1 Introduction

Obesity is a critical global health issue with rising prevalence across all age groups and regions. It is a leading risk factor for numerous chronic conditions, including cardiovascular disease, diabetes, and certain cancers, making it a significant burden on healthcare systems worldwide. As obesity is influenced by complex interactions between genetic, environmental, and behavioral factors, understanding these contributors is crucial for designing effective public health interventions.

This study aims to explore the relationship between lifestyle factors and obesity levels, focusing on how these factors vary across different age groups. Specifically, the analysis uses data from a questionnaire to quantify the effects of eating habits and physical conditions as predictors of obesity. By employing a Bayesian hierarchical model, it incorporate uncertainties and account for variability at both the individual and group levels, enabling a deeper understanding of these relationships.

2 Data Preparation

There are several preprocessing procedures performed to guarantee the dataset was prepared correctly for this project. The goal of these procedures was to convert variables into suitable formats while maintaining their meaning and structure. Dataset was categorized into synthetic and real-world, then manipulation was performed on the latter to allow for more real-life implications.

NObeyesdad, the goal variable, refers to seven ranked categories of obesity levels, from "Insufficient Weight" to "Obesity Type III." In order to represent the growing severity of obesity, this variable's ordinal nature was converted into a numeric scale. "Insufficient Weight" was given the lowest value (1) and "Obesity Type III" the highest value (7). The categories were encoded in a progressive fashion. Given the ordinal nature of the response variable even after transforming them into numerical scale, I transform obesity levels into an ordered response and perform a probit regression on the response, since the probit model is well-suited for analyzing outcomes with an inherent order.

Age was grouped into 2 meaningful brackets to capture group-level effects: 14-21 years and 21-61 years. The breakpoint at 21 holds important empirical meaning not only as a boundary of adolescents, but also as indicator of legal alcohol consumption. These age brackets enable me to account for the variability in obesity patterns across different life stages. Each group was assigned a numeric value to allow for its inclusion as a group-level predictor in the hierarchical model.

For consistency, binary category variables —— which have two possible outcomes —— were converted to numerical values. For instance, women were encoded as 1, men otherwise; those who had a family history of obesity received a score of 1, while those who did not received a score of 0. Frequent caloric food consumption (FAVC) was encoded binary numerical values too, those who ate a lot of calories were assigned a score of 1, and those who didn't were assigned a score of 0. Similar transformations were done to variables such as calorie consumption monitoring (SCC) and smoking status (SMOKE). The dataset's variables, including food intake between meals (CAEC) and alcohol consumption frequency (CALC), were regarded as ordered categorical predictors. These variables were encoded with numeric values that maintained their ordinal linkages and describe behaviors on a scale from "Never" to "Always." As an illustration, "Never" received the lowest score, while "Sometimes," "Frequently," and "Always" received the greatest scores. When integrating these categories into the analysis, this encoding makes sure the model takes into account their inherent ordering. Moreover, continuous predictors, such as physical activity frequency (FAF), daily water intake (CH2O), and the number of main meals consumed per day (NCP), were standardized. This transformation ensures that all continuous predictors are on a comparable scale, which helps improve the stability and efficiency of the Bayesian model. Once all transformations were completed, the dataset was thoroughly reviewed to confirm that all variables were correctly encoded and ready for analysis.

3 Methodology

The project employed a Bayesian hierarchical model with three layers to analyze the relationship between individual-level factors (high-calorie food intake, physical activity, alcohol consumption, and family history of overweight) and obesity levels across different age groups. The model incorporates both fixed effects and random effects for age brackets to capture grouplevel variability. The posterior distributions of the parameters were estimated using Markov Chain Monte Carlo (MCMC) sampling with the Gibbs sampling algorithm, implemented in R using the rjags package to interface with JAGS. Hierarchical structure specification is as followed:

1. At the observation level, the response variable Y_i represents the ordinal categorical variable for obesity levels with $K = 7$ categories. The probability of observing Y_i is determined by the thresholds τ_k and the latent continuous variable Z_i:

$$P(Y_i = k) \quad \begin{cases} \Phi(\tau_1 - \mu_i) & \text{if } k = 1 \\ =\Phi(\tau_k - \mu_i) - \Phi(\tau_{k-1} - \mu_i) & \text{if } k = 2,\ldots, K - 1\ 1 - \\ \Phi(\tau_{K-1} - \mu_i) \quad \text{if } k = K & \end{cases}$$

where:

- $\Phi(\cdot)$ is the CDF of the standard normal distribution.
- $\tau_1 < \tau_2 < \ldots < \tau_{K-1}$ are monotonically increasing thresholds.

