Bayesian classification of obesity levels: A sequential logit approach

— Project Report —

Advanced Bayesian Data Analysis

Sathish Kabatkar Ravindranth

April 11, 2024

TU Dortmund University

Contents

1 Introduction

The pervasive rise of overweight and obesity constitutes a critical global health crisis, evident in the staggering toll of 5 million noncommunicable disease (NCD) deaths attributed to elevated body mass index (BMI) levels in 2019. Over the past few decades, both adults and children have experienced a profound increase in obesity rates, a trend that transcends economic boundaries and challenges traditional perceptions of health disparities (WHO, 2024).

For effective management of obesity and precise classification of its severity levels, logistic regression, commonly used for binary outcomes, is unsuitable. Instead, more intricate sequential models are required due to the multicategorical and ordinal nature of the response variables in obesity classification. Sequential models surpass simple logistic regression in complexity, allowing for better handling of the ordinal nature of the response variable. This underscores the necessity of employing advanced modelling techniques to ensure precise and dependable classification of obesity levels.

The problem of separation occurs when specific combinations of predictor variables perfectly predict outcome categories, resulting in unreliable estimates and model instability. In the context of obesity classification, this scenario frequently arises when factors such as weight and height distinctly determine the obesity class. Consequently, frequentist logistic regression is inadequate for accurately classifying obesity levels (Gelman et al., 2022, p. 412).

In this project, the application of the sequential model approach to classify seven different levels of obesity will be explored. The intricacies of the modelling process, including the formulation of Bayesian models to handle separation issues and the choice of appropriate priors to inform our analyses, will be delved into. By adopting a Bayesian framework, prior knowledge and uncertainty can be effectively incorporated into our model, thereby enhancing the robustness and reliability of our results.

The structure of this report is as follows: In Section 1, an overview of the dataset used in our study is presented. In Section 2, the problem of separation encountered in our analysis is discussed. Following this, an overview of the statistical methods employed, including the Bayesian framework and the sequential logit model, is provided in Section 3. Section 4 delves into the analysis process, covering topics such as prior specification, posterior sampling, model comparison, predictive performance, prior sensitivity analysis, and posterior predictive check. The limitations of our study and potential avenues for improvement are addressed in Section 5. Finally, concluding remarks on the findings and implications of our research are offered in Section 6.

2 Dataset

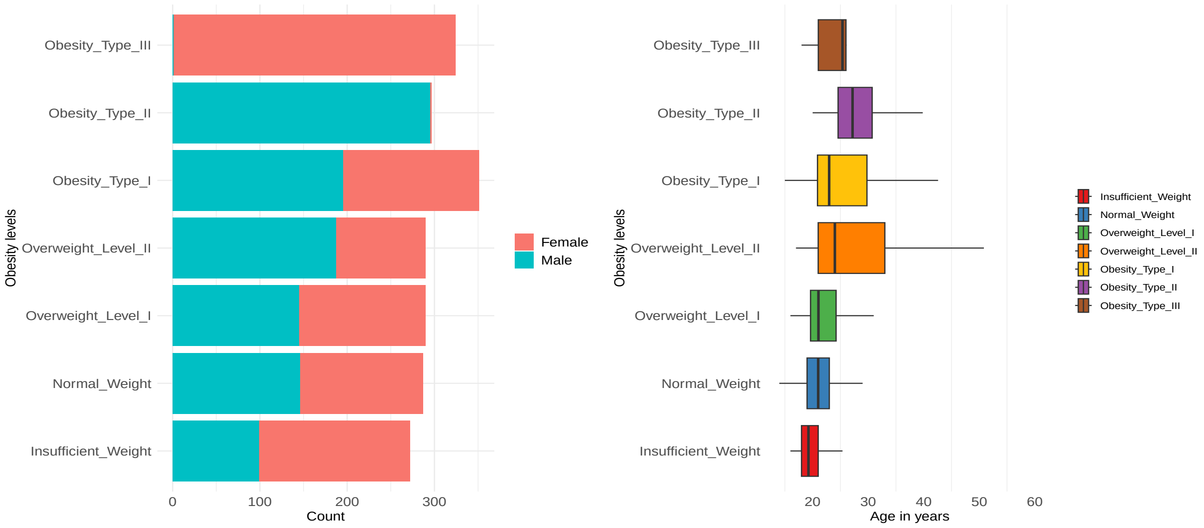
2.1 Overview

The dataset used in this study contains information collected from people in Mexico, Peru, and Colombia to estimate obesity levels based on their eating habits and physical condition. Collected via a web survey, responses from anonymous users yielded 17 attributes and 2111 records (Palechor and de la Hoz Manotas, 2023). Attributes related to eating habits include the frequency of consuming high-caloric foods, vegetables, main meals, food between meals, water intake, and alcohol consumption. Physical status attributes include monitoring calorie expenditure, frequency of physical activity, time spent on technology devices, and mode of transportation. Demographic variables such as gender, age, height, and weight are also included, along with a class variable `NObeyesdad` for true obesity levels classified. This dataset offers valuable insights into the relationship between lifestyle factors and obesity across diverse populations. Additionally, Table 2 provides names, abbreviations, and data types of variables, which are given on page 16 of the Appendix A.

2.2 Exploratory data analysis

Figure 1(a) illustrates the distribution of obesity levels by gender and shows distinct patterns. Notably, there is a large overrepresentation of males in the Obesity Type II category compared to females. This disparity suggests a potential gender-specific predisposition to severe obesity. Conversely, the Insufficient Weight and Normal Weight categories exhibit relatively balanced distributions between males and females. These findings highlight the importance of considering gender-specific factors in addressing obesity-related health disparities.

