Understanding Drinking Behavior Through Bayesian Analysis

Santiago Vessi, 1958879

Introduction

This project explores the correlation between alcohol consumption and various health indicators using a dataset from the National Health Insurance Service in Korea [1]. Employing Bayesian statistical methods, particularly logistic and probit regression models, we analyze a sample of 5,000 observations to predict drinking status based on demographic, physiological, and biochemical markers. The study leverages Bayesian techniques to key findings that include significant associations between drinking and variables such as sex, age, liver enzymes, and smoking status, providing insights into the health implications of alcohol consumption.

Dataset

The dataset analyzed was sourced from the National Health Insurance Service (NHIS) in Korea and consists of health-related data from a significant sample of the Korean population. The initial dataset contained 991,346 observations across 24 variables. For computational efficiency and analysis, there were randomly sampled 5,000 observations from this larger dataset. The summary statistics reveal a wide range of values, particularly for variables like triglycerides and cholesterol, indicating variability in health metrics across the population.
No missing values were reported in the sampled dataset, which is a positive aspect for our analysis integrity.

```
sum(colSums(is.na(data)))
```

```
## [1] 0
```

The dataset under analysis focuses on predicting drinking behavior, represented by the binary dependent variable **DRK_YN**, which indicates whether an individual is a drinker (1) or a non-drinker (0).
The dataset includes a range of independent variables that provide physiological and lifestyle information about the individuals. Among the independent variables, several are continuous, capturing measurements such as **age**, **height**, **weight**, systolic and diastolic blood pressure (**SBP** and **DBP**), blood glucose levels (**BLDS**), **cholesterol**, **triglyceride**, **hemoglobin**, and liver biochemical markers **(SGOT_AST, SGOT_ALT, gamma_GTP)**. These variables offer insights into the physical and metabolic health of the participants.
In addition, the dataset includes categorical variables that describe characteristics such as **sex**, **smoking status**, **urine protein levels**, and **hearing ability** in both the left and right ears. Smoking status, in particular, is divided into three categories: individuals who have never smoked (0), those who used to smoke but have quit (1), and those who currently smoke (2).
The combination of continuous and categorical predictors allows for a comprehensive exploration of the factors influencing drinking behavior.

[1] https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset

EDA

```
table(data$DRK_YN)/length(data$DRK_YN)*100
```

```
##
##     0     1
## 50.58 49.42
```

The dataset used in this analysis consists of individuals classified as either drinkers or non-drinkers, with 50.58% identified as non-drinkers and 49.42% as drinkers. This nearly balanced distribution provides a solid foundation for exploring patterns and relationships associated with drinking behavior.

Relationships Between Variables

We start by checking correlations between numeric variables and **DRK_YN**.

```
cor_results = corr.test(data,data$DRK_YN, method = "pearson")
cor_values = cor_results$r[, 1]
p_values = cor_results$p[,1]

selected_vars=names(cor_values[abs(cor_values)>0.05 & p_values<0.05])
```

To identify the most relevant predictors for drinking status, a **Pearson correlation** analysis was conducted. Variables were selected based on having an absolute correlation coefficient greater than 0.05 and a p-value less than 0.05, ensuring statistical significance. This analysis highlighted several key variables associated with drinking behavior. These included demographic attributes such as sex and age, anthropometric measures like height, weight, and waistline, as well as physiological and biochemical indicators including diastolic blood pressure (DBP), LDL cholesterol, triglyceride levels, hemoglobin, serum creatinine, liver enzymes (SGOT_AST, SGOT_ALT, and gamma_GTP), and smoking status. Sensory measures, including vision (sight_left, sight_right) and hearing ability (hear_left, hear_right), were also identified as relevant.
To standardize the data and facilitate meaningful comparisons, all continuous variables were scaled. This transformation ensures that variables measured on different scales contribute equally during modeling and analysis, enhancing the robustness of statistical techniques employed. By incorporating this preprocessing step and focusing on statistically significant predictors, the dataset is well-prepared for advanced modeling approaches to uncover patterns and associations related to drinking behavior.