2. At the latent variable layer, the ordered nature of Y_i is modeled through a latent continuous variable Z_i, which follows a normal distribution:

$$Z_i \sim N(\mu_i, \sigma_z 2)$$

The mean μ_i of Z_i is defined by a linear predictor that includes fixed and random effects:

$\mu_i = \beta_1 + \beta_2 \cdot \text{FAVC}_i + \beta_3 \cdot \text{FAF}_i + \beta_4 \cdot \text{CALC}_i + \beta_5 \cdot \text{family\_history}_i + \gamma\text{AgeBracket}_{[i]} + \delta \cdot \text{Gender}_i$

where:

- $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ are the fixed effect coefficients for the covariates.
- $\gamma\text{AgeBracket}_{[i]}$ represents the random effect for age group j, capturing group-level variability.
- δ is the fixed effect coefficient for gender.

3.At the group level, the step accounts for effects for age brackets, as proposed:

$$\gamma_j \sim N(0,\tau_\gamma), \quad \tau_\gamma = \frac{1}{\sigma_\gamma^2}, \quad j = 1,2$$

where σ_γ represents the standard deviation of the random effects, quantifying variability across age groups.

As for the priors and hyperparameters, I specified weakly informative priors to allow the data to influence the posterior distributions while maintaining minimal prior assumptions:

- Fixed Effects Priors:

$$\beta_b \sim N(0,0.001), b = 1,\ldots,5, \quad \delta \sim N(0,0.001)$$

- Random Effect Variance Prior:

$$\sigma_\gamma \sim \text{Uniform}(0,10),$$

- Latent Variable Variance Prior: $\sigma_z \sim \text{Uniform}(0,10)$, $\tau_\gamma = \sigma\overline{1_\gamma 2}$

z

$\tau_z = \sigma\overline{1 2}$ • Threshold

Parameters Prior:

$$\tau_1 \sim \text{Uniform}(-10,10), \tau_k \sim \text{Uniform}(\tau_{k-1},10), k = 2,\ldots, K-1$$

The posterior distributions of the parameters in the Bayesian hierarchical ordered probit model were estimated using Gibbs sampling, a Markov Chain Monte Carlo (MCMC) technique. I implemented this process in R using JAGS. Gibbs sampling is well-suited for the model because it allows us to iteratively sample from the full conditional distributions of each parameter, given the current values of all other parameters, enabling efficient posterior estimation. I will talk about initialization for parameters in later section of the report, with the understanding that it provides starting points for each MCMC chain in mind. In addition, the iterative nature of Gibbs sampling enabled us to systematically update each parameter based on its full conditional distribution, leveraging the hierarchical structure of the model to incorporate fixed effects, random effects, and uncertainty in threshold parameters. After ensuring convergence when done running the model, I summarized the posterior distributions of the parameters by extracting the posterior means, credible intervals, and density plots. The posterior means provided point estimates for each parameter, while the 95% credible intervals quantified the uncertainty associated with these estimates. This comprehensive MCMC approach, combined with rigorous convergence diagnostics, ensured that the posterior estimates were reliable and robust, providing a solid foundation for interpreting the results.

4 Computational Details

1. Initialization for Parameters

The hierarchical ordered probit model was initialized iteratively to achieve better convergence. The key initialization choices were:

- Fixed Effects (β): Initialized as small random values from N(0,0.1), reflecting weakly informative priors centered around zero.
- Random Effects (γ): Random effects corresponding to age groups were initialized as small values from N(0,0.1).
- Threshold Parameters (τ): Thresholds were initialized as increasing sorted values sampled from $Uniform(-2,2)$, ensuring the monotonic constraint required for ordered probit models.
- Variance Parameters (σ and σγ): Initialized to 1, reflecting non-informative priors.

The initialization values were refined based on preliminary diagnostic outputs, such as traceplots and convergence metrics.

2. Iterations

After utilizing three parallel chains for the MCMC sampling process to ensure robust exploration of the posterior parameter space, each chain was run with 20,000 iterations, with the first 5,000 iterations discarded as burn-in to remove early unstable estimates. To reduce autocorrelation in the sampled values, a thinning interval of 10 was applied, meaning every 10th iteration was retained. After thinning, each chain yielded a total of 1,500 samples, resulting in a combined set of 4,500 posterior samples. Multiple chains were initialized with different starting values to further ensure that the posterior parameter space was thoroughly explored.

