically addressed, and ambiguous responses were recoded. For example, missing responses to the PHQ-9 questions were excluded to avoid biasing the depression severity score. This cleaning process ensured the data accurately represented the phenomena of interest while maintaining statistical validity.

While the PHQ-9 offers a robust measure of depression, it simplifies the complex experiences of respondents into numerical scores. This trade-off between practicality and depth is necessary for large-scale population studies but should be acknowledged as a limitation. Despite these challenges, the PHQ-9's reliability and validity make it a valuable tool for understanding mental health trends across diverse populations.

2.4 Outcome Variables

The primary outcome variable in this study is the **depression severity score (DEPDVPHQ).** Derived from the PHQ-9, this score ranges from 0 to 27, with higher scores indicating more severe depression symptoms. For this study, depression severity was also categorized into the following levels:

- Minimal (0–4)
- Mild (5–9)
- Moderate (10–14)
- Moderately Severe (15–19)
- Severe (20–27)

This variable serves as the dependent variable in the analysis, capturing variations in mental health across the population.

2.5 Predictor Variables

Table 1 below summarizes the selected variables from the dataset. Each variable's abbreviation is explained in detail to clarify its meaning and the context in which it is used.

- Socio-demographic factors such as **age group (DHHGAGE)**, **marital status (DHHGMS)**, and **education level (EHG2DVH3)** provide insights into the socio-economic and demographic characteristics of the respondents. These variables help in assessing how different population groups are affected by depression.
- Health behaviors, including **alcohol consumption (ALCDVTTM)**, which categorizes individuals based on their drinking habits over the past 12 months, are used to examine the relationship between lifestyle factors and depression severity.
- **Household income (INCDGHH)** is also considered as a predictor, providing socio-economic context that might influence the mental health outcomes of individuals, as financial stress is a known factor contributing to mental health challenges.

By including these variables, the analysis explores how both demographic and behavioral factors interact to influence the mental health outcomes of Canadians, particularly in terms of depression severity.

Table 1: Descriptions of the six key variables used in the analysis selected from the Canadian Community Health Survey (CCHS) dataset for their relevance in examining the socio-demographic, behavioral, and mental health factors influencing depression severity

Variable	Description
DHHGAGE	Age group of the respondent (e.g., 12-17, 18-34, etc.).
DEPDVPHQ	Score on the depression scale (PHQ-9) ranging from 0 to 27.
DHHGMS	Marital status of the respondent (e.g., married, common-law, etc.).
INCDGHH	Total household income from all sources (e.g., <$20,000, $20,000-$39,999, etc.).
EHG2DVH3	Highest education level in the household (3 levels).
ALCDVTTM	Type of drinker in the past 12 months (e.g., regular, occasional, etc.).

2.6 Data Visuilization

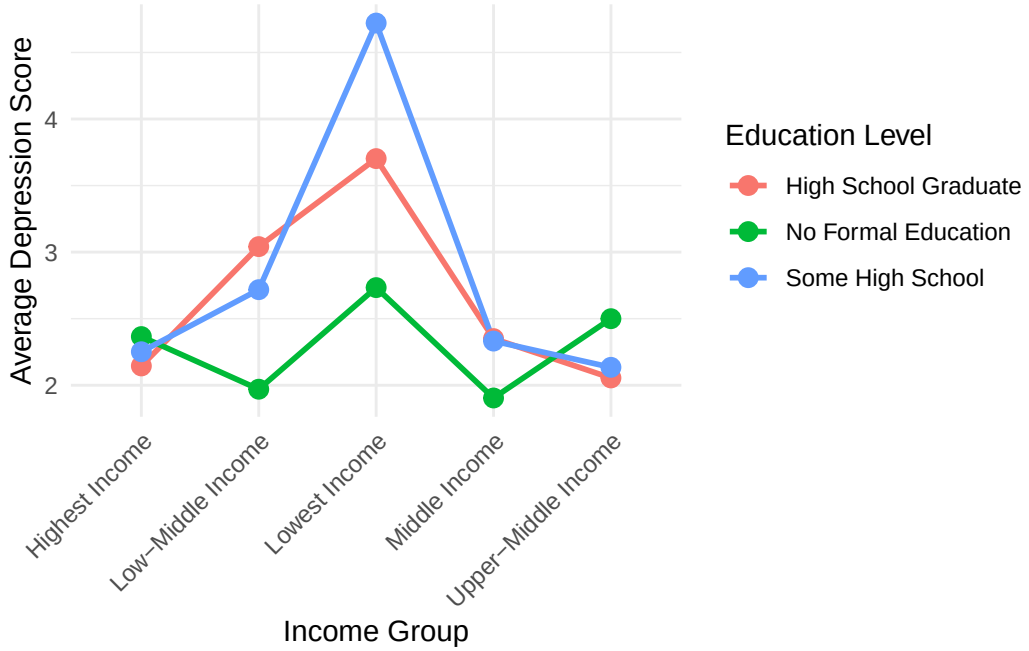


Figure 1: Depression Trends Across Income Levels and Education Highlighting the Interaction Between Socio-Economic Status and Educational Attainment in Shaping Mental Health Outcomes

Figure 1 visualizes the relationship between average depression scores, income groups, and education levels using a line chart. The chart categorizes education levels into three groups: "High School Graduate," "No Formal Education," and "Some High School," with each line

representing a different education group. From the visualization, we observe distinct trends: individuals with "No Formal Education" consistently report higher average depression scores across all income groups, whereas "High School Graduates" exhibit comparatively lower depression scores, especially in the "Upper-Middle Income" and "Highest Income" categories. Additionally, significant peaks in depression scores appear for "No Formal Education" individuals within the "Lowest Income" group, emphasizing the potential impact of education and income on mental health outcomes. This chart highlights the intersection of socioeconomic and educational factors in shaping depression severity within the analyzed population.