Visualizations

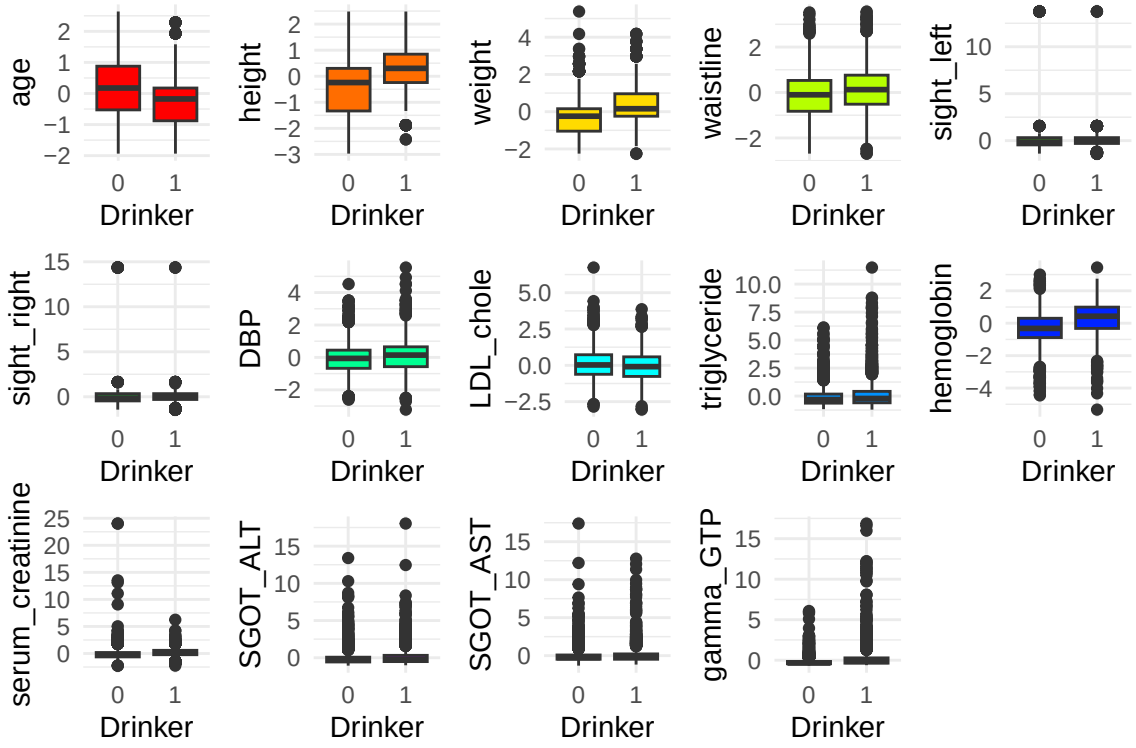

Figure 1: Boxplots

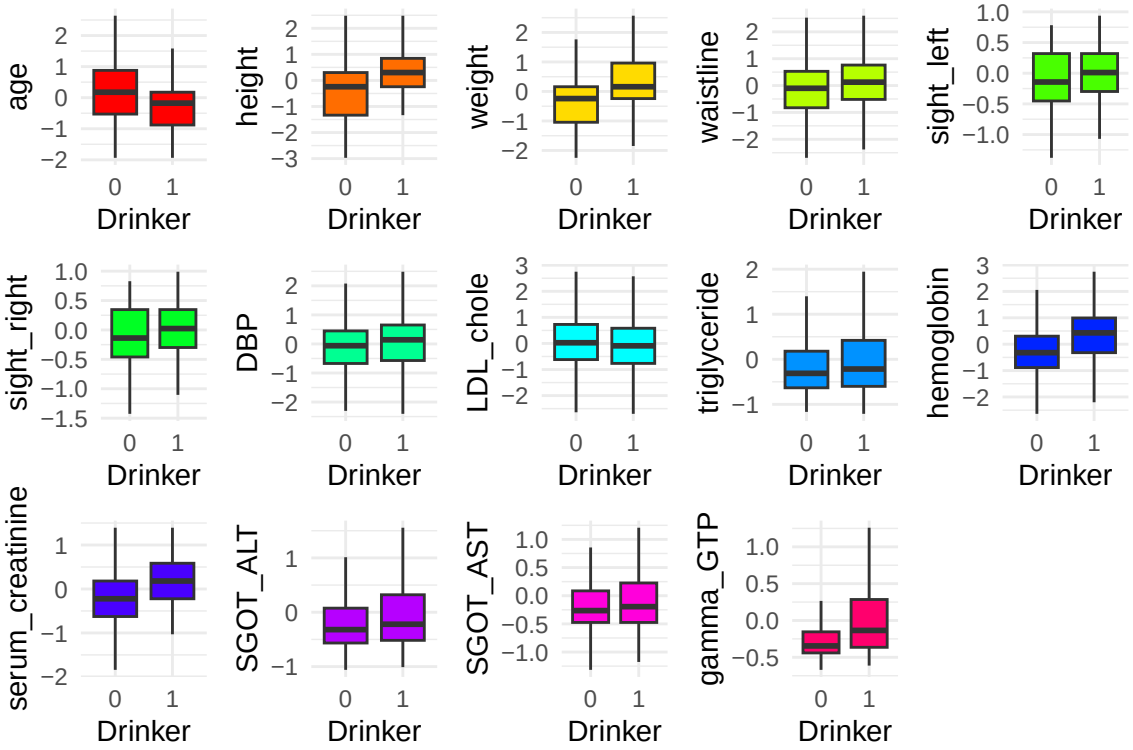

Figure 2: Boxplot without outliers

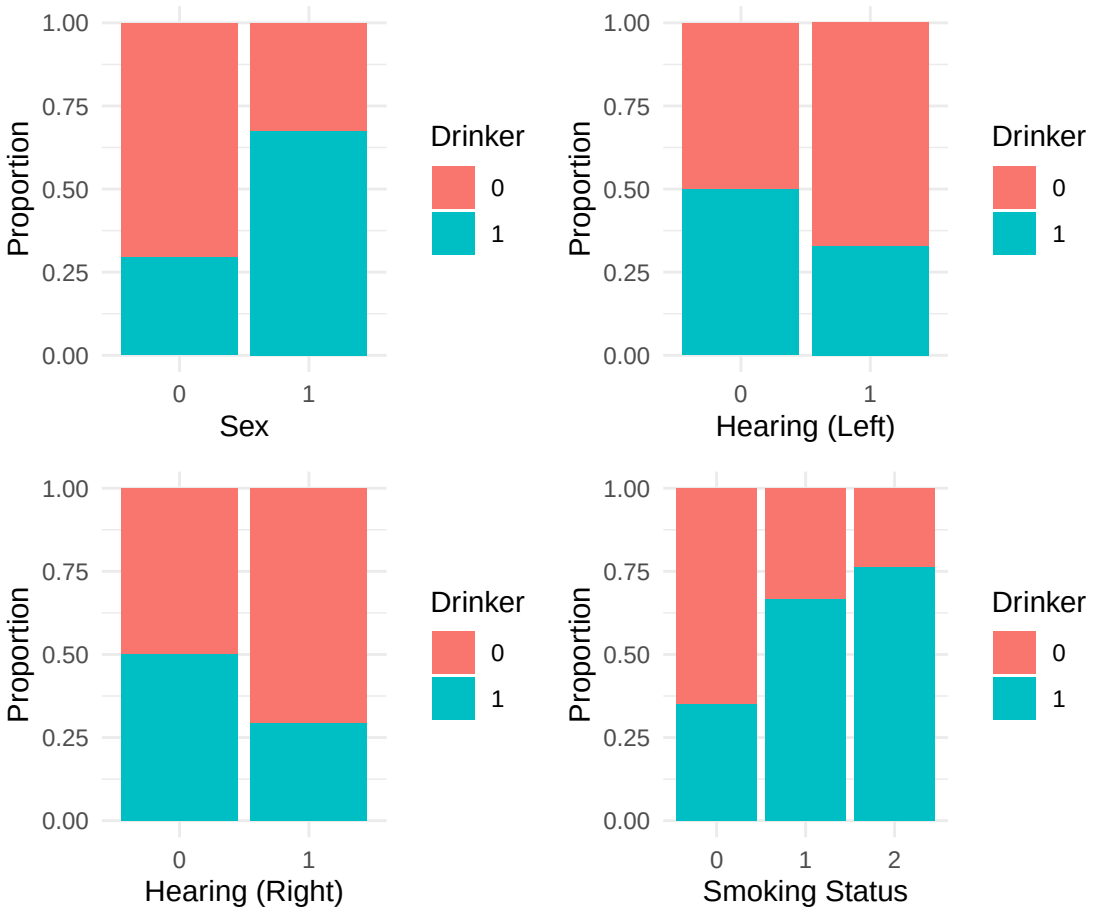

Figure 3: Proportion of Drinkers by Sex, Hearing Ability (Left and Right), and Smoking Status

A series of visualizations were created to explore the relationships between drinking behavior and various predictors in the dataset. These plots provide insights into the distribution and patterns of key variables, highlighting differences between drinkers and non-drinkers.