3. Convergence Diagnostics

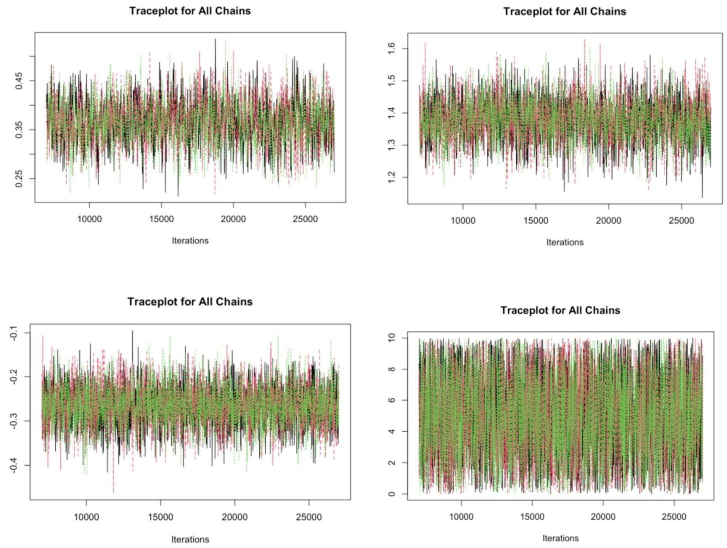


Figure 1. Traceplots displaying MCMC sampling behavior.

The traceplots presented above display the MCMC sampling behavior for the β coefficients (fixed effects) and the δ parameter (gender effect) across 3 chains. The traceplots exhibit dense oscillations with good mixing across iterations, where all three chains (represented in different colors) appear to overlap significantly. This behavior indicates that the chains are exploring the posterior distribution thoroughly without getting stuck in local modes. 4. Autocorrelation Check

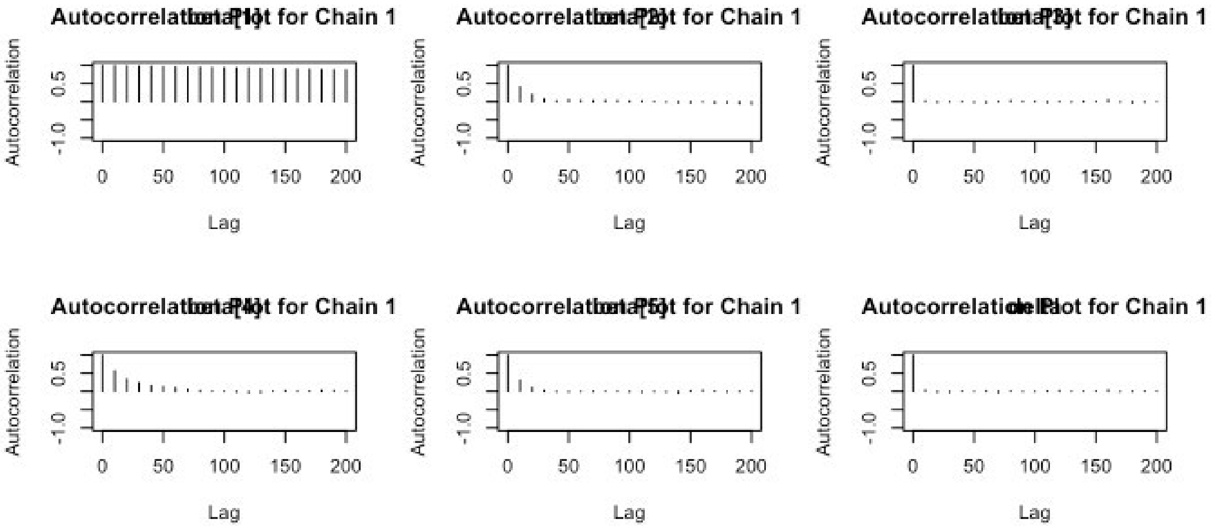
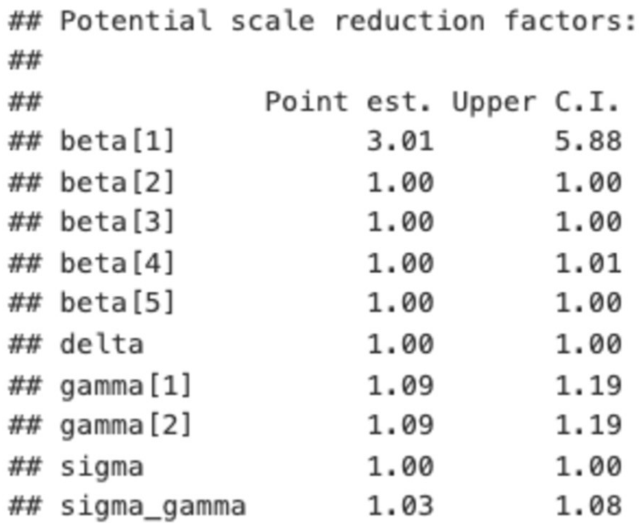


Figure 2. Autocorrelation Checks for Key Parameters.