Figure 1(b) depicts the distribution of age across different obesity levels and reveals discernible patterns. Generally, as the severity of obesity increases from insufficient weight to obesity type III, there is a trend toward higher median ages observed. Specifically, Obesity Type II exhibits the highest median age among all obesity levels, suggesting that individuals with more severe forms of obesity tend to be older on average. Furthermore, there is considerable variability in age within each obesity level, as evidenced by the

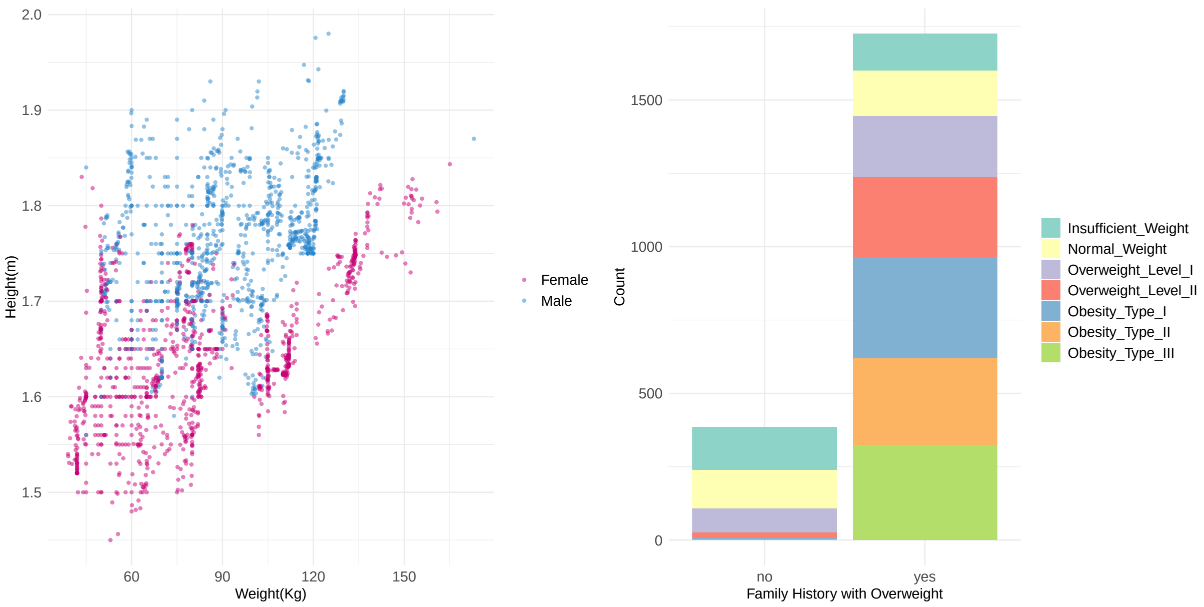


(a) No. of Persons in each obesity level

(b) Distribution of age in obesity levels

Figure 1: Distribution of obesity levels and age across the dataset.

spread of the interquartile ranges (Q1 to Q3) and the range of ages represented by the whiskers. These findings underscore the complex relationship between age and obesity severity, highlighting the potential influence of age-related factors on obesity outcomes.



(a) Weight vs Height by Gender

(b) Obesity levels with Family history

Figure 2: Scatter plot of Weight vs Height by Gender and stacked bar chart of Obesity levels by Family history.

Figure 2(a) presents a scatter plot depicting weight on the x-axis and height on the y-axis, with data points color-coded by gender. The visualization reveals clear trends: males exhibit greater height and lower weight compared to females, as indicated by the clustering of data points. This finding highlights a gender-based disparity in body composition, with males commonly demonstrating taller stature and lower weight relative to females.

Figure 2(b) shows how family history of being overweight relates to different obesity levels. People with a family history of overweight (yes) are more common in higher obesity categories like Obesity Type II and III. Meanwhile, those without a family history (no) have a more even distribution across obesity levels, with fewer cases of severe obesity. This fact suggests that having a family history of being overweight might increase the chances of developing severe obesity.

3 Problem of seperation

Nonidentifiability presents a prevalent obstacle in logistic regression, especially when compounded by factors like collinearity and separation. Separation occurs when a combination of predictors accurately predicts an outcome, leading to instability in discrete data regression.

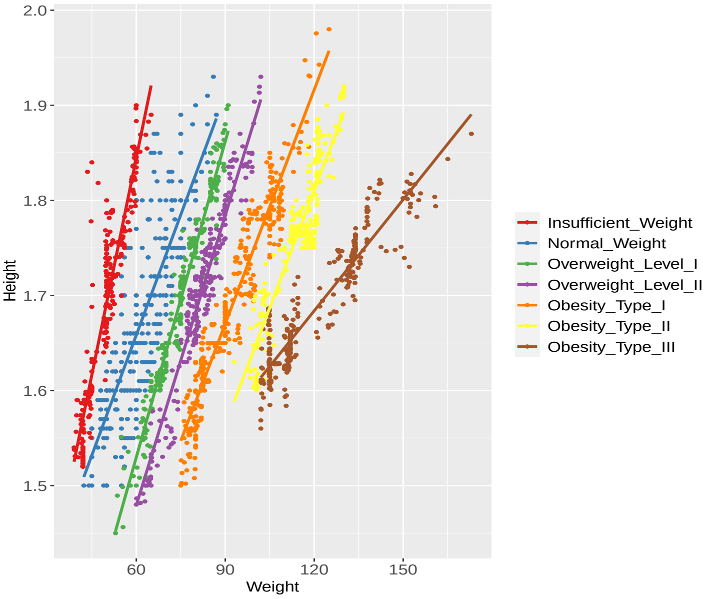


Figure 3: Weight vs Height by Obesity levels

This problem is common in logistic regression, especially when dealing with binary predictors, and is often poorly managed in practical applications. Common strategies, such as eliminating predictors to achieve model identifiability, can lead to the removal of important predictors, thereby exacerbating the problem. In the case of separation, maximum likelihood estimates may become uncertain or unreasonably biased (Gelman et al., 2022, p. 412).