Boxplots (**Figure 1 & 2**) were generated for continuous variables, such as age, height, weight, waistline, blood pressure, and biochemical markers, to examine differences in their distributions based on drinking status. For example, variables like **age** and **LDL cholesterol** showed higher values among non-drinkers, suggesting that older individuals and those with higher cholesterol levels may be less likely to drink. Conversely, most other variables, including **waistline**, **triglycerides**, **hemoglobin**, **serum creatinine**, and liver enzymes (**SGOT_ALT**, **SGOT_AST**, and **gamma_GTP**), exhibited higher values among drinkers. These patterns may reflect metabolic and physiological differences associated with drinking habits. Additionally, the boxplots revealed a large number of outliers for liver enzymes, particularly **SGOT_ALT**, **SGOT_AST**, and **gamma_GTP**, which could indicate liver function abnormalities in drinkers.

For categorical variables, bar plots (**Figure 3**) were used to assess proportional differences. When analyzing **sex**, it was observed that males were more likely to be drinkers compared to females. Similarly, **smoking status** was strongly associated with drinking behavior. Current smokers (coded as 2) had the highest proportion of drinkers, followed closely by former smokers (coded

as 1), while those who never smoked (coded as 0) had a significantly lower proportion of drinkers. These trends suggest a strong link between smoking habits and alcohol consumption.
Hearing ability, measured separately for the left and right ears, also showed notable patterns. Individuals with normal hearing (coded as 0) were more likely to drink compared to those with abnormal hearing (coded as 1). This may indicate potential lifestyle or health differences between the groups.
Overall, the visualizations effectively highlight how drinking behavior correlates with demographic, physiological, and lifestyle factors. The observed trends suggest that drinkers tend to exhibit metabolic profiles and habits linked to higher health risks, emphasizing the need for further analysis to quantify these relationships.

Statistical Model

Model Specification

The logistic regression model is defined within the **JAGS** framework, which allows for Bayesian inference through Markov Chain Monte Carlo (MCMC) simulations. The model assumes that the probability of being a drinker follows a **Bernoulli distribution**:

$$y_i \sim Bernoulli(p_i)$$

where p_i is the probability that individual i is a drinker. The logit function is used to model this probability as a linear combination of the predictors:

$$logit(p_i) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK}$$

Here:

- α: Intercept term
- β: Coefficients for each predictor
- X: Matrix of predictor values

Priors Weakly informative priors are placed on the coefficients (β) and the intercept (α):

- $\alpha \sim N(0, 0.01)$: This assumes no strong prior knowledge about the intercept but allows it to vary widely.
- $\beta_j \sim N(0, 0.01)$: Similar assumptions apply to the regression coefficients, enabling the data to primarily inform the posterior distributions.

Code Implementation The dataset is preprocessed, and the predictors are scaled to standardize values across variables. The **JAGS** model is implemented as a text file named `"logistic_model.jags"`, which is read into R for simulation. Initialization sets the starting values for α and β as 0.
The MCMC simulation is executed with:

- **3 chains** to assess convergence,
- **1,000 adaptation iterations** for initial tuning,
- **4,000 burn-in iterations** to allow the chains to stabilize, and
- **1500 sampling iterations** with thinning (keeping every 10th sample) to reduce autocorrelation.

Table 1: Summary

	lower	upper	posterior_means	significance	prop
alpha	-0.83	-0.52	-0.68	1	0.00
sex	0.57	1.07	0.82	1	1.00
age	-0.71	-0.54	-0.63	1	0.00
height	0.08	0.33	0.21	1	1.00
weight	-0.13	0.19	0.03	0	0.63
waistline	-0.08	0.18	0.05	0	0.76
sight_left	-0.07	0.08	0.00	0	0.53
sight_right	-0.14	0.01	-0.06	0	0.06
hear_left	-0.32	0.61	0.10	0	0.66
hear_right	-0.85	0.12	-0.31	0	0.11
DBP	-0.03	0.11	0.03	0	0.84
LDL_chole	-0.18	-0.05	-0.11	1	0.00
triglyceride	-0.11	0.05	-0.03	0	0.19
hemoglobin	0.01	0.19	0.10	1	0.99
serum_creatinine	-0.21	-0.04	-0.12	1	0.00
SGOT_AST	0.20	0.48	0.34	1	1.00
SGOT_ALT	-0.72	-0.45	-0.57	1	0.00
gamma_GTP	0.63	0.96	0.79	1	1.00
SMK_stat_type_cd	0.38	0.58	0.48	1	1.00