Autocorrelation was evaluated for key parameters. For well-behaved parameters (e.g., β2,β3), autocorrelation declined quickly after a few lags. 5. Gelman-Rubin Diagnostics for Convergence



```
## Potential scale reduction factors:
##
##              Point est. Upper C.I.
## beta[1]            3.01       5.88
## beta[2]            1.00       1.00
## beta[3]            1.00       1.00
## beta[4]            1.00       1.01
## beta[5]            1.00       1.00
## delta              1.00       1.00
## gamma[1]           1.09       1.19
## gamma[2]           1.09       1.19
## sigma              1.00       1.00
## sigma_gamma        1.03       1.08
```

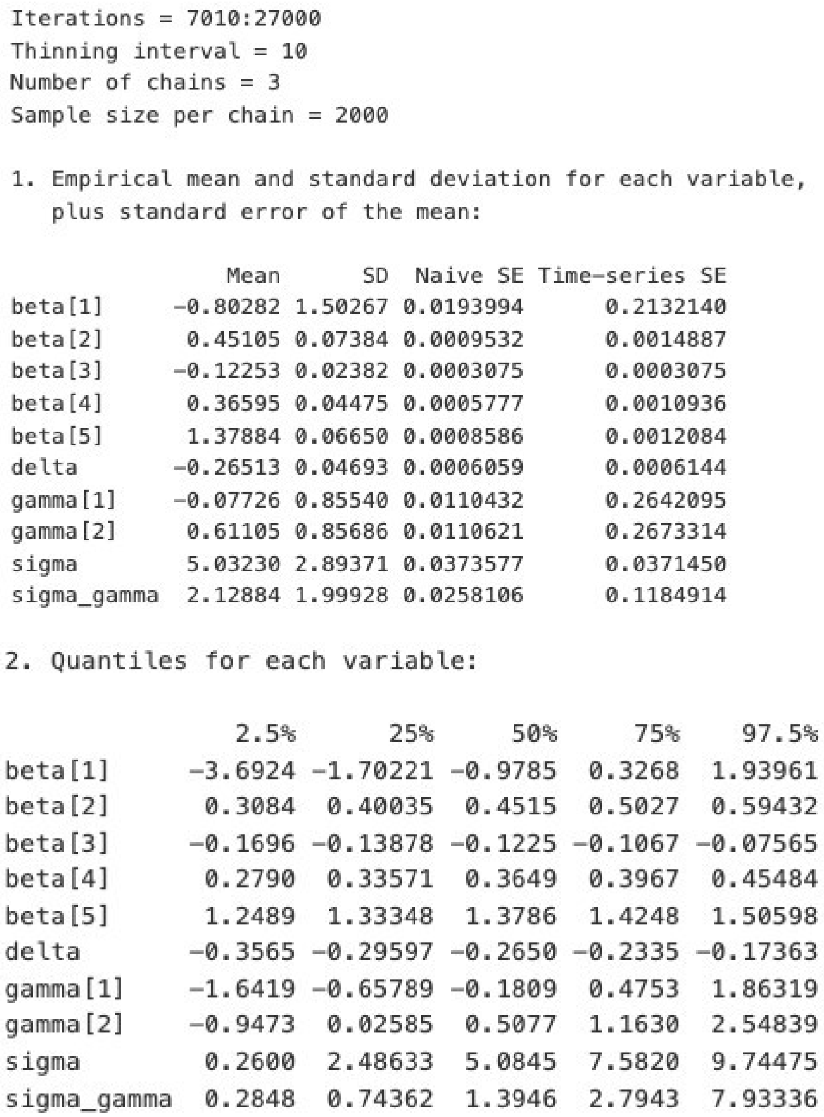
Table 1. Gelman-Rubin Statistics for Paramters.

The Gelman-Rubin statistic was computed for all parameters to confirm convergence. Parameters such as β2,β3 (I skipped discussion for β_1 most of the times because it is the intercept term) and δ had PSRF values close to 1, indicating convergence. Threshold parameters τ had higher PSRF values, suggesting slow convergence.