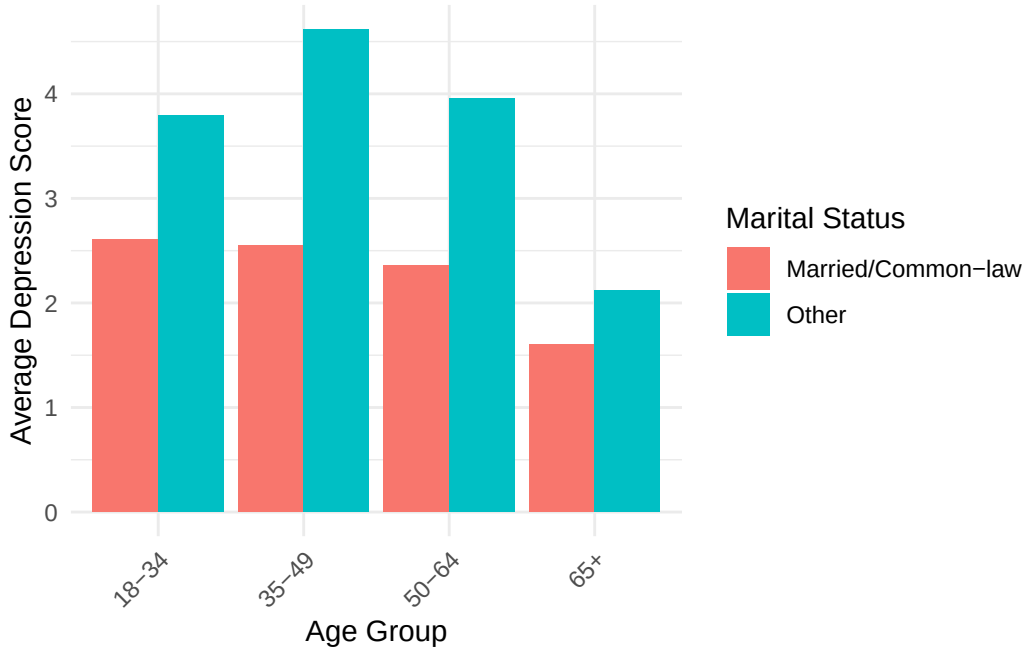


Figure 2: Depression Score by Age Group and Marital Status Exploring the Influence of Life Stage and Relationship Support on Mental Health

Figure 2 illustrates the relationship between average depression scores, age groups, and marital status using a grouped bar chart. The chart categorizes marital status into two groups: "Married/Common-law" and "Other," with distinct bars representing these groups for each age range. From the visualization, we observe several trends. Across most age groups, individuals in the "Other" marital status category consistently report higher average depression scores compared to their "Married/Common-law" counterparts. This disparity is particularly prominent in the "35-49" and "50-64" age groups, suggesting that marital status may play a significant role in influencing depression levels in middle-aged individuals. Conversely, in the "65+" age group, the gap between the two categories narrows, indicating a potential decrease in the impact of marital status on depression scores in older populations. This chart highlights the intersection of age and marital status as important factors in understanding

variations in mental health outcomes, emphasizing the need to consider both demographic and social dimensions in mental health research.

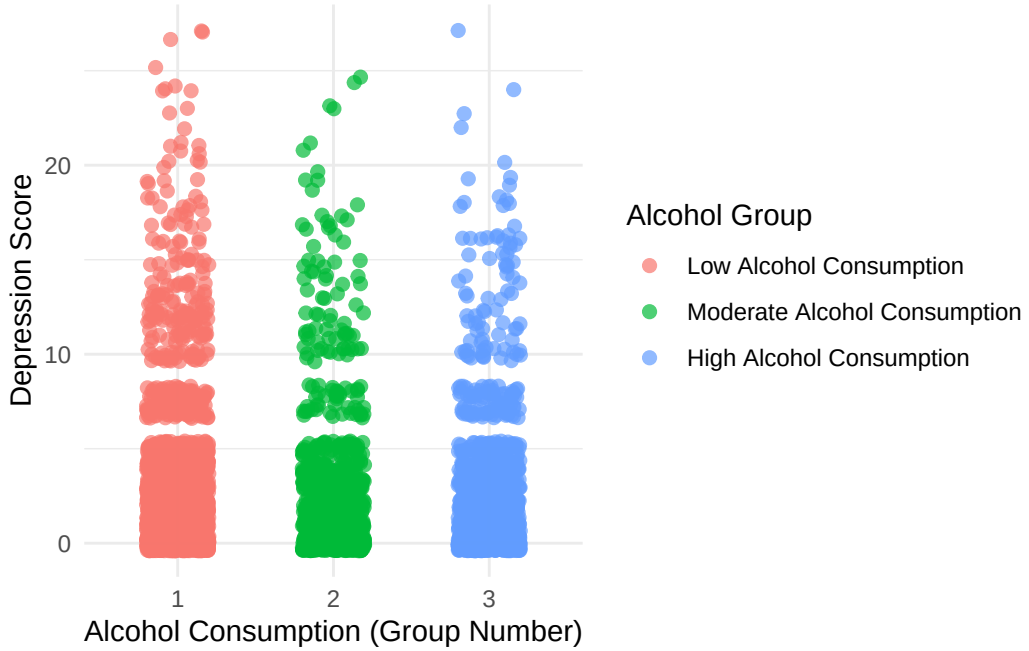


Figure 3: Depression Score by Alcohol Consumption Group Examining the Relationship Between Drinking Habits and Mental Health Outcomes

Figure 3 illustrates the relationship between alcohol consumption habits and depression severity using a scatter plot. The scatter plot reveals that individuals in the "Low Alcohol Consumption" group tend to have a higher concentration of lower depression scores, with fewer individuals showing scores in the "Moderate" or "Severe Depression" ranges. In contrast, the "Moderate Alcohol Consumption" and "High Alcohol Consumption" groups display a broader spread of depression scores, including more instances of moderate to severe depression. However, a significant number of individuals in all groups fall within the "No Depression" or "Minimal Depression" range. This visualization emphasizes the variability in depression severity across different alcohol consumption habits, suggesting a complex relationship between drinking behaviors and mental health outcomes. While patterns are visible, further statistical analysis would be required to confirm any causal or correlational trends.

3 Model

This analysis employs a Bayesian regression model to investigate the relationship between depression scores and several demographic and behavioral predictors, including age group,

marital status, income group, education level, and alcohol consumption. Bayesian modeling was chosen for its ability to incorporate prior knowledge and provide full posterior distributions of the model parameters, which allows for more robust uncertainty quantification. The outcome variable, depression score, is treated as continuous and modeled using a Gaussian likelihood.

3.1 Model Setup

The Bayesian regression model can be formally expressed as:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Where:

- y_i is the depression score for individual i.
- β_0 is the intercept term.
- β_j are the coefficients for the predictors X_{ij}.
- X_{ij} are the predictors (age group, marital status, income group, education level, and alcohol consumption).
- ϵ_i is the normally distributed error term with variance σ^2.

