Bayesian classification of obesity levels: A sequential logit approach

Sathish Kabatkar Ravindranth

February 13, 2024

Introduction to Dataset

- **Attributes**
  - **Eating Habits:**
    - FAVC: Frequent Consumption of High Caloric Food
    - FCVC: Frequency of Consumption of Vegetables
    - NCP: Number of Main Meals
    - CAEC: Consumption of Food Between Meals
    - CH20: Consumption of Water Daily
    - CALC: Consumption of Alcohol
  - **Physical Condition:**
    - SCC: Calories Consumption Monitoring
    - FAF: Physical Activity Frequency
    - TUE: Time Using Technology Devices
    - MTRANS: Transportation Used
  - **Variables:** Gender, Age, Height, Weight and Family history

- **Target Variable (NObesity)**
  - Insufficient_weight
  - Normal_weight
  - Overweight_level_I
  - Overweight_level_II
  - Obesity_type_I
  - Obesity_type_II
  - Obesity_type_III

- **Dataset Overview**
  - **Countries:** Mexico, Peru, Colombia
  - **Age Range:** 14-61
  - **Records:** 2111

Exploratory data analysis

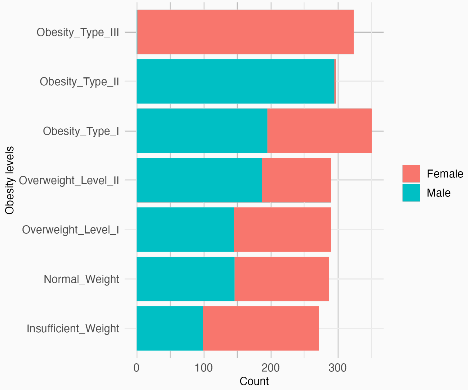

Figure 1: Bar chart for Obesity level counts

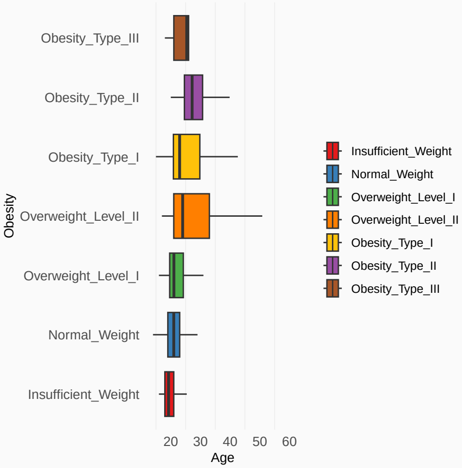

Figure 2: Distribution of age in obesity levels

Exploratory data analysis - continued

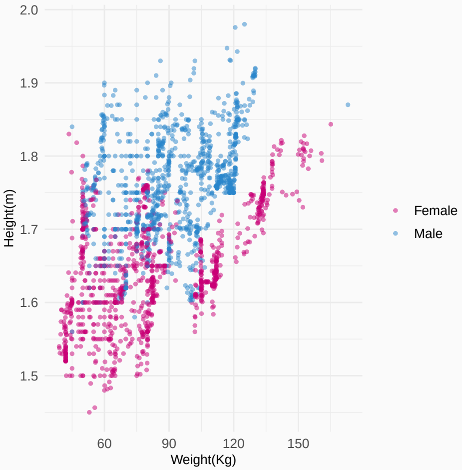

Figure 3: Weight vs Height by Gender

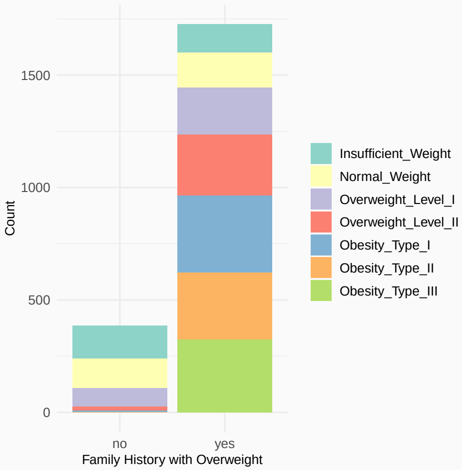

Figure 4: Obesity levels with Family history

Sequential model (Bürkner and Vuorre., 2019, p. 32)

- Ordinal response $Y \in \{1, \ldots, K\}$
- Insufficient Weight $\rightarrow$ Normal Weight $\rightarrow$ Overweight Level I $\rightarrow$ Overweight Level II $\rightarrow$ Obesity Type I $\rightarrow$ Obesity Type II $\rightarrow$ Obesity Type III
- The transition between $\{r\}$ and $\{r+1, \ldots K\}$ is a binary decision which is modelled via a latent continuous variable $\tilde{Y}_r$, with a global coefficient vector $\boldsymbol{\beta}$ and category specific intercepts $\beta_{01}, \ldots, \beta_{0(K-1)}$ as a linear predictor $\eta = \beta_{0r} + \mathbf{x}^T \boldsymbol{\beta}$ is:

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

- Assuming error ϵ_r has mean 0 with cumulative distribution function $F(\cdot)$
- Considering a threshold τ for each level, the process stops $\tilde{Y}_r \leq \tau_r$ and the result is $Y = r$

$$P(Y = r | Y \geq r, \mathbf{x}) = P(\tilde{Y}_r \leq \tau_r | \eta)$$

$$= F(\tau_r - \eta)$$

Sequential logit model

- Using conditional probability:

$$P(Y = r|\eta) = F(\tau_r - \eta)\prod_{j=1}^{r-1}(1 - F(\tau_j - \eta)$$

- The $F(\cdot)$ is logistic distribution function $F(\cdot) = \frac{exp(\cdot)}{1+exp(\cdot)}$

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

- For category K, $P(Y = K|Y \geq K, \mathbf{x}) = 1$, the max is reached.

The problem of separation (Gelman et al., 2022, p. 412)

- A linear combination of predictors perfectly predicts outcomes
- Removing predictors, may lead to excluding influential variables.
- Bayesian inference with weakly informative priors for stable and meaningful coefficient estimates

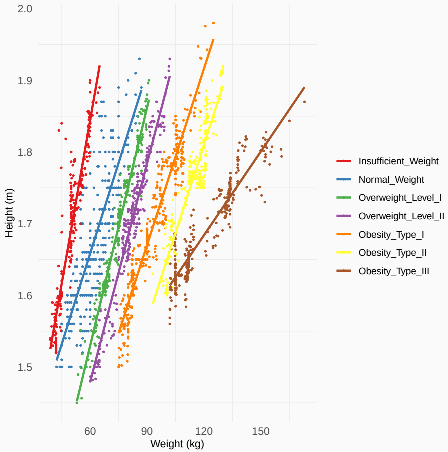

Figure 5: Obesity levels with Weight vs Height

Bayesian learning (Gelman et al., 2022, p. 7)

Bayesian machine learning is a systematic method for estimating the posterior predictive distribution of new data, allowing the derivation of uncertainty properties in predictions.

To fit a multinomial classification model, there is a feature input matrix $\mathbf{X} \in \mathbb{R}^{nxd}$ (where $n \in \mathbb{N}$ is the number of inputs, and $d \in \mathbb{N}$ is the number of features), a 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 $\mathbf{X}$, is given as:

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

- $P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$: Posterior distribution of parameters
- $P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$: Likelihood of labels given inputs and weights
- $P(\mathbf{y}|\mathbf{X})$: Evidence, a normalization constant
- $P(\boldsymbol{\theta}|\mathbf{X})$: Prior distribution on model weights, assumed independent of inputs.

Prior specification (Gelman et al., 2022, p. 416)

- We start with a weakly informative prior for all coefficients and intercepts $\boldsymbol{\beta}, \beta_{0r} \sim normal(0, 5)$
- We try two different scales of cauchy priors, $\boldsymbol{\beta} \sim cauchy(0, 2.5), \beta_{0r} \sim normal(0, 5)$ and $\boldsymbol{\beta} \sim cauchy(0, 0.1), \beta_{0r} \sim normal(0, 5)$
- We split the data into training (70%) and test (30%) sets
- Pre-processing step is to standardise input variables
- We fit the model using *brms* package with *sratio*() family

Comparison of models

- For all models, rhat values are equal to 1
- Lesser the leave one out information criterion (looci) better the model fit
- Considering the fit with *cauchy*$(0, 2.5)$ prior as reference, the other two fits will perform worse in new data
- It is also evident from the predictive accuracy on test set as well

Table 1: LOO values and differences between three models

Model	elpd_diff	se_diff	looic	se_looic
cauchy(0,2.5)	0.000	0.000	757.020	43.720
normal(0,5)	-1.223	0.998	759.466	42.933
cauchy(0,0.1)	-3.816	4.023	764.652	42.730