6. Monte Carlo Error

Monte Carlo Standard Errors (MCSE) were evaluated for all parameters to assess the precision of the posterior estimates (table in next section). Parameters like β2, β3, β4, δ had small MCSEs relative to their posterior standard deviations, ensuring stable estimates. Threshold parameters τ exhibited higher MCSEs, reflecting slow mixing and autocorrelation.

5 Final Results and Key Findings



```
Iterations = 7010:27000
Thinning interval = 10
Number of chains = 3
Sample size per chain = 2000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean      SD  Naive SE Time-series SE
beta[1]     -0.80282 1.50267 0.0193994      0.2132140
beta[2]      0.45105 0.07384 0.0009532      0.0014887
beta[3]     -0.12253 0.02382 0.0003075      0.0003075
beta[4]      0.36595 0.04475 0.0005777      0.0010936
beta[5]      1.37884 0.06650 0.0008586      0.0012084
delta       -0.26513 0.04693 0.0006059      0.0006144
gamma[1]    -0.07726 0.85540 0.0110432      0.2642095
gamma[2]     0.61105 0.85686 0.0110621      0.2673314
sigma        5.03230 2.89371 0.0373577      0.0371450
sigma_gamma  2.12884 1.99928 0.0258106      0.1184914

2. Quantiles for each variable:

               2.5%      25%     50%     75%    97.5%
beta[1]     -3.6924 -1.70221 -0.9785  0.3268  1.93961
beta[2]      0.3084  0.40035  0.4515  0.5027  0.59432
beta[3]     -0.1696 -0.13878 -0.1225 -0.1067 -0.07565
beta[4]      0.2790  0.33571  0.3649  0.3967  0.45484
beta[5]      1.2489  1.33348  1.3786  1.4248  1.50598
delta       -0.3565 -0.29597 -0.2650 -0.2335 -0.17363
gamma[1]    -1.6419 -0.65789 -0.1809  0.4753  1.86319
gamma[2]    -0.9473  0.02585  0.5077  1.1630  2.54839
sigma        0.2600  2.48633  5.0845  7.5820  9.74475
sigma_gamma  0.2848  0.74362  1.3946  2.7943  7.93336
```

Table 2. Summary for Model Results.

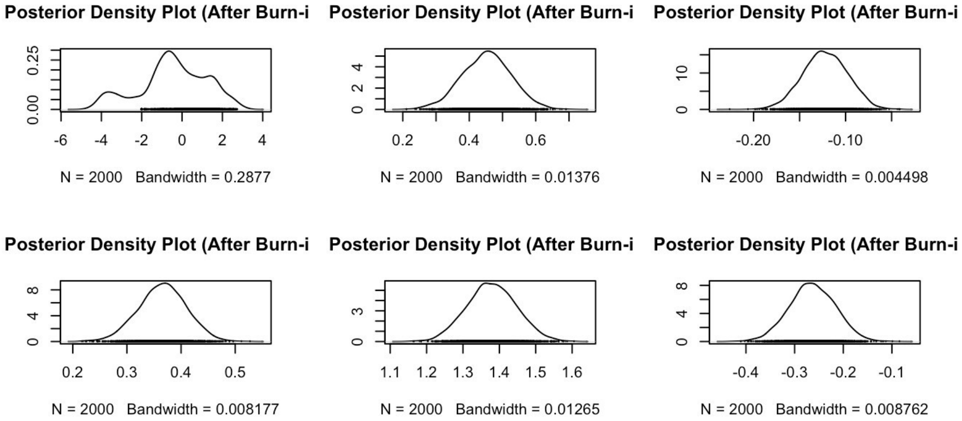


Figure 3. Posterior Density Plots (Only presenting the key factors due to excessive outputs).