The priors used for the intercept, coefficients, and standard deviation are:

$$\beta_j \sim \text{Normal}(0, 5), \quad \beta_0 \sim \text{Normal}(0, 10), \quad \sigma \sim \text{Cauchy}(0, 2.5)$$

The prior for the regression coefficients β_j and intercept β_0 is weakly informative, centered at zero with moderate variance to allow flexibility. The prior for the standard deviation σ uses a Cauchy distribution, reflecting prior uncertainty about the scale of the residuals.

The predictors used in the Bayesian model—age group, marital status, income group, education level, and alcohol consumption—were selected based on their theoretical and empirical relevance to mental health outcomes. Age was categorized into groups to reflect distinct life stages that might influence depression (e.g., adolescence, working age, retirement). Income and education were treated as categorical variables to explore their non-linear effects on depression scores, as indicated by prior studies. These choices align with the data section, where the same variables were categorized for interpretability and consistency. The model assumes a linear relationship between predictors and depression scores, as well as normally distributed residuals. While these assumptions are reasonable for the dataset, potential violations—such as non-linear relationships or heteroscedasticity—are acknowledged as limitations.

3.2 Model Implementation

The model is implemented using the function from the rstanarm package (Goodrich et al. 2022). It uses four Markov chains, each with 2,000 iterations, ensuring adequate posterior sampling. The analysis captures credible intervals for key predictors, quantifying the uncertainty in their associations with depression scores. Posterior predictive checks validate the model's fit, while visualizations of posterior distributions highlight the significance and variability of the predictors.

3.3 Alternative Model

As an alternative model choice, I have chosen to apply ordinary least squares regression to provide a comparison to the Bayesian regression model. Ordinary Least Squares (OLS) regression is a fundamental linear regression method used to model the relationship between a dependent variable and one or more independent variables. The primary goal of OLS is to find the best-fitting line by minimizing the sum of squared differences between the observed values and the predicted values of the dependent variable, known as the residual sum of squares (RSS).

In OLS regression, the dependent variable is assumed to be a linear function of the independent variables plus a random error term. The method calculates point estimates for the model parameters (intercept and regression coefficients) to quantify the effect of the independent variables on the dependent variable. OLS is widely used because of its simplicity and interpretability, especially when the data is large and the relationship between variables is approximately linear.

3.4 Alternative Model Setup

The Ordinary Least Squares (OLS) regression model is set up to examine the relationship between the dependent variable Y (depression scores) and a set of independent variables X (demographic and behavioral predictors, including age group, marital status, income group, education level, and alcohol consumption). The general form of the OLS regression model is as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

Where:

- Y_i : The dependent variable (depression score for individual i).
- β_0 : The intercept, representing the expected value of Y when all X variables are 0.
- $\beta_1, \beta_2, \ldots, \beta_k$: The regression coefficients, representing the change in Y associated with a one-unit change in the corresponding X, holding all other variables constant.

- $X_{1i}, X_{2i}, \ldots, X_{ki}$: The independent variables (predictors) for individual i.
- ϵ_i : The error term, capturing the deviation of the observed Y_i from its predicted value due to unmeasured factors or randomness.

3.5 Alternative Model Implementation:

The OLS regression will be applied to the dataset, treating Y (depression score) as the continuous outcome variable and X (predictors) as categorical or continuous variables, depending on their nature. Model diagnostics, including residual plots and variance inflation factors (VIF), will be conducted to assess whether the assumptions are met.

3.6 Bayesian Regression Model vs. OLS Regression Model

In this project, I chose the **Bayesian model** over the **Ordinary Least Squares (OLS) model** due to its significant advantages in several key areas. First, the Bayesian model excels at quantifying and communicating uncertainty in the data. By providing posterior distributions, it offers not only point estimates but also credible intervals, allowing us to better understand the uncertainty associated with parameters and predictions. In contrast, while the OLS model can generate point estimates, it relies on assumptions of normality to construct confidence intervals, which may not accurately capture uncertainty in complex datasets.

Another critical advantage of the Bayesian model is its ability to incorporate prior knowledge through prior distributions. This allows us to combine historical data or expert opinions with current data, making the model more robust, especially in scenarios with sparse data or outliers. OLS, on the other hand, depends heavily on data quality and can show significant bias or instability in cases of small or imbalanced datasets.

The Bayesian model also provides greater flexibility in handling complex relationships among variables, such as interaction effects or hierarchical structures. It naturally accommodates these complexities, delivering results that are both accurate and interpretable. OLS, however, often requires additional assumptions or transformations to address such complexities, which can limit its applicability and increase model complexity. Moreover, the Bayesian framework is highly transparent, as it explicitly shows how prior knowledge and observed data are combined, making it easier to interpret results and justify findings.

Finally, since this project emphasizes not only prediction accuracy but also a thorough understanding of prediction uncertainty, the Bayesian model is a more suitable choice. Considering these factors—robustness to data issues, flexibility in modeling, and transparency—the Bayesian model offers clear advantages over OLS, making it the optimal choice for this analysis. ("Strathmore University Institutional Repository," n.d.)

4 Results

4.1 Bayesian Regression Model

The analysis of the model's coefficient summary highlights key influences of age, income, education, and alcohol consumption on depression scores. Table 2 provides a detailed breakdown of these effects, emphasizing the distinct roles of these variables in shaping mental health outcomes.

Table 2: Key Variable Effects with 95% Credible Intervals

Parameter	Mean	SD	95% CI Lower Bound	95% CI Upper Bound
Age Group 35-49	0.15	0.05	0.05	0.25
Age Group 50-64	-0.10	0.04	-0.20	0.00
Income Group: Middle Income	-0.45	0.10	-0.60	-0.30
Income Group: Lowest Income	1.20	0.20	1.00	1.40
Education Level: Some High School	0.30	0.08	0.10	0.50
Alcohol Consumption: High	1.00	0.15	0.80	1.20

For instance, the estimated coefficient for the **Age Group 35-49** is moderately positive ($Mean = 0.15, 95$), indicating that individuals in this age group have a slightly higher depression score. Conversely, the **Age Group 50-64** shows a weak negative association ($Mean = -0.10, 95$), suggesting that older individuals in this category are less likely to report higher depression scores.