Figure 3 illustrates a scatter plot of weight against height, with color-coding representing obesity levels. Clear patterns emerge within each obesity category, and there are clear distinctions between them. The presence of clear boundaries suggests that combinations of weight and height can accurately predict specific obesity levels. Consequently, the issue of separation complicates model estimation and may introduce unreliable estimates in predicting obesity levels based on weight and height.

4 Statistical methods

4.1 Bayesian framework

In the realm of Bayesian statistics, inference follows a structured three-step process revolving around the posterior distribution. Initially, a comprehensive probability model is developed, covering all pertinent observable and unobservable quantities within the problem domain, while ensuring alignment with the underlying scientific inquiry and data collection methodology. Subsequent analysis involves conditioning on the observed data, entailing the computation and interpretation of the posterior distribution, which signifies the conditional probability distribution of the unobserved quantities of interest given the observed data. Lastly, an assessment of the model's fit and the implications of the posterior distribution is conducted, evaluating the congruence between the model and the data, the rationality of the derived conclusions, and the sensitivity of the results to the employed assumptions (Gelman et al., 2022, p. 3-4).

To fit a multi categorical classification model, there is a feature input matrix $\mathbf{X} \in \mathbb{R}^{nxd}$, where $n \in \mathbb{N}$ is the number of observations, and $d \in \mathbb{N}$ is the number of features, K class labeled vector $\mathbf{y} \in \{0, 1, .., K\}^n$, and $\boldsymbol{\theta} \in \mathbb{R}^d$ a d-dimensional vector of the model parameters. Using Bayes' theorem, conditioned on the observed data $\mathbf{X}$ and $\mathbf{y}$, is defined as:

$$P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) \cdot P(\boldsymbol{\theta}|\mathbf{X})}{P(\mathbf{y}|\mathbf{X})}, \tag{1}$$

where the posterior distribution $P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$ updates beliefs about the model's parameters after observing the data. The likelihood function $P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$, indicating the probability of observing the data given the parameters and input data. The evidence $P(\mathbf{y}|\mathbf{X})$, or marginal likelihood, functions as a normalizing factor. The prior distribution on model weights $P(\boldsymbol{\theta}|\mathbf{X})$ is taken into account, representing beliefs about the parameters before observing any data, and allowing for the incorporation of prior knowledge or assumptions about the parameters into the analysis. These model weights are assumed to be independent of the input features, enabling the specification of the prior distribution without consideration of the input data. Generating the posterior distribution analytically is rarely feasible, necessitating Bayesian statistics to rely on Markov Chain Monte Carlo (MCMC) methods for obtaining samples from the posterior distribution. These sampling algorithms are computationally intensive, resulting in Bayesian model fitting generally being slower than frequentist approaches (Bürkner and Vuorre, 2019, p. 12).

4.2 Sequential logit model

Consider an ordinal response variable Y with K ordered categories, denoted as, $\{1, 2, \ldots, K\}$, representing different class levels that can only be reached sequentially. The sequential accessibility of the categories is explicitly represented through a series of binary transitions. The model aims to capture this sequential nature of transitions between these categories using latent continuous variables and thresholds (Fahrmeir et al., 2013, p. 337-339).

The transition between level $\{r\}$ and above $\{r+1, \ldots K\}$ is a binary decision which is modelled via a latent continuous variable,

$$\tilde{Y}_r = \eta + \epsilon_r, \;\; r \in \{1, \ldots, K-1\}, \tag{2}$$

where ϵ_r is category specific error term. Considering the influence of a predictor is constant across r (i.e., global coefficient vector $\boldsymbol{\beta}$) and category specific intercepts $\beta_{01}, \ldots, \beta_{0(K-1)}$, the η is defined as:

$$\eta = \beta_{0r} + \mathbf{x}^T \boldsymbol{\beta}, \;\; r \in \{1, \ldots, K-1\}. \tag{3}$$

The sequential logit model is obtained if the error term ϵ_r is assumed to follow a logistic distribution function $F(\cdot)$ i.e., cumulative distribution function (CDF) with an expected mean of 0. Then, by the definition of CDF:

$$P(\epsilon_r \leq z) = F(z) \tag{4}$$

Consider a threshold τ_r for each category, the sequential process starts at $Y = 1$, the conditional probability of the response remaining in $Y = 1$ and the transition $Y \geq 1$ is defined as:

$$P(Y = 1|Y \geq 1, \eta) = P(\tilde{Y_1} \leq \tau_1|\eta) = F(\tau_1 - (\beta_{01} + \mathbf{x}^T\boldsymbol{\beta})). \tag{5}$$

If the condition $\tilde{Y_1} \leq \tau_1$ holds, the response category is $Y = 1$, the process is stopped (i.e., stopping ratio model). If the condition fails, the process is continued until all categories are validated sequentially (Bürkner and Vuorre, 2019, p. 32-33). Accordingly, Y being in category r is given by:

$$P(Y = r|Y \geq r) = F(\tau_r - (\beta_{0r} + \mathbf{x}^T\boldsymbol{\beta})), \;\; r \in \{1, \ldots, K-1\}. \tag{6}$$