The table above summarizes the posterior means and standard deviations for key parameters. Based on the posterior summaries and quantiles provided, some conclusions can be drawn regarding the significance and relationships of the variables in the model. Variables are considered significant if their 95% credible intervals (2.5% to 97.5%) do not include zero:

- The analysis of β2, which represents high-calorie food consumption, reveals a posterior mean of 0.451 with a 95% credible interval ranging from [0.3084, 0.5943]. This result indicates a significant positive relationship between high-calorie food consumption and obesity levels. Specifically, individuals who frequently consume high-calorie food are more likely to fall into higher obesity categories. The positive association is supported by the credible interval, which does not include zero, affirming the robustness of this finding.
- The analysis of β5, which represents family history of overweight, shows a posterior mean of 1.3788 with a 95% credible interval of [1.2489, 1.5059]. This result highlights a strong positive relationship between having a family history of overweight and higher obesity levels. The significant association, as evidenced by the credible interval not containing zero, suggests that family history of overweight is one of the most important predictors in the model, indicating a higher likelihood of individuals with such a background being categorized into elevated obesity levels.
- The analysis of δ delta, which represents gender, reveals a posterior mean of -0.2651 with a 95% credible interval of [-0.3650, -0.1736]. This indicates a significant negative relationship between gender and obesity levels. Specifically, males (coded as 1) are less likely to fall into

higher obesity categories compared to females. The negative posterior mean and the credible interval excluding zero confirm the robustness of this relationship, suggesting that gender is an influential predictor in determining obesity levels.