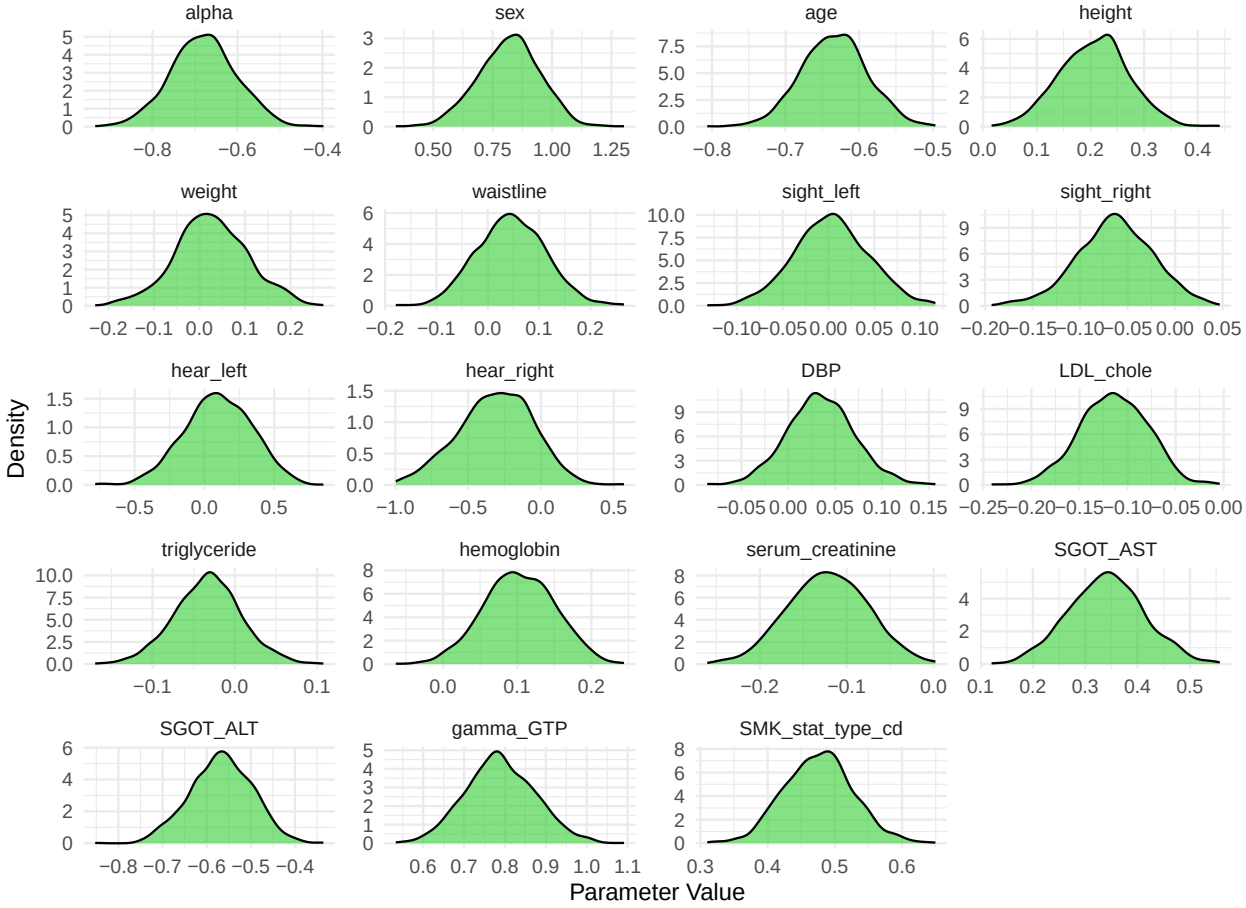

Figure 4: Posterior Distributions of Parameter

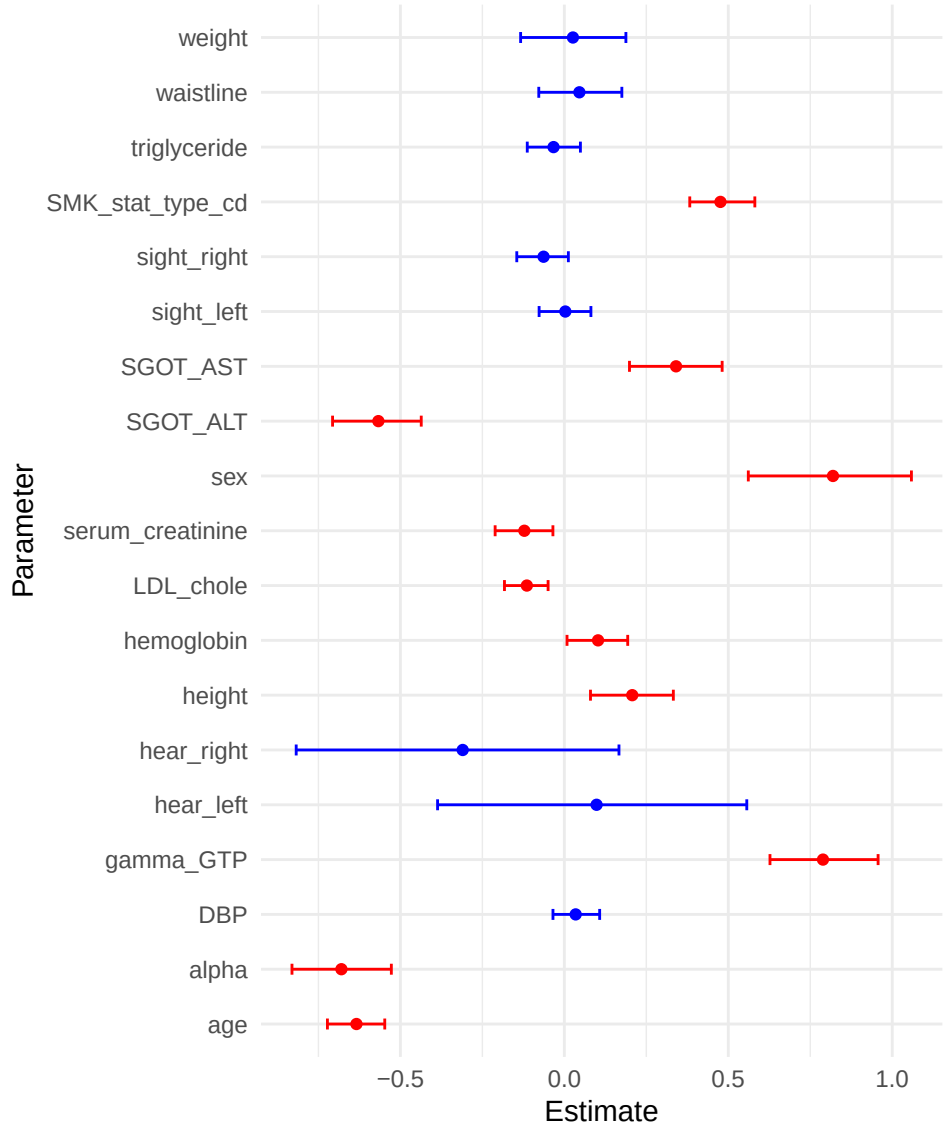

Figure 5: 95% Credible Intervals for Model Parameters

The model's baseline, represented by the **intercept** (-0.68), shows that the odds of drinking are relatively low when all other predictors are absent. This reflects a starting point where, in the absence of other influences, individuals are less likely to engage in drinking.

One of the most significant findings is the role of **sex**. The strong positive coefficient (0.82) highlights that men are far more likely to drink than women, with the model showing complete confidence in this result (prop = 1.00). **Age**, in contrast, has a substantial negative relationship with drinking. As people get older, their likelihood of drinking decreases significantly (coefficient: -0.63), which could reflect shifts in lifestyle, increasing health concerns, or changing social patterns with age.

Height emerges as another significant factor, with a small positive association (coefficient: 0.21). This suggests that taller individuals are slightly more likely to drink, though the reasons behind this relationship are possibly related to the sex variable. in this model.