Using conditional probability, equation (6) can be rewritten with marginal probabilities:

$$P(Y = r) = F(\tau_r - (\beta_{0r} + \mathbf{x}^T\boldsymbol{\beta}))\prod_{j=1}^{r-1}(1 - F(\tau_j - (\beta_{0j} + \mathbf{x}^T\boldsymbol{\beta})) \tag{7}$$

and for the K^{th} category (i.e., reference category) :

$$P(Y = K) = 1 - \sum_{r=1}^{K-1} P(Y = r). \tag{8}$$

Replacing $F(\cdot)$ with the logistic distribution function $F(\cdot) = \frac{exp(\cdot)}{1+exp(\cdot)}$, the conditional probability from (6) is given as:

$$P(Y = r|Y \geq r) = \frac{exp(\tau_r - (\beta_{0r} + \mathbf{x}^T\boldsymbol{\beta}))}{1 + exp(\tau_r - (\beta_{0r} + \mathbf{x}^T\boldsymbol{\beta}))}, \;\; r \in \{1, \ldots, K-1\} \tag{9}$$

or equivalently,

$$log(\frac{P(Y = r|Y \geq r)}{1 - P(Y = r|Y \geq r)}) = \tau_r - (\beta_{0r} + \mathbf{x}^T\boldsymbol{\beta}). \tag{10}$$

It is important to note that these individual transitions are not directly observed, but rather only the realized category of the response variable is observed (Fahrmeir et al., 2013, p. 340).

4.3 Leave one out cross validation

Leave-One-Out Cross-Validation (LOO-CV) is a method utilized for comparing models and assessing predictive performance. It entails iteratively fitting the model to the dataset while excluding one observation at a time and predicting the omitted observation using the remaining data. The LOO-CV procedure yields the LOO Information Criteria (LOOIC), calculated from the log pointwise predictive density (lppd) for each observation. LOOIC is the sum of lppd values adjusted for the effective number of parameters in the model. This process generates out-of-sample predictions that are compared to observed values, enabling evaluation of the model's predictive accuracy. Lower LOOIC values indicate higher prediction performance. By comparing LOOIC values across models, one can identify the model that strikes the best balance between complexity and accuracy (Vehtari et al., 2016).

5 Analysis

The analysis and results of the models are presented in this section. All Bayesian techniques applied in this report are performed using the statistical programming software `R` (R Core Team, 2022) and the package `brms` (Bayesian Regression Models using Stan) (Bürkner, 2017). Plots are created using the packages `ggplot2` (Wickham, 2016) and `matplotlib` (Hunter, 2007).

Before conducting the analysis, preprocessing steps are applied to the dataset to prepare it for Bayesian modeling. First, all numeric variables are standardized to achieve a mean of 0 and a standard deviation of 1, ensuring uniform scaling. Subsequently, the dataset is partitioned into training and test sets, with 70% of the observations randomly sampled for the training set `datatrain`, while the remaining 30% are assigned to the test set `datatest`. This division allows for model training on one subset and evaluation on another, facilitating the assessment of model performance on unseen data.

5.1 Model assumptions

The assumption that obesity levels are sequentially progressed through in the sequential logit model is supported by real-life observations. It is often observed that individuals gradually transition through various stages of obesity over time. This progression reflects the natural course of obesity development, with individuals typically starting with smaller weight gains and progressing to higher levels of obesity. The use of a global coefficient vector of predictors in a sequential logit model ensures standardised modelling of the predictor effects of transitions in obesity levels, thereby enhancing interpretability and reducing the risk of overfitting in the analysis of this comprehensive dataset.

5.2 Prior specification and posterior sampling

All models were fitted using the No-U-Turn Sampler (NUTS) as the default sampling technique in the brms package (Hoffman and Gelman, 2011). Each model was run with 4 chains, with a total of 2000 iterations per chain, including a warm-up phase of 1000 iterations, resulting in a total of 4000 post-warm-up draws per model. This consistent sampling strategy ensures reliable estimates of the posterior distribution and model convergence across all analyses.

The inclusion of prior information assists in addressing the problem of separation as these priors are selected to provide some constraint on the estimates, preventing them from reaching unrealistic extremes while still allowing for flexibility in the model. For the first model (`bay_fit_1`) , weakly informative normal priors are employed for all regression coefficients, including an intercept term, with the aim of improving model convergence. Setting a weakly informative prior aims to strike a balance between incorporating prior knowledge and allowing the data to inform the estimate.

```
prior_1      <-  prior(normal(0, 5), class = "b") +
                 prior(normal(0, 5), class = "Intercept")
bay_fit_1    <-  brm(NObeyesdad ~ ., data = datatrain,
                 family = sratio(), prior = prior_1, seed = 100)
```

For the second model (`bay_fit_2`), the challenge of determining an appropriate scale for the prior distribution is addressed. Here, default prior distributions are centered at zero to reflect the absence of problem-specific information regarding the direction of coefficients. To accommodate the scale of the coefficients, a hierarchical approach is employed, approximating the appropriate scaling parameter from the data. Specifically, independent Cauchy prior distributions with a center at zero and a scale of 2.5 are assigned to each regression coefficients in the sequential logit model. This choice of prior distribution, combined with standardization, ensures that the absolute difference in logit probability remains within a reasonable range across different levels of the predictor variables (Gelman et al., 2022, p. 416).

```
prior_2      <-  prior(cauchy(0, 2.5), class = "b") +
                prior(normal(0, 5), class = "Intercept")
bay_fit_2   <-  brm(NObeyesdad ~ ., data = datatrain,
                family = sratio(), prior = prior_2, seed = 100)
```

In the third model (bay_fit_3), sensitivity of the regression coefficients is explored by utilizing a reduced scale parameter. This approach aims to shrink the regression coefficients towards zero, facilitating the assessment of the prior's impact on the estimates. Through the examination of different scale parameters, insights are gained into how the choice of prior influences the resulting estimates and model outcomes.

```
prior_3     <-  prior(cauchy(0, 0.1), class = "b") +
                prior(normal(0, 5), class = "Intercept")
bay_fit_3   <-  brm(NObeyesdad ~ ., data = datatrain,
                family = sratio(), prior = prior_3, seed = 100)
```

Rhat values for all three fitted models indicate convergence, with all values 1.00. Furthermore, the effective sample size for each posterior estimate was at least 1000, further confirming convergence. This convergence can be verified by calling the summary(bay_fit_2) function, details of which are given on page 18 of the Appendix C.