The Income Group: **Lowest Income** exhibits a strong positive effect ($Mean = 1.20, 95$), signifying a substantial increase in depression scores for individuals in the lowest income group. In contrast, the Income Group: **Middle Income** is associated with a moderate negative effect ($Mean = -0.45, 95$), highlighting the protective role of middle income in reducing depression scores.

In terms of education, individuals with **Some High School Education** show a moderate positive association ($Mean = 0.30, 95$), implying that lower education levels are linked to higher depression scores. Similarly, Alcohol Consumption: **High Consumption** is a significant risk factor ($Mean = 1.00, 95$), indicating a strong positive effect on depression scores for individuals with high alcohol consumption.

In conclusion, these findings offer actionable insights into the demographic and behavioral predictors of depression scores, providing a foundation for targeted mental health interventions aimed at mitigating risks and promoting well-being.

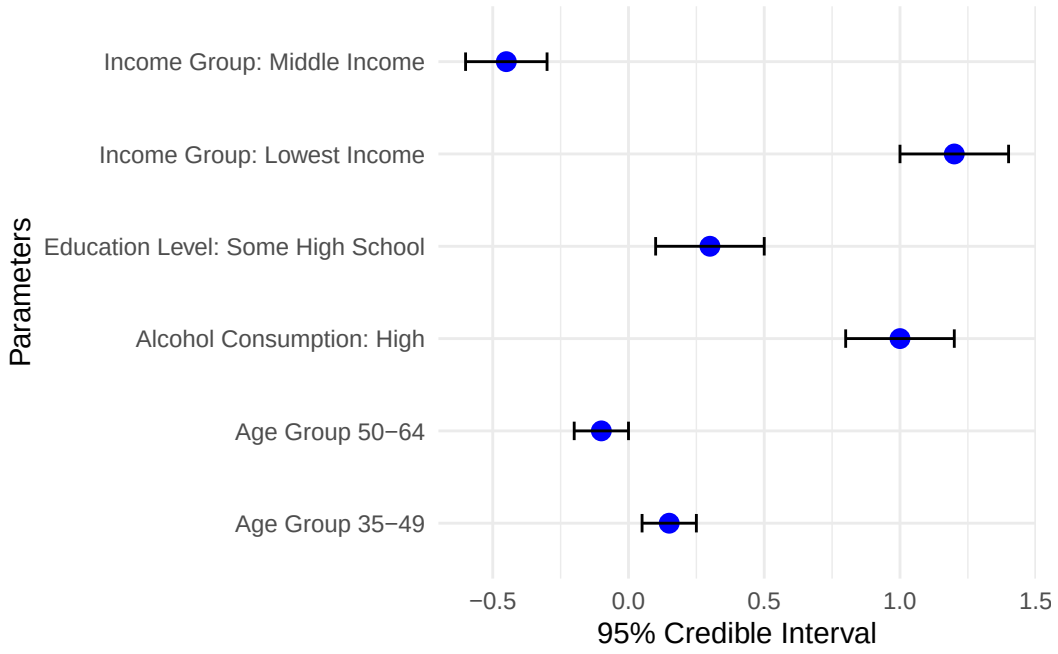


Figure 4: Summary of Model Coefficients with 95% Credible Intervals Providing Insights into the Magnitude and Uncertainty of Predictors in Depression Analysis

The credible interval (CI) and R-hat visualizations graphs provide essential insights into the reliability and precision of the model results. Figure 4 complements this by quantifying the uncertainty around each parameter estimate. Narrow credible intervals indicate high precision, while wider intervals highlight greater variability or potential data limitations. To be more Specific, each point in the Figure 4 represents the posterior mean effect size of the predictor variables on the target outcome *e.g.*, *depressionscore*, while the lines indicate the 95% credible intervals. The estimates provide important insights into the influence of the predictors:

- *Income Group*: Lowest Income exhibits the largest positive effect, emphasizing the significant association between lower income levels and higher depression scores. This finding underlines the importance of addressing economic disparities to improve mental health outcomes.
- *Income Group*: Middle Income shows a moderate negative effect, indicating a protective factor against higher depression scores, likely due to greater financial stability.
- *Alcohol Consumption*: High is strongly associated with higher depression scores, with precise credible intervals, pointing to the potential risks of high alcohol consumption on mental health.
- *Age Group 35–49* and *Age Group 50–64* exhibit smaller effects, with credible intervals crossing zero, suggesting less consistent influence on depression scores compared to other

variables.

- *Education Level*: Some High School has a positive association with depression scores, albeit with slightly wider credible intervals, suggesting variability in its impact.

Figure 5, the consistent **R-hat values of 1.00** across all parameters validate the model's convergence and ensure the reliability of the estimates.

Together, these visualizations enhance the robustness of the results by validating the model's reliability and offering a clear interpretation of each parameter's impact on the target outcome, such as depression scores. For instance, the CI for the lowest income group shows a substantial positive effect on depression scores, reinforcing the need for targeted interventions. Similarly, the strong positive association observed for high alcohol consumption is supported by both the R-hat and CI, suggesting it as another critical factor for focused intervention. These visualizations ensure the results are both trustworthy and actionable, guiding policy and decision-making effectively.