Table 2: Accuracy on test set

Model	Accuracy
cauchy(0,2.5)	0.850
normal(0,5)	0.839
cauchy(0,0.1)	0.837

Posterior distribution of estimated coefficients

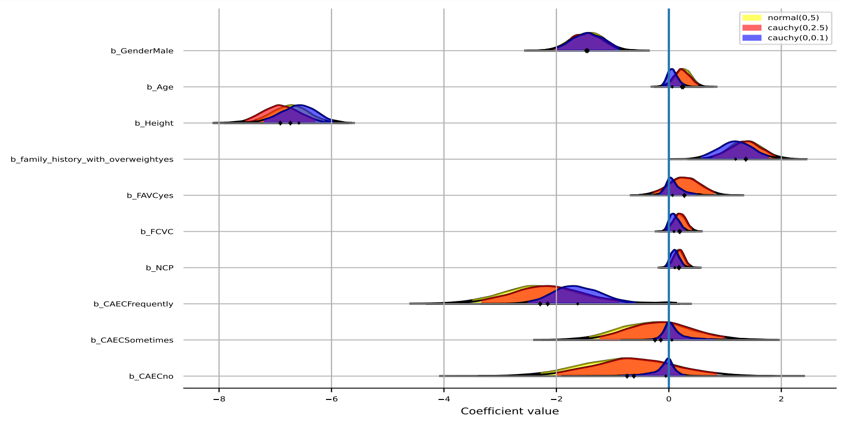

Figure 6: Posterior distribution of estimated coefficients

Posterior distribution of estimated coefficients - contd.

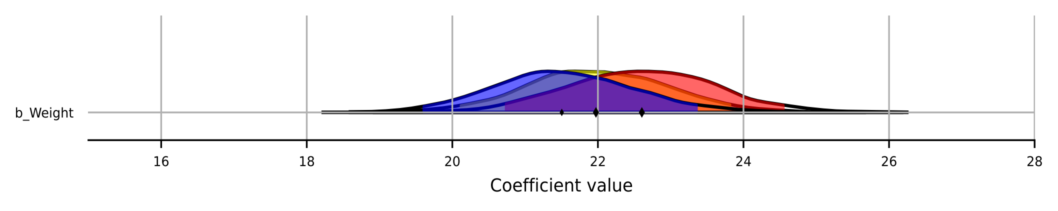

Figure 7: Posterior distribution of estimated coefficient for predictor Weight

Posterior distribution of estimated coefficients - contd.

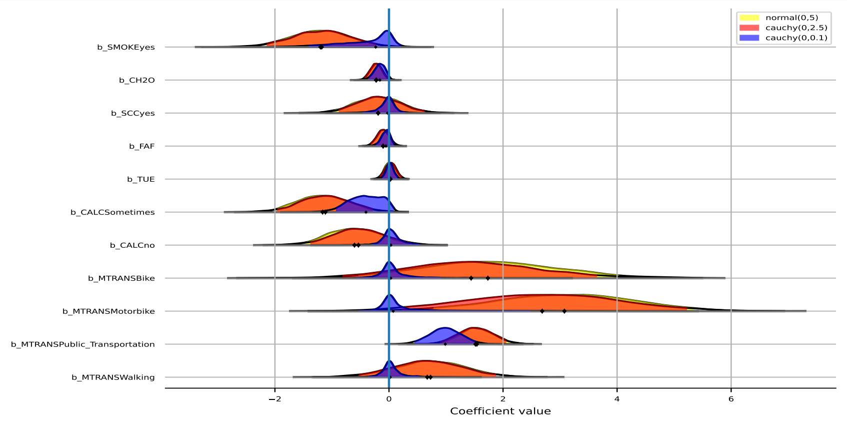

Figure 8: Posterior distribution of estimated coefficients

Posterior predictive check

- The model is capturing distinct modes associated with 7 ordinal levels, suggest the model is sensitive to the variations among these levels
- Model is complex enough to capture the nuances in the data, essential to ensure the complexity is not leading to over-fitting

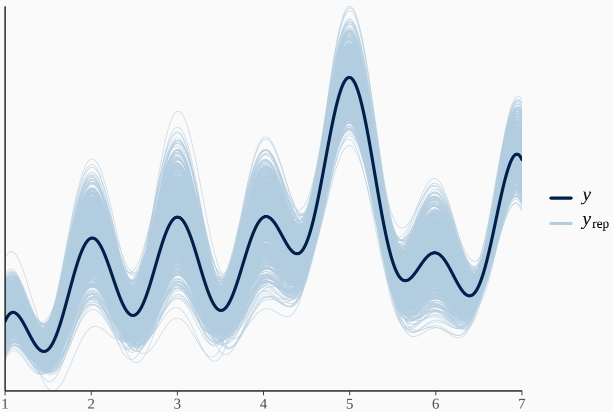

Figure 9: Posterior predictive check for 1000 draws

Summary

- Effectively addressed separation issues using Bayesian modeling, preserving crucial predictors
- Explored the Sequential logit model's suitability on the dataset
- Investigate further with category-specific effects
- Explore sensitivity to different priors, studying robustness in varied contexts

References

- Gelman, A., B Carlin, J., S Stern, H., B Dunson, D, Vehtari, A., B Rubin, D. Bayesian Data Analysis Third edition, 2022.
- Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. Regression Models, Methods and Applications, Second edition, Springer, 2013. ISBN 978-3-662-63882-8.
- Bürkner, P. and Vuorre, M. Ordinal Regression Models in Psychology: A Tutorial, 2019.