Health markers reveal intriguing patterns. Lower **LDL** cholesterol and **serum creatinine** levels are associated with higher odds of drinking (coefficients: -0.11 and -0.12, respectively), which may point to differences in diet, metabolism, or overall lifestyle between drinkers and non-drinkers. Conversely, higher **hemoglobin** levels (coefficient: 0.10) are positively linked to drinking. This might reflect associations with physical activity or dietary habits.
The role of liver function markers is particularly striking. **SGOT_AST** and **gamma_GTP**, both of which are known to rise with alcohol consumption, show strong positive associations with drinking (coefficients: 0.34 and 0.79, respectively). These results align with established medical knowledge about the effects of alcohol on the liver. Interestingly, **SGOT_ALT** shows a negative association (coefficient: -0.57), suggesting that different aspects of liver health or other confounding factors could influence this relationship.
Smoking status also plays a major role, with smokers significantly more likely to drink (coefficient: 0.48). This may support the behavioral link between smoking and drinking, where one often predicts the other.
Not all predictors showed significant effects. Variables like **weight**, **waistline**, **vision**, **hearing**, and diastolic blood pressure (**DBP**) had coefficients near zero, with credible intervals that included zero, indicating their influence on drinking is either negligible or masked by other factors. For example, while weight and waistline have weak positive coefficients (0.03 and 0.05, respectively), their low significance suggests limited relevance in explaining drinking behavior.
The model also provides an indication of certainty through the "prop" column, this offers a probabilistic measure of the directionality (positive/negative) of each parameter's effect. It is particularly useful when interpreting Bayesian models, as it reflects the uncertainty in the posterior distribution directly. Predictors like **sex**, **age**, and **gamma_GTP** have very high prop values (close to 1.00), indicating robust and reliable effects. Conversely, predictors like **hearing** and **triglycerides**, with prop values near 0.5, show less certainty or weaker relationships with drinking habits.
This analysis offers valuable insights into factors influencing drinking behavior. Key demographic characteristics, such as sex and age, emerge as strong predictors. Health-related variables, including liver function markers, hemoglobin levels, and cholesterol, also play significant roles, highlighting the connection between physical health and drinking habits. In conclusion, this Bayesian logistic regression analysis shows the various influences on drinking habits, with demographic factors, health markers, and smoking status emerging as key predictors.

Simulation

To validate and understand the robustness of the initial Bayesian logistic regression model, a simulation study was conducted. Using the true parameter estimates from the original model, synthetic data were generated to mimic the observed relationships and evaluate how well the model can recover these known parameters.
The simulation involved creating 5000 observations with predictor variables sampled from a standard normal distribution. Using the true parameter values for the intercept (α) and regression coefficients (β) derived from the original model, a linear predictor (η) was computed. Probabilities were then calculated using the logistic link function ($p = 1/(1+\exp(-\eta))$), and binary outcomes (y) were simulated based on these probabilities. This process yielded a dataset with a structure similar to the original data.

The simulated dataset was analyzed using the same Bayesian logistic regression model, and posterior samples of the parameters were obtained to compare their estimates to the true values.

Table 2: Summary Simulation

	True_Value	Mean	Lower	Upper
	-0.68	-0.71	-0.78	-0.64
sex	0.82	0.82	0.74	0.90
age	-0.63	-0.64	-0.71	-0.57
height	0.21	0.21	0.14	0.29
weight	0.03	0.02	-0.05	0.09
waistline	0.05	0.05	-0.02	0.13
sight_left	0.00	-0.04	-0.11	0.03
sight_right	-0.06	-0.01	-0.08	0.06
hear_left	0.10	0.10	0.03	0.17
hear_right	-0.31	-0.31	-0.38	-0.24
DBP	0.03	0.12	0.05	0.20
LDL_chole	-0.11	-0.13	-0.20	-0.06
triglyceride	-0.03	-0.02	-0.08	0.05
hemoglobin	0.10	0.11	0.03	0.18
serum_creatinine	-0.12	-0.13	-0.20	-0.07
SGOT_AST	0.34	0.39	0.33	0.46
SGOT_ALT	-0.57	-0.61	-0.68	-0.54
gamma_GTP	0.79	0.78	0.71	0.86
SMK_stat_type_cd	0.48	0.51	0.43	0.59