5.3 Model comparison and predictive performance

Comparing the LOOIC values across models in Table 1 reveals differences in predictive performance. The model bay_fit_2 has the lowest LOOIC value of 756.156, indicating better predictive accuracy compared to the other models. In contrast, bay_fit_1 and bay_fit_3 have higher LOOIC values of 759.496 and 763.936, respec- tively. These higher values indicate reduced prediction accuracy relative to bay_fit_2.

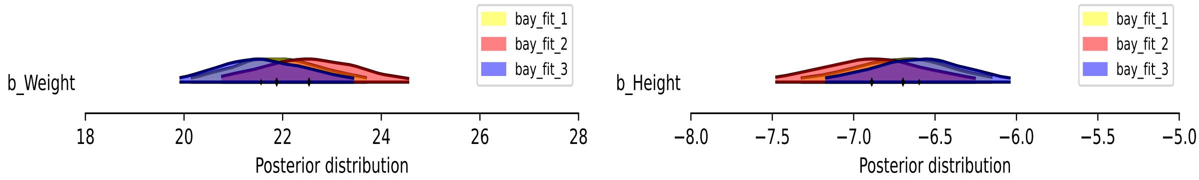
Table 1: LOOIC values and differences between three models.

Model	elpd_diff	se_diff	looic	se_looic
bay_fit_2	0.000	0.000	756.156	43.617
bay_fit_1	-1.670	1.035	759.496	42.755
bay_fit_3	-3.890	3.983	763.936	42.753

Further examination of the accuracy on the test set corroborates these findings. bay_fit_2 achieves the highest accuracy of 0.858, followed by bay_fit_1 with an accuracy of 0.845, and bay_fit_3 with an accuracy of 0.826. These accuracy metrics are consistent with the LOOIC values, reinforcing the superiority of bay_fit_2 in predicting results on unseen data.

5.4 Prior sensitivity analysis

Analysis of the 95% credible interval significance of the estimated coefficients reveals insights into their statistical significance in the model. For all three models, the interval range for the predictor `Height` in Figure 4(b) is approximately -7.5 to -6, and the interval range for the `Weight` predictor in Figure 4(a) is approximately 20 to 26. These intervals are consistent across all models, suggesting that the priors used in the analysis are insensitive and that the coefficients are primarily determined by the data themselves. Furthermore, the negative value of the height coefficient is consistent with expectations based on the BMI formula, where weight in kilograms is divided by height in metres squared.



(a) Weight

(b) Height

Figure 4: Posterior distribution of predictors Weight and Height with 95% credible intervals of all three models.

Additionally, several predictor variables in Figure 6 on page 17 of Appendix B exhibit confidence intervals that do not include zero, indicating their importance in predicting the outcome variable. For example, predictors such as `GenderMale` , `CAECFrequently`, `family_history_with_overweightyes` , `SMOKEyes` exhibit confidence intervals that exclude zero, indicating that they are affecting the response variable importance of aspects.

Moreover, the consistency of these estimates across the three models further supports the robustness of these predictors' effects on the outcome. Despite slight variations in the magnitude of the coefficients, the directionality and significance of their effects remain consistent, underscoring their relevance in predicting the outcome variable across different model specifications.

5.5 Posterior predictive check

The posterior predictive check (PPC) conducted using the Bayesian model `bay_fit_2` provides valuable insights into the model's ability to replicate the observed data. By generating replicated draws of responses based on the posterior distribution of the model parameters, the PPC allows for an assessment of how well the model fits the observed data.

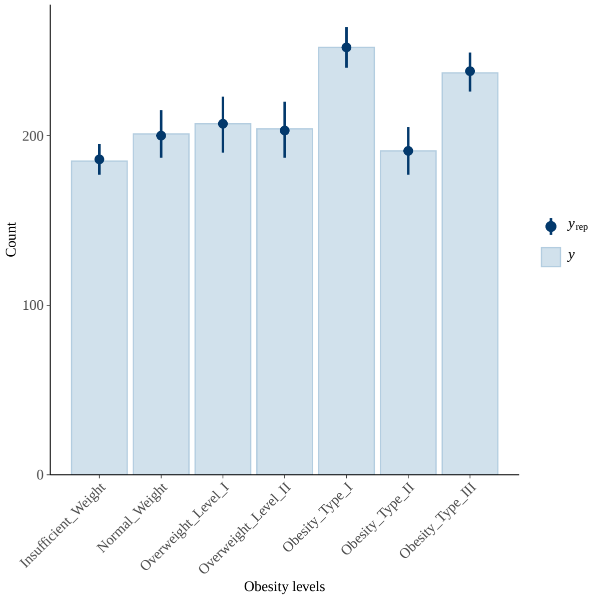


Figure 5: Posterior predictive check with 2000 draws.