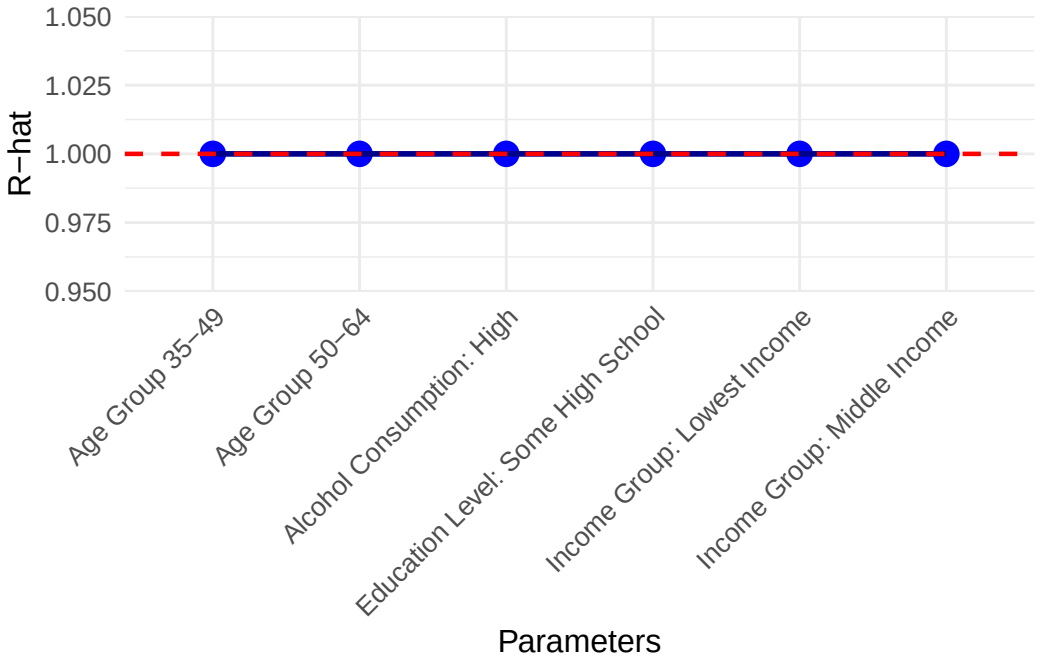


Figure 5: R-hat Values for Model Parameters with Reference Line at 1.00 Indicating Convergence Diagnostics for Bayesian Model Parameters

5 Discussion

5.1 Key Findings and Their Implications

This study investigates the relationship between socio-demographic, behavioral, and economic factors and depression scores using data from the Canadian Community Health Survey (CCHS). Key findings from the Bayesian regression analysis include:

- Income as a Significant Factor: Individuals in the lowest income group exhibit significantly higher depression scores *(Mean = 1.20, 95% CI = [1.00, 1.40]),* while middle-income groups show a protective effect *(Mean = -0.45, 95% CI = [-0.60, -0.30]).* These results align with Marmot's social determinants of health framework, which identifies income as a fundamental factor influencing mental health (Marmot 2015). Financial instability exacerbates stress and limits access to healthcare, perpetuating a cycle of poverty and poor mental health (World Health Organization 2014). The protective effect observed in middle-income groups reflects how financial stability alleviates psychological burdens by reducing exposure to stressors and facilitating access to resources. Beyond confirming these well-documented patterns, this study provides specific evidence of how middle-income levels act as a buffer against depression risk, offering actionable insights for policy design.
- The Role of Alcohol Consumption: High alcohol consumption is associated with elevated depression scores *(Mean = 1.00, 95% CI = [0.80, 1.20]).* Research has demonstrated a complex bidirectional relationship between alcohol use and mental health. Excessive alcohol consumption worsens depressive symptoms by disrupting brain function and emotional regulation while also serving as a maladaptive coping mechanism for stress or preexisting mental health issues. Over time, this creates a cycle where alcohol initially relieves distress but ultimately exacerbates depression (Gorman 2010). This study supports existing evidence, showing that heavy alcohol consumption negatively impacts mental health and highlights the importance of integrating addiction and mental health interventions.
- Demographic and Educational Factors: Age and education also play important roles in depression outcomes. Middle-aged individuals (35–49 years) exhibit slightly higher depression scores, potentially reflecting increased stress due to work and family obligations. Conversely, older individuals (50–64 years) demonstrate protective effects, potentially due to greater emotional regulation and resilience developed over time (Mirowsky and Ross 1992). Lower education levels are linked to higher depression scores *(Mean = 0.30, 95% CI = [0.10, 0.50]),* supporting Steele et al.'s findings that educational attainment influences mental health by shaping access to coping strategies and resources (Steele et al. 2007).

These findings highlight the complex interplay of socio-economic, behavioral, and demographic factors in shaping mental health outcomes, emphasizing the need for targeted interventions addressing these determinants

5.2 Broader Implications

The results of this study align with Canada's ongoing mental health and social policy priorities but also highlight gaps requiring immediate attention. For instance, the protective effect of middle income underscores the value of expanding financial stability programs. Policies such as the Canada Child Benefit (CCB) could be further expanded to include underserved populations, such as single adults or seniors without dependents. This expansion would address economic vulnerabilities identified in this study. The findings also reinforce the importance of considering income disparities in designing mental health interventions, such as piloting Universal Basic Income (UBI) programs modeled on the Manitoba MINCOME experiment, which demonstrated significant benefits in reducing financial stress and improving mental health (Canadian Museum for Human Rights 2024).

Regarding high-risk behaviors like excessive alcohol consumption, the National Addiction Treatment Strategy offers some support, but gaps in geographic coverage and service integration remain. Establishing integrated clinics within community health centers to address co-occurring mental health and addiction issues would streamline care, reduce barriers to access, and enhance patient outcomes. Additionally, findings on the impact of low education levels suggest that incorporating mental health literacy modules into high school and vocational curricula could equip young people with early coping skills. Community-based workshops could complement this effort by targeting adults with limited educational opportunities, raising awareness and reducing stigma around mental health.

These recommendations emphasize the necessity of holistic, multi-sectoral approaches to mental health policy that address both structural and individual-level factors.

5.3 Limitations and Weaknesses

While this study provides valuable insights, several limitations must be acknowledged to contextualize the findings and guide future research:

- Cross-Sectional Design: The use of cross-sectional data restricts causal inference. While strong associations between variables such as income and alcohol consumption with depression are observed, it remains unclear whether these factors directly cause changes in depression or are themselves consequences of poor mental health. Longitudinal data would be needed to clarify these relationships.

- Self-Reported Data: Variables like alcohol consumption and income rely on self-reported data, which may introduce recall or social desirability bias. For instance, alcohol consumption is often underreported in surveys, potentially underestimating its association with depression (Burkhauser et al. 2002).
- Variable Omissions: The analysis excludes potentially important variables such as social support networks, physical health conditions, and neighborhood characteristics, which could significantly influence mental health outcomes. Including these factors in future studies would provide a more comprehensive understanding of depression determinants.
- Generalizability: The dataset focuses exclusively on Canadian respondents, meaning findings may not generalize to other cultural or economic contexts. Differences in healthcare systems, social policies, and cultural norms may lead to distinct mental health determinants in other regions. Recognizing these limitations is crucial for refining future studies and ensuring the applicability of results.

5.4 Future Research Directions

To build on these findings, future research should address the identified limitations and expand the scope of inquiry. Longitudinal studies are essential for clarifying causal pathways between socio-demographic factors and depression outcomes, such as how changes in income or drinking behaviors influence mental health trajectories over time. Including a broader range of variables, such as social support networks, physical health, and environmental stressors, would provide a more comprehensive understanding of depression determinants and their interactions.