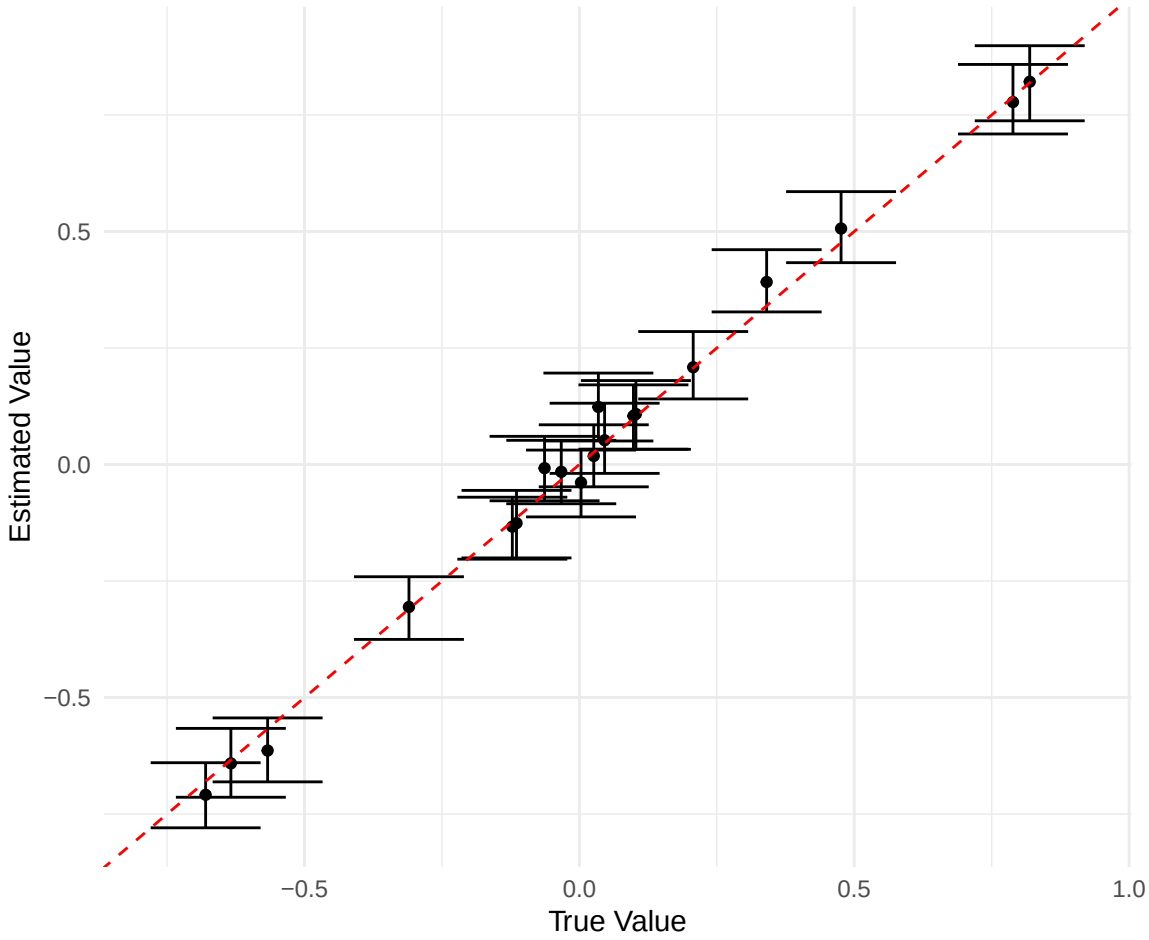

Figure 6: True vs Estimated Parameters

Simulation Results

The results show (**Table 2**) that the model successfully recovered the true parameter values with high precision:
- **Intercept (**α): The estimated mean (-0.71) closely matches the true value, with a narrow 95% credible interval (-0.78, -0.64), indicating strong agreement.
- **Regression Coefficients (**β): For most predictors, the posterior means closely align with their true values, with small standard deviations. The credible intervals for all significant predictors captured the true parameter values, reinforcing the model's ability to recover the underlying relationships.
The following trends were observed:
1. Predictors such as **sex** (strong positive effect, mean = 0.82), and **gamma_GTP** (positive effect, mean = 0.78) showed clear influence on the outcome, with their posterior distributions narrowly centered around the true values.
2. Variables like **age** (mean = -0.64) and **SGOT_ALT** (mean = -0.61) exhibited strong negative associations, consistent with the true parameter estimates.
3. Smaller effects, such as **weight** (mean = 0.02) and **triglyceride** (mean = -0.02), were also

recovered, though their influence is weaker and less certain due to broader credible intervals.
The 95% credible intervals for all significant predictors successfully captured the true values, indicating a high level of model reliability. For instance:
- **sex**: 95% CI (0.74, 0.90)
- **age**: 95% CI (-0.71, -0.57)
- **height**: 95% CI (0.14, 0.29)
These results confirm the consistency and reliability of the Bayesian framework in recovering parameter estimates from data with known underlying structures.
By successfully recovering the true parameter values, the model demonstrates its reliability in capturing the relationships between predictors and the outcome. This builds confidence in the initial analysis and suggests that the observed relationships in the real-world data are likely reflective of genuine associations.
Additionally, the simulation underscores the importance of parameter uncertainty, as smaller effects (**sight_left**,**sight_right**) were less precisely estimated, which could reflect noise or limited predictive power of these variables.
This simulation study validates the Bayesian logistic regression model used for the original analysis. The close alignment of estimated and true parameter values reinforces the model's robustness and suggests that it provides a reliable framework for understanding the predictors of drinking behavior.

Alternative Model

Model Specification

The Probit regression model is defined within the JAGS framework, this model assumes that the probability of being a drinker follows a Bernoulli distribution:

$$y_i \sim Bernoulli(p_i)$$

where y_i represents whether individual i is a drinker (1) or not (0), and p_i is the probability that individual i is a drinker. The probit function, which is the cumulative distribution function (CDF) of the standard normal distribution, is used to model this probability as a linear combination of the predictors:

$$\Phi^{-1}(p_i) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK}$$

or equivalently,

$$p_i = \Phi(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_K X_{iK})$$

Where:

- α: Intercept term
- β: Coefficients for each predictor where β_j ($j = 1, 2, \ldots, K$) corresponds to the j-th predictor
- X: Matrix of predictor values where X_{ij} is the value of the j-th predictor for the i-th individual
- Φ: The standard normal CDF (probit function)

Priors

Weakly informative priors are placed on the coefficients (β) and the intercept ($alpha$):

$$\alpha \sim N(0, 0.01)$$

This assumes no strong prior knowledge about the intercept but allows it to vary widely. The precision (inverse of variance) of 0.01 corresponds to a variance of 100, indicating a broad range of potential values for the intercept.

$$\beta_j \sim N(0, 0.01) \quad \text{for} \quad j = 1, 2, \ldots, K$$

Similar assumptions apply to all regression coefficients. These priors encourage shrinkage towards zero but do not strongly constrain the coefficients, allowing the data to primarily inform the posterior distributions. The choice of precision here (0.01) again implies a variance of 100 for each coefficient.

Analysis of DIC Results

```
## [1] "Logistic:"

## Mean deviance:  5450
## penalty 19.09
## Penalized deviance: 5469

## [1] "Probit"

## Mean deviance:  5457
## penalty 18.35
## Penalized deviance: 5475
```

The Deviance Information Criterion (DIC) values for the logistic and probit models provide insight into their relative performance. DIC balances goodness-of-fit with model complexity, where lower values indicate a better trade-off.
For the **logistic model**, the DIC is **5469**, with a mean deviance of **5450** and a penalty for complexity of **19.09** effective parameters. In comparison, the **probit model** has a DIC of **5475**, with a mean deviance of **5457** and a slightly lower complexity penalty of **18.35** effective parameters. The difference in DIC between the two models is minimal ($\Delta DIC = 6$), suggesting that both models exhibit similar performance.
The logistic model's slightly lower mean deviance indicates a marginally better fit to the observed data. However, the probit model compensates with a slightly reduced complexity penalty, reflecting fewer effective parameters. This difference may be due to the inherent characteristics of the probit link function, which handles extreme probabilities differently than the logistic link.
In practical terms, the small difference in DIC implies that both models are nearly equivalent in predictive capability. The choice between them may differ on theoretical considerations or the context of the analysis. Logistic regression is often preferred for its interpretability, as coefficients are directly related to odds ratios. Conversely, the probit model may be more appropriate in

scenarios where a latent variable framework or a cumulative normal distribution assumption aligns better with the underlying data generation process.
In conclusion, while the logistic model shows a slightly better balance of fit and complexity, the probit model offers an alternative perspective with comparable performance. For most applications, the logistic model remains the default choice unless there is a compelling theoretical reason to prefer the probit link function.

MCMC Diagnostics

Finally it was evaluated the performance and convergence of MCMC sampling using a combination of visualizations and diagnostic metrics.
The trace plots (Figure 7) visualize the sampled parameter values across iterations, helping to assess the stability and mixing of the chains. These plots were generated using `ggplot`, where each parameter's trace is shown as a line, and the parameters are faceted for clarity. The trace plots exhibit good mixing, with no visible trends, sudden jumps, or plateaus. This suggests that the chains have stabilized and are sampling effectively from the posterior distribution.