The result of the PPC, where the replicated draws exhibit similar counts to the true labels in the training set, indicates a good fit of the model to the data. Specifically, the observation that the counts of each obesity type in the replicated draws closely match the frequencies of the true labels suggests that the model captures the underlying patterns and variability in the data effectively.

Furthermore, the PPC highlights the model's ability to accurately predict the distribution of patients across different obesity types. The fact that the most common obesity type (Obesity type 1) in the true labels also has the highest count in the replicated draws reinforces the model's ability to capture the dominant characteristics of the data. Correspondingly, the replication of the second most common obesity type (Obesity type 3) and the consistency in the counts of other obesity types further validate the model's predictive accuracy across different categories.

6 Limitations and potential improvements

While the project demonstrates promising results in modeling the relationship between various predictors and obesity types using Bayesian methods, there are several limitations and potential areas for improvement to consider. These include the need for incorporating additional relevant predictors such as dietary habits, genetic predispositions, or socioeconomic factors to enhance the model's predictive accuracy and provide a more comprehensive understanding of obesity determinants. Moreover, addressing issues related to data quality and quantity, including acquiring larger and more comprehensive datasets with higher quality data, could improve the robustness and generalizability of the model. Even more than that, conducting sensitivity analyses to assess the model's performance under different assumptions, exploring more flexible modeling techniques to capture complex interactions and nonlinear relationships, and conducting external validation using independent datasets are essential steps to ensure the model's reliability and generalizability.

For the sequential model, considering predictor-specific effects instead of global coefficients for the predictors could make the model more flexible. This adjustment might offer a deeper understanding of how individual predictors influence the sequential transitions between obesity types. Beyond that, incorporating more detailed information on predictors and refining the model architecture could lead to improved predictions and a more nuanced understanding of obesity progression.

7 Conclusion

The project extensively explored Bayesian methods to dissect the intricate relationship between various predictors and obesity types, revealing nuanced insights into the determinants of this complex health condition. By employing ordinal regression models, the study delved into an array of predictors spanning demographic factors, lifestyle choices, and familial influences. Noteworthy predictors included gender, age, height, weight, family history of overweight, dietary habits, physical activity levels, and transportation modes. The models exhibited robust convergence, affirming the reliability of the estimates derived from the Bayesian framework. Particularly striking were the significant associations unveiled between predictors such as height and weight and the different obesity types, corroborating expectations based on standard BMI calculations.

Analysis of the 95% confidence intervals of the estimated coefficients highlights their statistical significance in the model. Across all models, consistent interval ranges for `Height` and `Weight` predictors imply that the choice of priors used are insensitive, indicating that the coefficients are primarily determined by the data themselves. Notably, the negative `Height` coefficient is consistent with BMI expectations. Additionally, predictors like Gender, CAEC, family_history_with_overweight, and SMOKE have 95% credible intervals excluding zero, underscoring their significance in predicting the outcome variable.

Furthermore, the project conducted thorough posterior predictive checks, which served as a litmus test for the models' goodness of fit. The replicated draws closely mirrored the distribution of true label frequencies, providing compelling evidence of the models' efficiency in capturing the underlying patterns in the data. However, amidst these successes, the project also identified several limitations warranting consideration. One notable aspect is the potential for further refinement by incorporating additional predictors that could enhance the models' predictive power and explanatory capacity. Additionally, the project acknowledged the need for improvements in data quality and quantity, as these factors can profoundly influence the accuracy and generalizability of the findings.

Reflecting on personal learning experiences, the project offered valuable insights into the intricacies of Bayesian modeling and its real-world applications. Through hands-on experimentation with Bayesian techniques such as prior sensitivity analysis and posterior predictive checks, a deeper appreciation for the flexibility and interpretability of this approach was gained. Furthermore, grappling with the complexities of modeling obesity underscored the interdisciplinary nature of data science and its potential to drive impactful research in public health and clinical domains. Overall, the project underscored the transformative potential of Bayesian methods in deciphering complex phenomena and guiding evidence-based decision-making.

Bibliography

Bürkner, P.-C. (2017), ‘brms: An R package for Bayesian multilevel models using Stan’, *Journal of Statistical Software* **80**(1), 1–28.

Bürkner, P.-C. and Vuorre, M. (2019), ‘Ordinal regression models in psychology: A tutorial’, *Advances in Methods and Practices in Psychological Science* .

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013), *Regression Models, Methods and Applications, Second edition*, Springer. ISBN 978-3-662-63882-8.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2022), *Bayesian Data Analysis, Third edition.*
URL: *http://www.stat.columbia.edu/ gelman/book/*

Hoffman, M. D. and Gelman, A. (2011), ‘The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo’.
URL: *https://arxiv.org/abs/1111.4246v1*

Hunter, J. D. (2007), ‘Matplotlib: A 2d graphics environment’, *Computing in Science & Engineering* **9**(3), 90–95.

Palechor, F. M. and de la Hoz Manotas, A. (2023), ‘Obesity or cvd risk (classify/regressor/cluster)’.
URL: *https://www.kaggle.com/dsv/7009925 (visited on 17th March 2024)*

R Core Team (2022), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Vehtari, A., Gelman, A. and Gabry, J. (2016), ‘Practical bayesian model evaluation using leave-one-out cross-validation and waic’.
URL: *https://arxiv.org/abs/1507.04544*

WHO (2024), ‘Obesity, world health organisation’.
URL: *https://www.who.int/health-topics/obesitytab=tab_1 (visited on 17th March 2024)*

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag. R package version 3.4.0.

A Additional tables

Table 2: List of variables with data types.