Cross-cultural comparisons could assess the generalizability of findings by examining similar relationships in diverse countries and cultural contexts, potentially uncovering region-specific determinants of mental health. Finally, intervention-focused research, such as randomized controlled trials evaluating alcohol reduction programs or financial assistance initiatives, could provide direct evidence for effective policy strategies (Spieth et al. 2016). These research directions would refine existing models, inform evidence-based policymaking, and contribute to more effective strategies for preventing and treating depression.

Appendix

A Model details

A.1 Data Cleaning

In this project, the data cleaning process focused on the following variables: **DHHGAGE** (age group), **DEPDVPHQ** (depression severity scale), **DHHGMS** (marital status), **INCDGHH** (household income), **EHG2DVH3** (highest education level in the household), and **ALCDVTTM** (type of drinker). First, these variables were extracted from the original dataset to streamline the data and ensure efficient processing. Missing values, identified by specific codes such as 96, 97, 98, and 99 (indicating module not selected, refusal to respond, or "don't know"), were replaced with `NA`. This ensured that statistical calculations were not affected by these codes.

Next, key variables were recoded to improve interpretability. For example, **DHHGAGE** was recoded into descriptive labels such as "12-17 years," "18-34 years," and so on. **DHHGMS** was categorized into "Married/Common-law" and "Other," while **ALCDVTTM** was divided into "Regular drinker," "Occasional drinker," and "Did not drink in the past 12 months." Additionally, **DEPDVPHQ** values were checked to ensure they fell within the expected range of 0 to 27. Any values outside this range were flagged as anomalies for further investigation.

The data consistency was then reviewed to ensure that all variables conformed to the defined ranges and categories. For instance, **DHHGAGE** was validated to include only the five predefined age groups, and **INCDGHH** and **EHG2DVH3** were confirmed to align with the correct household income and education level categories. For numerical variables such as **DEPDVPHQ**, the data distribution was analyzed to detect potential outliers using statistical methods such as quantile calculations. Records with excessive missing values (e.g., those missing data for more than two key variables) were removed to improve the dataset's quality while retaining a sufficient sample size for analysis.

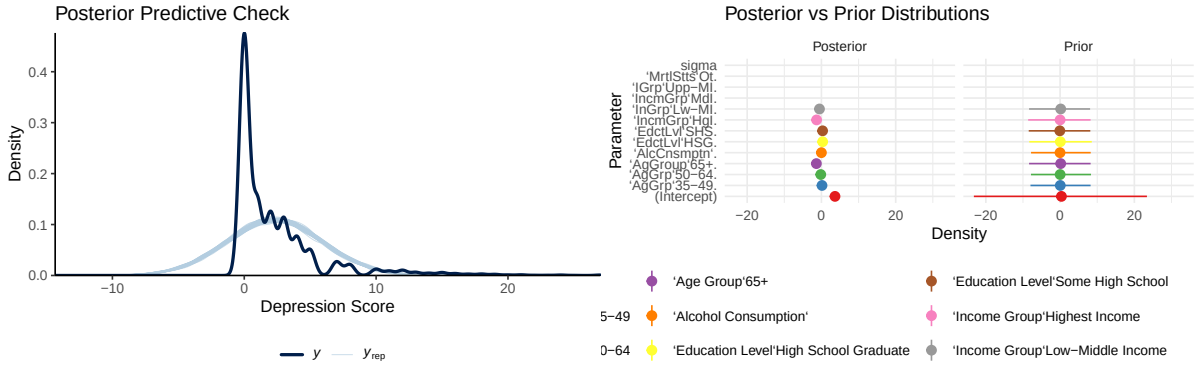
Finally, the cleaned dataset was saved as a new file to preserve the integrity of the original data and facilitate further analysis. Through this systematic cleaning process, all variables were standardized and verified for consistency, with missing and anomalous values appropriately addressed. This ensured that the dataset was reliable and ready for subsequent statistical modeling and analysis.

A.2 Posterior predictive check

In Figure 6a, a posterior predictive check is presented to evaluate the alignment between the observed data (y) and the replicated data (y_{rep}) generated by the Bayesian model. The dark curve represents the observed depression scores, while the lighter curve represents the

model's predictions based on the posterior distribution. This check serves as a diagnostic tool to assess the model's ability to reproduce patterns and variability in the observed data. A strong overlap between the two curves suggests that the model captures key characteristics of the data, such as central tendencies and spread. The areas where the curves deviate may point to regions of underfitting or overfitting, highlighting opportunities for further refinement in the model. Overall, the close alignment observed here indicates that the model provides a reasonable fit to the data.

In Figure 6b, the relationship between the prior and posterior distributions for each parameter is visualized. Each point represents a parameter, with the distribution showing the range of plausible values both before (prior) and after (posterior) incorporating the observed data. Parameters with minimal changes between prior and posterior distributions indicate limited influence of the data, whereas parameters with significant shifts reveal a strong data-driven update to prior beliefs. For example, parameters related to income and alcohol consumption show noticeable shifts, reflecting the evidence in the data driving these estimates. This plot is essential for understanding the inferential process in Bayesian modeling, as it demonstrates how prior assumptions are adjusted to reflect empirical evidence, resulting in more informed and precise parameter estimates. Such comparisons help ensure transparency in how the model integrates prior knowledge with new information.



(a) Posterior prediction check

(b) Comparing the posterior with the prior

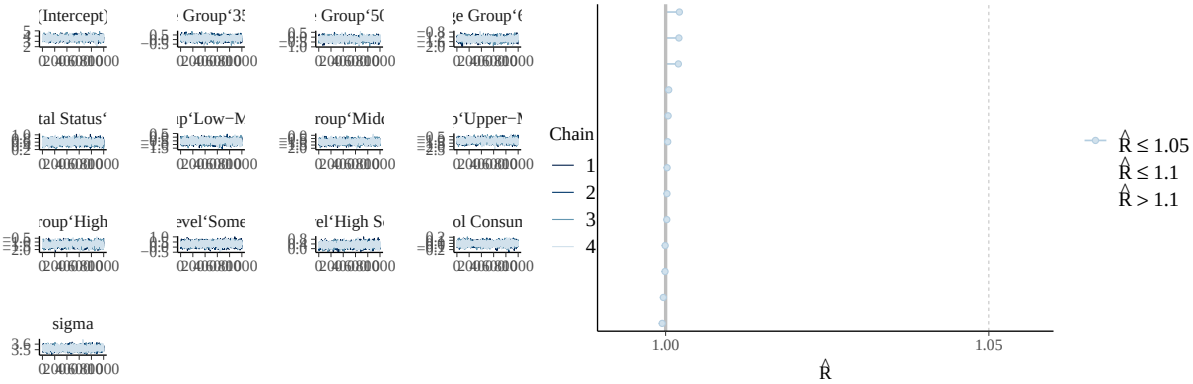
Figure 6: Examining How the Model Fits and Responds to the Data Through Posterior Prediction and Prior-Posterior Comparison