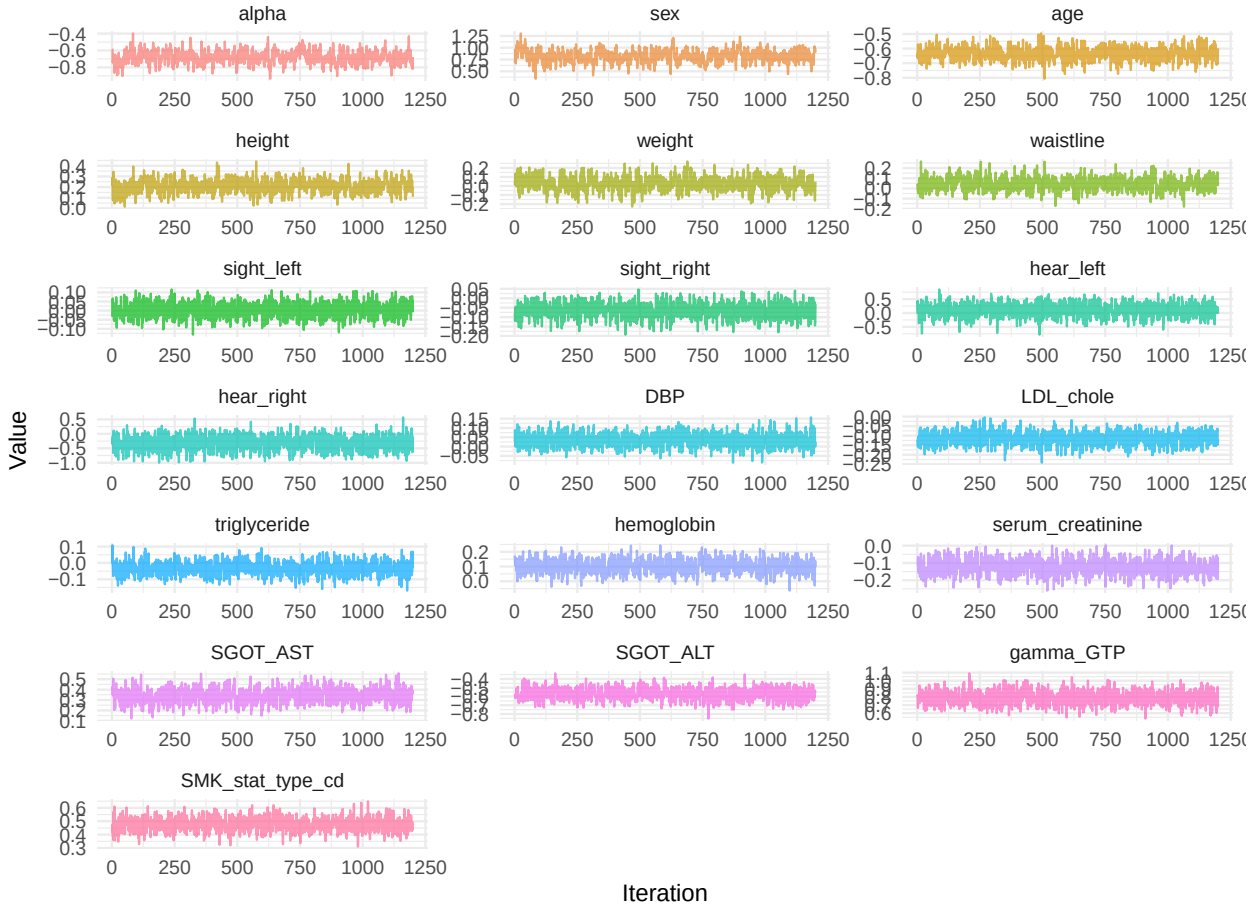

Figure 7: Trace Plots for MCMC Chains

The **Gelman-Rubin** diagnostic (**Table 3**) evaluates chain convergence by calculating the Po-

tential Scale Reduction Factor (PSRF). The PSRF values for all parameters are close to **1.0**, with point estimates ranging from **1.000** to **1.022**. The multivariate PSRF is **1.02**, further supporting the conclusion that the chains have converged. Values near 1.0 indicate that between-chain and within-chain variances are nearly identical, confirming convergence.

Table 3: Gelman Diagnostics

	psrf.Point.est.	psrf.Upper.C.I.
alpha	1.01	1.02
sex	1.00	1.01
age	1.00	1.00
height	1.00	1.01
weight	1.00	1.00
waistline	1.00	1.00
sight_left	1.00	1.01
sight_right	1.01	1.02
hear_left	1.00	1.00
hear_right	1.00	1.01
DBP	1.00	1.01
LDL_chole	1.00	1.01
triglyceride	1.00	1.00
hemoglobin	1.00	1.00
serum_creatinine	1.00	1.00
SGOT_AST	1.01	1.02
SGOT_ALT	1.01	1.02
gamma_GTP	1.00	1.00
SMK_stat_type_cd	1.00	1.00

The **effective sample size** (**Table 4**) quantifies how many independent samples the chains effectively represent, accounting for autocorrelation. The ESS values range from **307** to over **1400** across parameters. These high values reflect good chain mixing, minimal autocorrelation, and a sufficient number of independent samples for reliable inference.

Table 4: Effective Sample Size

	round.ess_values..2.
alpha	307.27
sex	270.65
age	714.97
height	494.59
weight	484.65
waistline	559.20
sight_left	1200.00
sight_right	1264.87
hear_left	1200.00
hear_right	1447.41
DBP	1316.68
LDL_chole	1064.34
triglyceride	1070.67
hemoglobin	779.38
serum_creatinine	996.82
SGOT_AST	586.62
SGOT_ALT	656.03
gamma_GTP	1166.55
SMK_stat_type_cd	905.32