Variable	Data type
Gender	Nominal
Age	Numeric
Height	Numeric
Weight	Numeric
Family history with overweight	Nominal
FAVC: Frequent Consumption of High Caloric Food	Nominal
FCVC: Frequency of Consumption of Vegetables	Numeric
NCP: Number of Main Meals	Numeric
CAEC: Consumption of Food Between Meals	Nominal
SMOKE: Smoker or not	Nominal
CH20: Consumption of Water Daily	Numeric
SCC: Calories Consumption Monitoring	Nominal
FAF: Physical Activity Frequency	Numeric
TUE: Time Using Technology Devices	Numeric
CALC: Consumption of Alcohol	Nominal
MTRANS: Transportation Used	Nominal
NObeyesdad: Obesity level deducted	Ordinal

B Additional figures

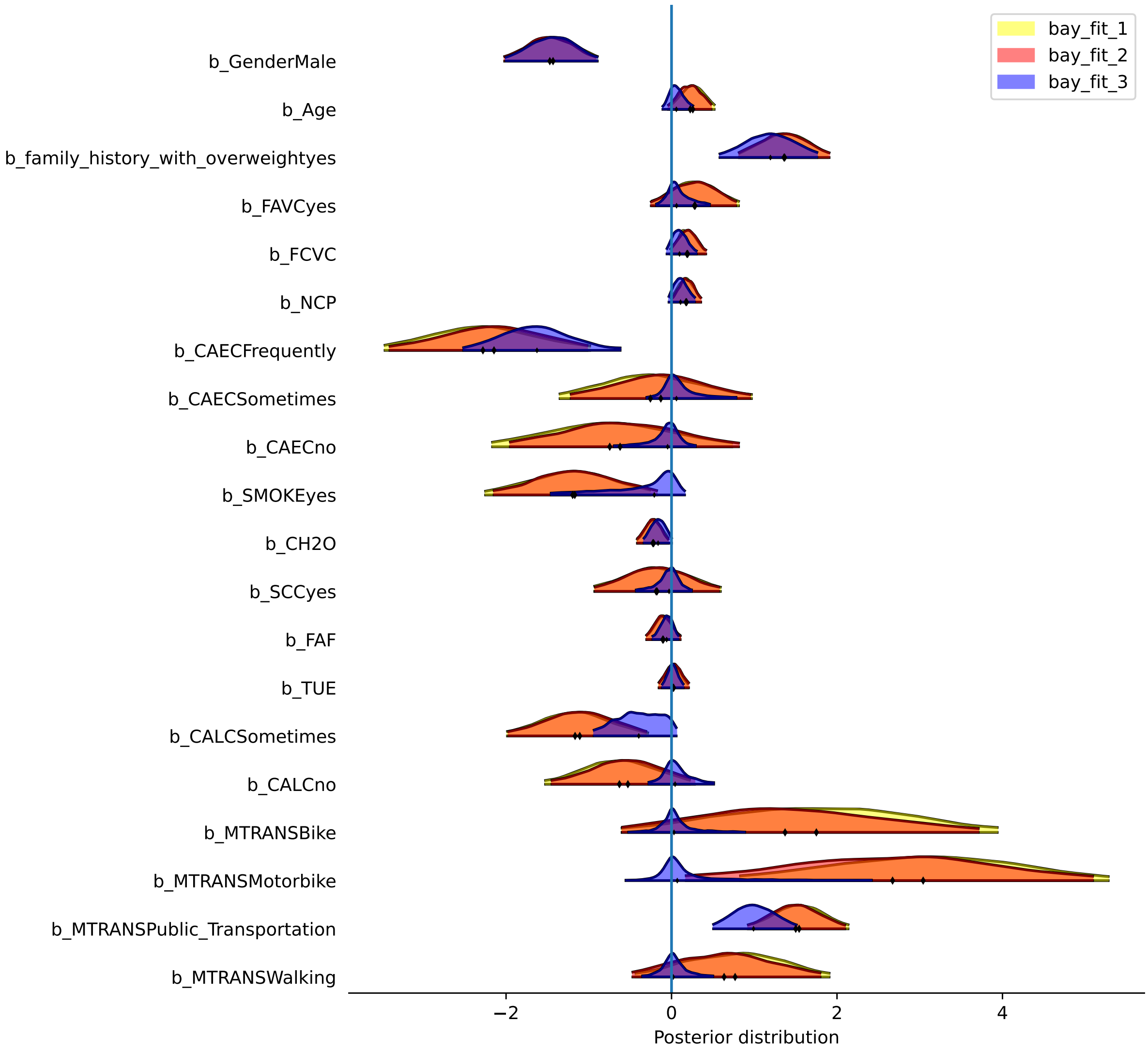
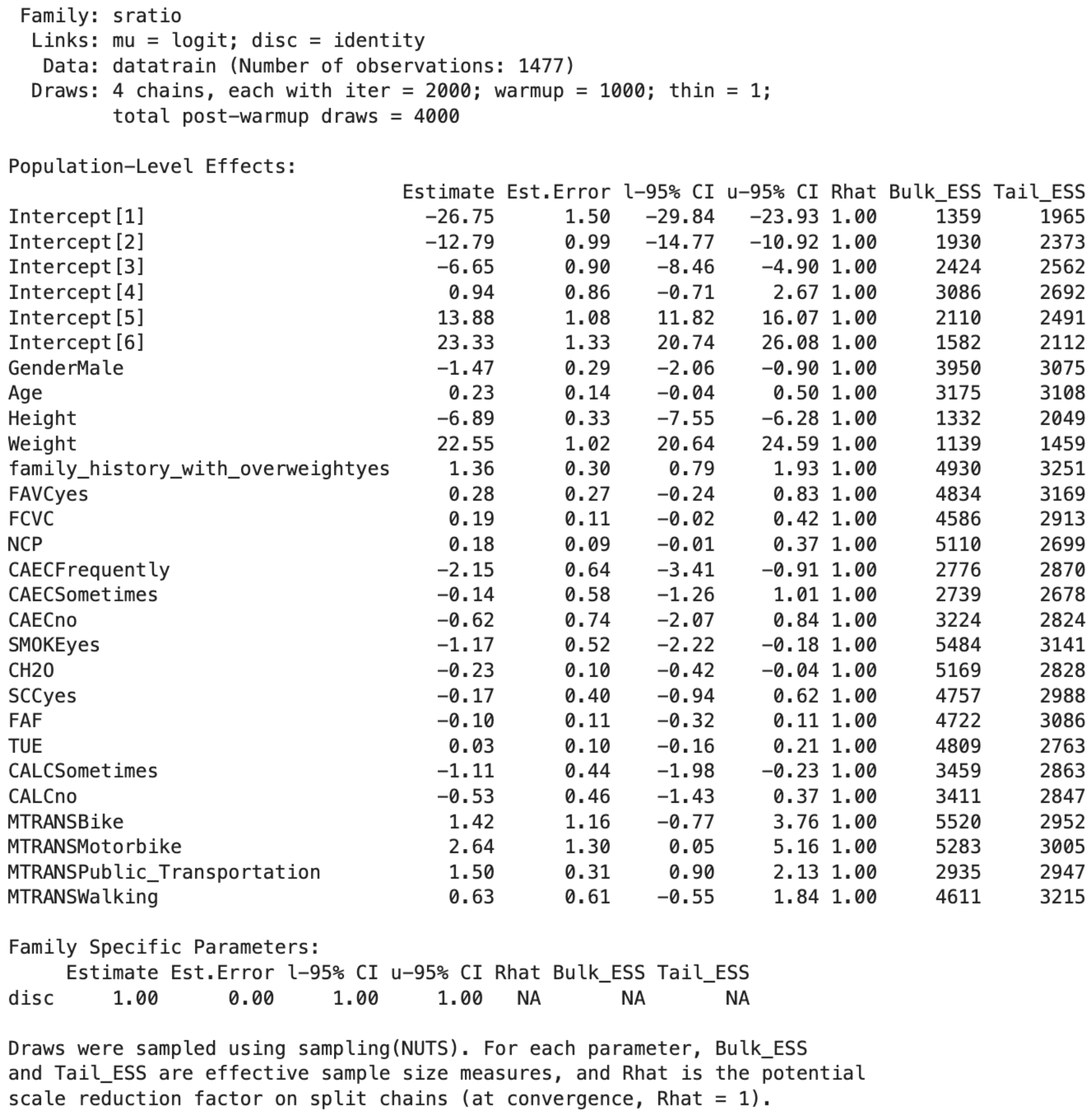