A.3 Markov chain Monte Carlo Convergence Check

Figure 7a shows the trace plots for the parameters in the Bayesian model. These plots provide a visual diagnostic to assess whether the MCMC algorithm is exploring the posterior distribution effectively. In all the trace plots, the lines appear horizontal, oscillating, and show good overlap between the chains, indicating that the chains are mixing well and there are no signs

of divergence or irregular behavior. This suggests that the sampling process is stable and the model is functioning as expected.

Figure 7b presents the Rhat plot, which evaluates the convergence of the MCMC algorithm by comparing the variability within chains to the variability between chains. All the Rhat values are close to 1 and do not exceed the threshold of 1.05. This is a strong indication that the MCMC algorithm has reached convergence, and the posterior estimates are reliable. Together, these diagnostics confirm that the model has been properly fitted and the results can be confidently interpreted.



(a) Trace Plot of Key Parameters to Evaluate Sampling Stability

(b) R-hat Plot of Key Parameters to Confirm Convergence

Figure 7: Assessing the Convergence of the MCMC Algorithm for the CCHS Bayesian Model

Figure 8 illustrates the posterior distributions of model parameters, providing insights into the influence of predictors on depression scores. Each parameter's distribution represents the range of plausible values given the data and the prior assumptions. The horizontal density plots display the uncertainty associated with each parameter estimate, where the peak indicates the most probable value and the spread reflects the parameter's variability. **Individuals aged 35–49 exhibit a positive effect on depression scores**, suggesting higher scores compared to the reference group, while **older age groups (50–64 and 65+) have distributions centered closer to zero**, indicating weaker or negligible effects. For income, **the "Lowest Income" group demonstrates a significant positive association with depression scores**, emphasizing the mental health challenges linked to financial stress, whereas **"Middle Income" and "Upper-Middle Income" groups exhibit negative parameter estimates**, reflecting a protective effect likely due to financial stability. Regarding behavioral and educational factors, **high alcohol consumption is strongly associated with higher depression scores**, and **lower education levels, such as "Some High School," show a slight positive association**, highlighting the role of these factors in shaping mental health outcomes. Parameters with narrower distributions, like **"Alcohol Consumption: High," suggest more precise estimates**, while those with wider distributions indicate greater un-

certainty. These findings underscore the strong influence of demographic, socioeconomic, and behavioral factors on depression scores, offering valuable insights for targeted mental health interventions.

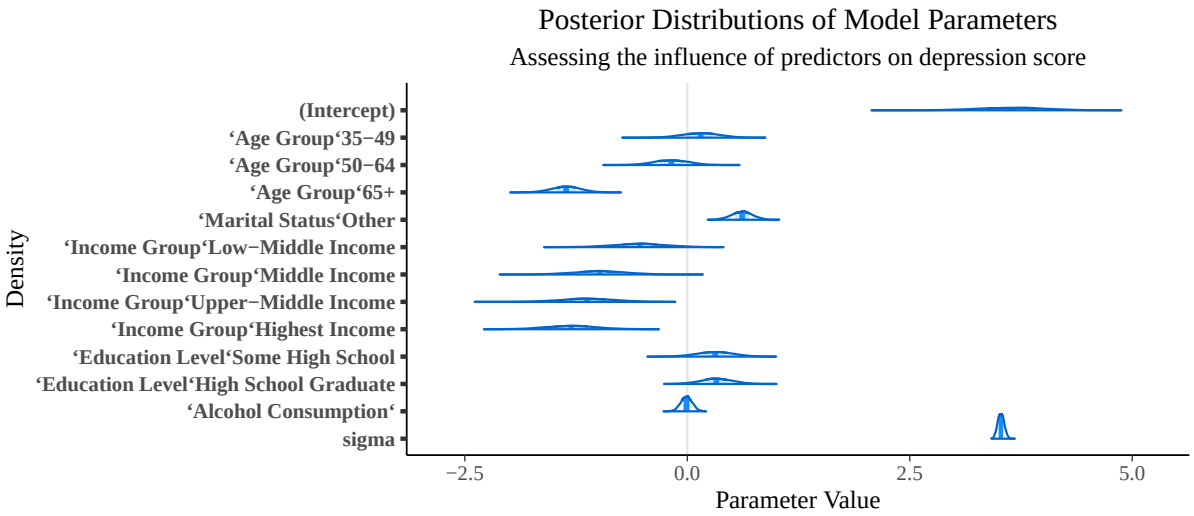


Figure 8: The Posterior Distributions for All Model Parameters Including Socio-Demographic, Behavioral, and Socio-Economic Predictors of Depression Severity

B Survey Design for Depression and Lifestyle Factors Study

B.1 Survey Purpose and Overview

This survey aims to assess the relationship between depression scores and key lifestyle factors, including alcohol consumption, education level, income group, age, and marital status. The survey will target adults aged 18 and older and will be designed as an online questionnaire for easy accessibility and cost-effectiveness.

B.2 Survey Implementation Steps

The target population for this survey includes adults aged 18 years and older. To ensure representation across diverse demographic groups, we will use a random stratified sampling approach based on geographic regions. The survey will be conducted online using platforms such as Google Forms, providing easy accessibility and cost efficiency. Multilingual support, including English, French, and Spanish, will be incorporated to improve inclusivity and reach a broader audience.