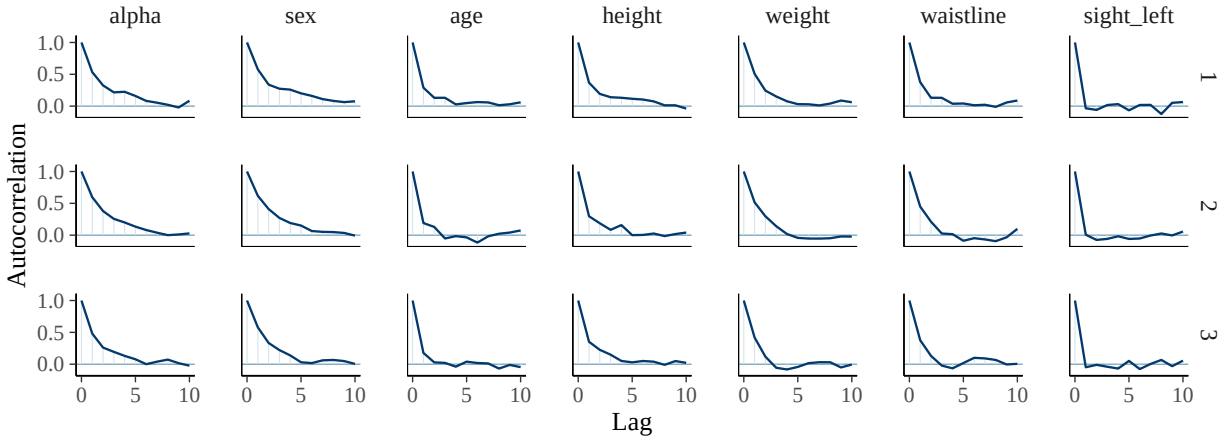

Figure 8: Autocorrelation 1/3

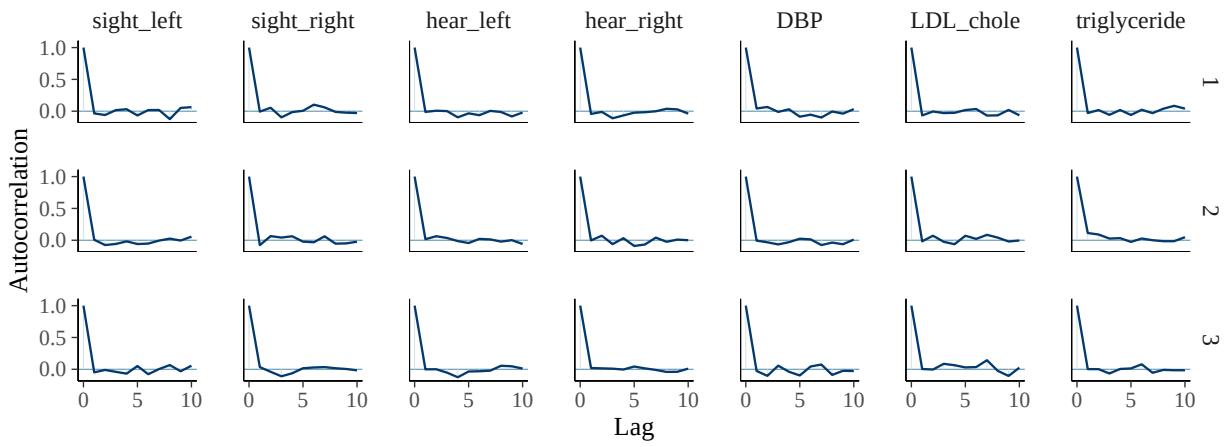

Figure 9: Autocorrelation 2/3

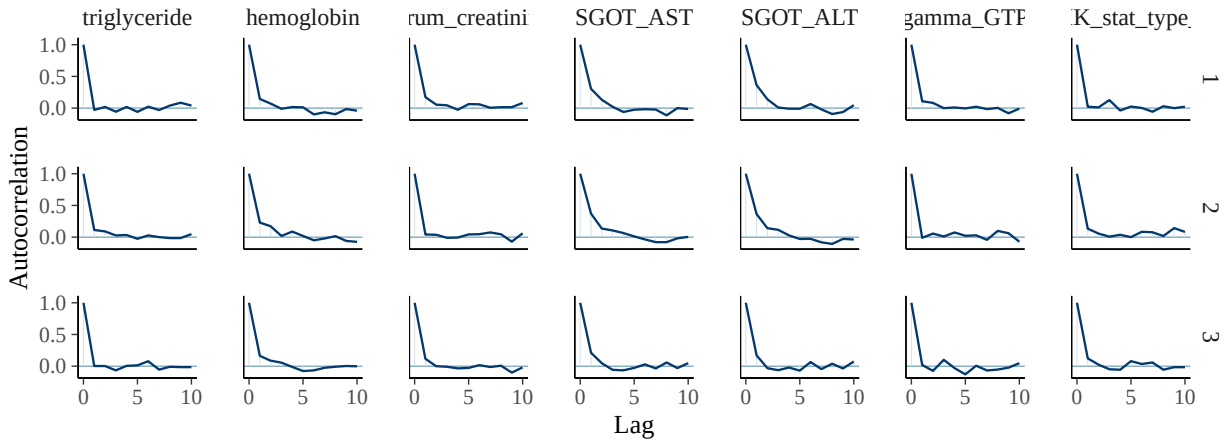

Figure 10: Autocorrelation 3/3

The autocorrelation plots (**Figures 8,9 & 10**) provide valuable insights into the behavior of the Markov Chain Monte Carlo sampling process for the Bayesian model parameters. These plots were generated to assess how well the chains mix and how quickly the dependency between successive samples diminishes over ten lags.

The absence of erratic spikes or slow declines further reinforces the notion that the MCMC chains are mixing effectively across all parameters. This uniformity across different subsets is a reassuring indicator of the sampler's overall reliability and performance.

The behavior observed across all three subsets is significant for several reasons. A rapid decline in autocorrelation means that the chains have reached a stationary distribution, allowing for a well-represented posterior. It also suggests that the sampler settings, including burn-in period, thinning, and the number of iterations, are appropriately configured for this analysis. These findings indicate that the chains provide reliable samples with minimal bias from autocorrelation. The convergence diagnostics for the MCMC sampling process collectively indicate that the model's posterior distributions have been reliably estimated.

Overall, these convergence diagnostics collectively validate the reliability and stability of the MCMC process. The chains have converged effectively, and the posterior estimates can be confidently used for inference and interpretation. This ensures that the Bayesian analysis conducted is robust and the results are a faithful representation of the data and the model.

Conclusion

The Bayesian logistic regression model, supplemented by a probit alternative, effectively identified key predictors of drinking behavior from the NHIS dataset. Significant predictors included sex, age, and various biochemical markers, particularly liver function indicators like SGOT_AST and gamma_GTP, which showed expected correlations with alcohol consumption. The simulation study confirmed the robustness of the model by accurately recovering true parameter values, enhancing confidence in the model's predictive power. The minimal difference in performance between logistic and probit models (as evidenced by DIC) suggests that while logistic regression offers better fit, probit could be considered for theoretical alignment with latent variable assumptions. Finally, MCMC diagnostics confirmed excellent convergence and mixing of chains, suggesting reliable posterior estimates.