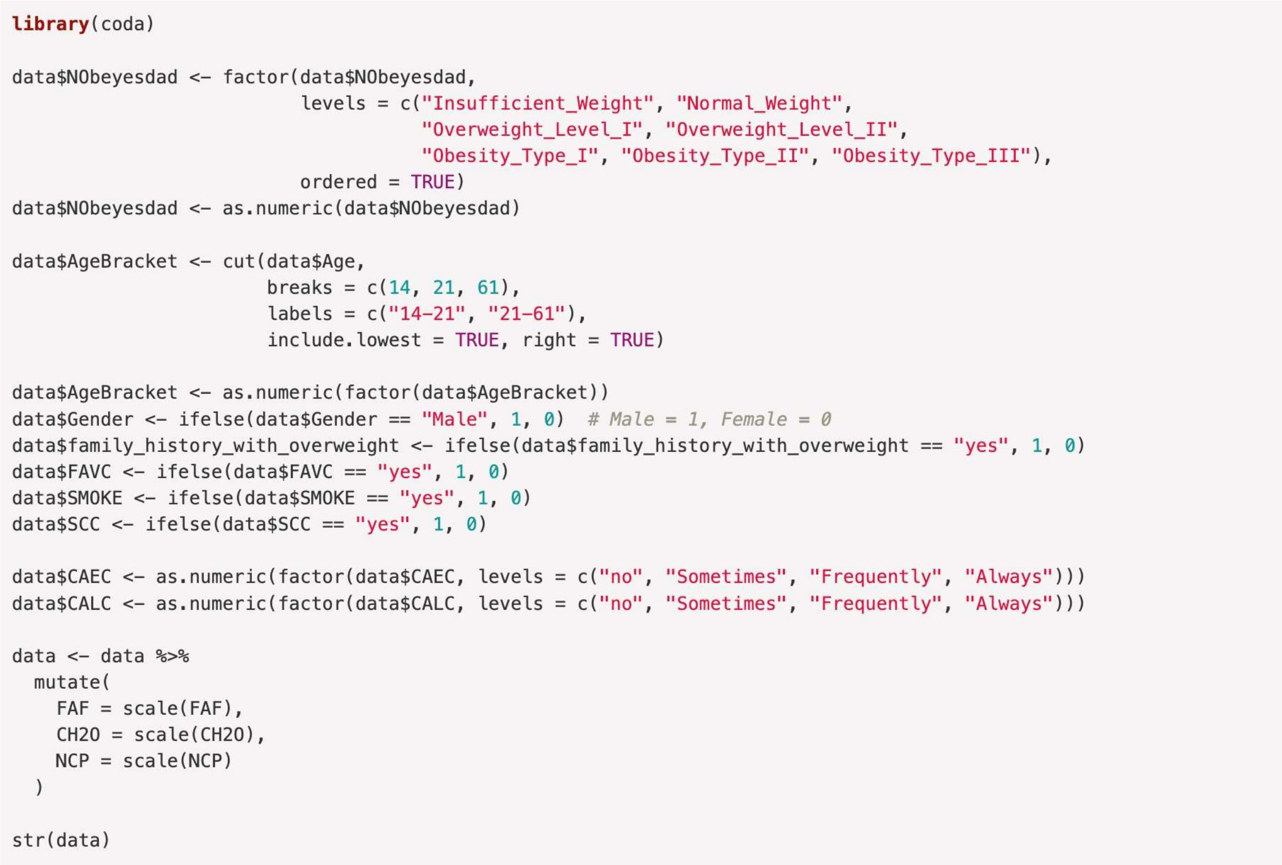
6 Conlusion and Further Steps

The Bayesian hierarchical model with ordered Probit for obesity categories identifies key covariates influencing obesity, but it exhibits inefficiencies for specific parameters. Though the model results appear to be trivial, there are some further steps I could take to improve validity. Specifically, To address convergence issues, particularly for parameters such as β1 , increasing the number of burn-in iterations and total samples can provide more time for the chains to stabilize and improve overall convergence. Sampling efficiency can also be improved by applying thinning, which reduces the correlation between successive samples by selecting every k-th iteration, thereby enhancing mixing and efficiency. To evaluate the model fit, posterior predictive checks will be conducted by simulating data from the posterior predictive distribution (PPD) and comparing summary statistics, such as means, variances, and proportions, across obesity categories. If the simulated data aligns closely with the observed data, it indicates that the model effectively captures the underlying patterns. Additionally, including interactive terms between covariates, such as physical activity (FAF) and family history, may help better capture complex relationships within the data. Finally, comparing the ordered probit model with alternative models, such as multinomial logistic regression, using comparison metrics like DIC can identify the best-performing approach for the dataset and ensure robust results.

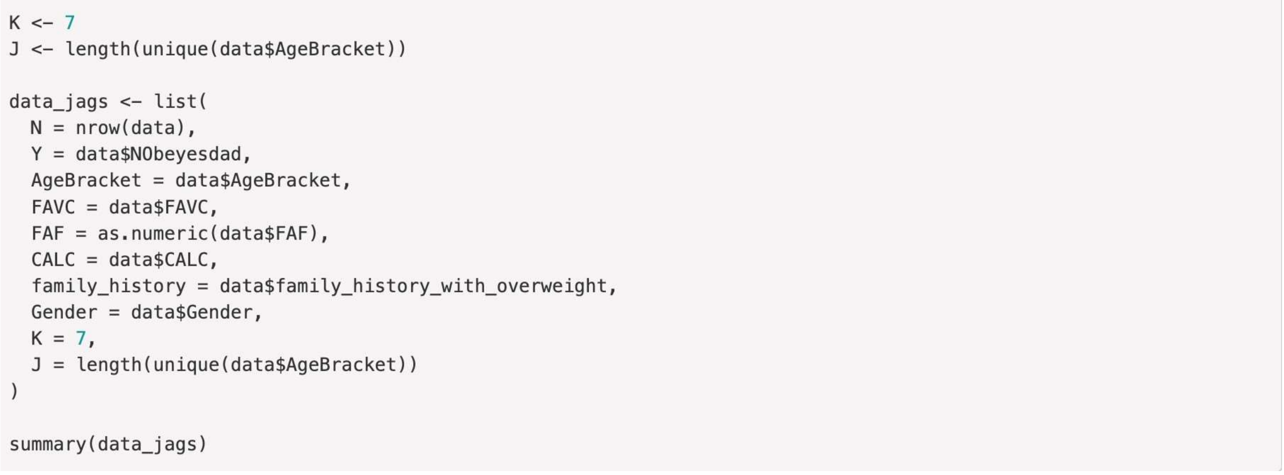
References

1. Lim, J., Akashi, Y., Song, D., Hwang, H., Kuwahara, Y., Yamamura, S., Yoshimoto, N., & Itahashi, K. (2018). Hierarchical Bayesian modeling for predicting ordinal responses of personalized thermal sensation: Application to outdoor thermal sensation data. Building and Environment, 142, 414–426. https://doi.org/10.1016/j.buildenv.2018.06.045
2. UCI Machine Learning Repository. (2019). Estimation of obesity levels based on eating habits and physical condition [Data set]. University of California, Irvine. https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

Appendix



```r
library(coda)

data$NObeyesdad <- factor(data$NObeyesdad,
                          levels = c("Insufficient_Weight", "Normal_Weight",
                                     "Overweight_Level_I", "Overweight_Level_II",
                                     "Obesity_Type_I", "Obesity_Type_II", "Obesity_Type_III"),
                          ordered = TRUE)
data$NObeyesdad <- as.numeric(data$NObeyesdad)

data$AgeBracket <- cut(data$Age,
                       breaks = c(14, 21, 61),
                       labels = c("14-21", "21-61"),
                       include.lowest = TRUE, right = TRUE)

data$AgeBracket <- as.numeric(factor(data$AgeBracket))
data$Gender <- ifelse(data$Gender == "Male", 1, 0)  # Male = 1, Female = 0
data$family_history_with_overweight <- ifelse(data$family_history_with_overweight == "yes", 1, 0)
data$FAVC <- ifelse(data$FAVC == "yes", 1, 0)
data$SMOKE <- ifelse(data$SMOKE == "yes", 1, 0)
data$SCC <- ifelse(data$SCC == "yes", 1, 0)

data$CAEC <- as.numeric(factor(data$CAEC, levels = c("no", "Sometimes", "Frequently", "Always")))
data$CALC <- as.numeric(factor(data$CALC, levels = c("no", "Sometimes", "Frequently", "Always")))

data <- data %>%
  mutate(
    FAF = scale(FAF),
    CH2O = scale(CH2O),
    NCP = scale(NCP)
  )

str(data)
```