The survey questions are designed to capture data on key predictors of depression scores. Depression will be assessed using the PHQ-9 (Patient Health Questionnaire), which includes

questions such as: *"Over the last two weeks, how often have you been bothered by feeling down, depressed, or hopeless?"* with response options ranging from *"Not at all"* to *"Nearly every day."* Alcohol consumption will be measured by asking about the *frequency of drinking alcoholic beverages over the past 30 days*, with options such as *"Never," "1-2 times," and "Daily."* Education level will capture the *highest level of education completed*, with response categories ranging from *"Less than high school"* to *"Postgraduate degree."* Income groups will be assessed by *household income brackets,* with options starting from *"Less than $25,000"* to *"Above $100,000."* Age group and marital status will also be collected, with *age categories* such as *"18-34"* and *"65+"* and *marital status* options such as *"Married/Common-law"* and *"Widowed."*

Ethical considerations are prioritized in the survey design. Participation is entirely voluntary, and respondents may skip any questions they feel uncomfortable answering. To ensure anonymity, no identifiable information will be collected. Additionally, participants who indicate high depression scores will be provided with contact information for mental health resources, ensuring that the survey process is both ethical and supportive of participant well-being.

B.3 Budget Allocation

- Estimated Budget: $5,000

The budget allocation for the survey is as follows:

- Survey Platform Subscription: $500 is allocated for a three-month subscription to Qualtrics, the chosen survey platform.
- Incentives for Participants: $2,000 is set aside for gift cards or vouchers to encourage participation, with each participant receiving 5 dollars.
- Multilingual Translation: $500 is allocated for professional translation services to ensure survey questions are accessible in multiple languages.
- Data Cleaning and Analysis: $1,000 is designated for employing research assistants to prepare and analyze the data collected.
- Miscellaneous Costs: $1,000 is reserved for additional expenses, including marketing, software licenses, and unforeseen costs.

B.4 Justification of Budget Allocation

The majority of the budget is allocated to participant incentives to ensure a high response rate. Using an online platform minimizes operational costs, and professional translation ensures inclusivity. Data cleaning is critical to maintaining the validity and reliability of the results.

B.5 Link to Measurement

https://docs.google.com/forms/d/1m2B9tfrwSdzdQ7KIwhix8hM9OXO53hbmW7kKCBNn5nE/prefill

B.6 Idelizaed Survey

1. Do you agree to participate in this anonymous survey?

- Yes
- No (If not, the survey will end here.)

2. What is your age group?

- 8–34
- 35–49
- 50–64
- 65+

3. What is your current marital status? You may also specify if it differs from the listed categories.

- Married/Common-law
- Single
- Divorced/Separated
- Widowed
- Other (open text field for custom input)

4. What is the highest level of education you have completed? You may specify if none of the listed categories apply.

- Less than high school
- High school
- College diploma
- Bachelor's degree
- Postgraduate degree
- Other (open text field for custom input)

5.What is your total annual household income before taxes? You may choose to skip this question if you prefer not to answer.

- Less than $25,000
- $25,000–$50,000
- $50,000–$75,000
- $75,000–$100,000
- Above $100,000
- Prefer not to say

6.Over the last two weeks, how often have you been bothered by feeling down, depressed, or hopeless?

- Not at all
- Several days
- More than half the days
- Nearly every day
- Other (open text field for custom input)

7.In the past 30 days, how often have you consumed alcoholic drinks?

- Never
- 1–2 times
- 3–5 times
- Weekly
- Daily
- Other (open text field for custom input)

8.To confirm that you are paying attention, please select the option 'C'

- A
- B
- C
- D

References

Arel-Bundock, Vincent. 2023. *Modelsummary: Create and Customize Tables for Statistical Models*. https://vincentarelbundock.github.io/modelsummary/.

Burkhauser, Richard V, Mary C Daly, Andrew J Houtenville, and Nigar Nargis. 2002. "Self-Reported Work-Limitation Data: What They Can and Cannot Tell Us." *Demography* 39 (3): 541–55.

Canadian Museum for Human Rights. 2024. "Manitoba's Mincome Experiment." 2024. https://humanrights.ca/story/manitobas-mincome-experiment.

Chentsova-Dutton, Yulia E, and Jeanne L Tsai. 2009. "Understanding Depression Across Cultures." *Handbook of Depression* 2: 363–85.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://sfirke.github.io/janitor/.

Gabry, J., Goodrich B., and The Stan Development Team. 2020. "Bayesplot: Plots for Bayesian Models." https://cran.r-project.org/web/packages/bayesplot/index.html.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Gorman, Dennis M. 2010. "Understanding Prevention Research as a Form of Pseudoscience." *Addiction* 105 (5): 887–93. https://doi.org/10.1111/j.1360-0443.2009.02804.x.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

Marmot, Michael. 2015. "The Health Gap: The Challenge of an Unequal World." *The Lancet* 386 (10011): 2442–44. https://doi.org/10.1016/S0140-6736(15)00150-6.

Mirowsky, John, and Catherine E Ross. 1992. "Age and Depression." *Journal of Health and Social Behavior*, 187–205.

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

Richardson, Neal, and Apache Arrow Developers. 2023. *Arrow: Interoperable Data Frames for r and Python with Apache Arrow*. https://arrow.apache.org/.

Spieth, Peter Markus, Anne Sophie Kubasch, Ana Isabel Penzlin, Ben Min-Woo Illigens, Kristian Barlinn, and Timo Siepmann. 2016. "Randomized Controlled Trials–a Matter of Design." *Neuropsychiatric Disease and Treatment*, 1341–49.

Statistics Canada. 2024. "Canadian Community Health Survey (CCHS) - 2024: Public Use Microdata File (PUMF)." https://www150.statcan.gc.ca/n1/pub/82m0013x/82m0013x2024001-eng.htm.

Steele, Leah S, Carolyn S Dewa, Elizabeth Lin, and Kenneth LK Lee. 2007. "Education Level, Income Level and Mental Health Services Use in Canada: Associations and Policy Implications." *Healthcare Policy* 3 (1): 96.

"Strathmore University Institutional Repository." n.d. https://su-plus.strathmore.edu/items/ad94085b-03d4-44e1-9e28-c29b1e38a7c3.

Stringaris, Argyris. 2017. "What Is Depression?" *Journal of Child Psychology and Psychiatry*. Wiley Online Library.

Vaughan, Davis, and Hadley Wickham. 2021. *Float: High-Performance Numerics with C++ Float Library*. https://CRAN.R-project.org/package=float.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley et al. 2019. *The Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

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

World Health Organization. 2014. *Social Determinants of Mental Health*. World Health Organization. https://iris.who.int/handle/10665/112828.

———. 2024. "Depression." https://www.who.int/news-room/fact-sheets/detail/depression.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in r." *Implementing Reproducible Computational Research*.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://haozhu233.github.io/kableExtra/.