Figure 6: Posterior distribution of all the predictors with 95% credible intervals for all three models.

C Additional snippets



```
 Family: sratio
  Links: mu = logit; disc = identity
   Data: datatrain (Number of observations: 1477)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
                                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept[1]                        -26.75      1.50   -29.84   -23.93 1.00     1359     1965
Intercept[2]                        -12.79      0.99   -14.77   -10.92 1.00     1930     2373
Intercept[3]                         -6.65      0.90    -8.46    -4.90 1.00     2424     2562
Intercept[4]                          0.94      0.86    -0.71     2.67 1.00     3086     2692
Intercept[5]                         13.88      1.08    11.82    16.07 1.00     2110     2491
Intercept[6]                         23.33      1.33    20.74    26.08 1.00     1582     2112
GenderMale                           -1.47      0.29    -2.06    -0.90 1.00     3950     3075
Age                                   0.23      0.14    -0.04     0.50 1.00     3175     3108
Height                               -6.89      0.33    -7.55    -6.28 1.00     1332     2049
Weight                               22.55      1.02    20.64    24.59 1.00     1139     1459
family_history_with_overweightyes     1.36      0.30     0.79     1.93 1.00     4930     3251
FAVCyes                               0.28      0.27    -0.24     0.83 1.00     4834     3169
FCVC                                  0.19      0.11    -0.02     0.42 1.00     4586     2913
NCP                                   0.18      0.09    -0.01     0.37 1.00     5110     2699
CAECFrequently                       -2.15      0.64    -3.41    -0.91 1.00     2776     2870
CAECSometimes                        -0.14      0.58    -1.26     1.01 1.00     2739     2678
CAECno                               -0.62      0.74    -2.07     0.84 1.00     3224     2824
SMOKEyes                             -1.17      0.52    -2.22    -0.18 1.00     5484     3141
CH2O                                 -0.23      0.10    -0.42    -0.04 1.00     5169     2828
SCCyes                               -0.17      0.40    -0.94     0.62 1.00     4757     2988
FAF                                  -0.10      0.11    -0.32     0.11 1.00     4722     3086
TUE                                   0.03      0.10    -0.16     0.21 1.00     4809     2763
CALCSometimes                        -1.11      0.44    -1.98    -0.23 1.00     3459     2863
CALCno                               -0.53      0.46    -1.43     0.37 1.00     3411     2847
MTRANSBike                            1.42      1.16    -0.77     3.76 1.00     5520     2952
MTRANSMotorbike                       2.64      1.30     0.05     5.16 1.00     5283     3005
MTRANSPublic_Transportation           1.50      0.31     0.90     2.13 1.00     2935     2947
MTRANSWalking                         0.63      0.61    -0.55     1.84 1.00     4611     3215

Family Specific Parameters:
     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
disc     1.00      0.00     1.00     1.00   NA       NA       NA

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```