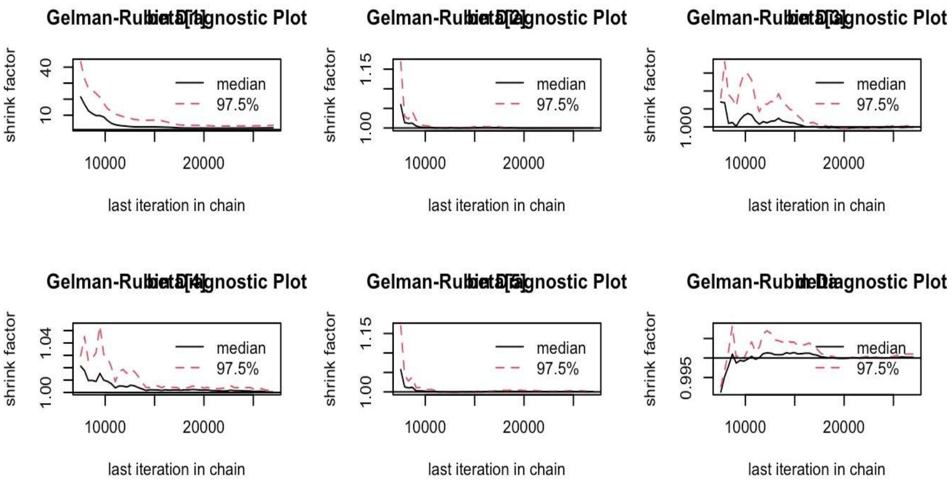


```r
K <- 7
J <- length(unique(data$AgeBracket))

data_jags <- list(
  N = nrow(data),
  Y = data$NObeyesdad,
  AgeBracket = data$AgeBracket,
  FAVC = data$FAVC,
  FAF = as.numeric(data$FAF),
  CALC = data$CALC,
  family_history = data$family_history_with_overweight,
  Gender = data$Gender,
  K = 7,
  J = length(unique(data$AgeBracket))
)

summary(data_jags)
```

```
model_string <- "
model {
  # 1. Likelihood: Ordered Probit Model
  for (i in 1:N) {
    Z[i] ~ dnorm(mu[i], tau_z)

    # Ordered Probit
    p[i, 1] <- phi(tau[1] - mu[i])
    for (k in 2:(K-1)) {
      p[i, k] <- phi(tau[k] - mu[i]) - phi(tau[k-1] - mu[i])
    }
    p[i, K] <- 1 - phi(tau[K-1] - mu[i])

    Y[i] ~ dcat(p[i, 1:K])

    mu[i] <- beta[1] + beta[2] * FAVC[i] + beta[3] * FAF[i] +
             beta[4] * CALC[i] + beta[5] * family_history[i] +
             gamma[AgeBracket[i]] + delta * Gender[i]
  }

  for (j in 1:J) {
    gamma[j] ~ dnorm(0, tau_gamma)
  }

  for (b in 1:5) {
    beta[b] ~ dnorm(0, 0.001)
  }
  delta ~ dnorm(0, 0.001)

  tau_gamma <- 1 / sigma_gamma^2
  sigma_gamma ~ dunif(0, 10)
  tau_z <- 1 / sigma^2
  sigma ~ dunif(0, 10)

  tau[1] ~ dunif(-10, 10)
  for (k in 2:(K-1)) {
  tau[k] ~ dunif(tau[k-1], 10)
  }
}
"
```

```
# Gelman-Rubin Diagnostic Plot
gelman.plot(samples, autoburnin = FALSE, main = "Gelman-Rubin Diagnostic Plot")
```

```
# Gelman-Rubin Diagnos
gelman.plot(samples, a
```


Gelman-Rubin Diagnostic Plot
beta[1]
beta[2]
beta[3]
beta[4]
beta[5]
delta
shrink factor
last iteration in chain
median