Bayesian final project

Yingmai Chen

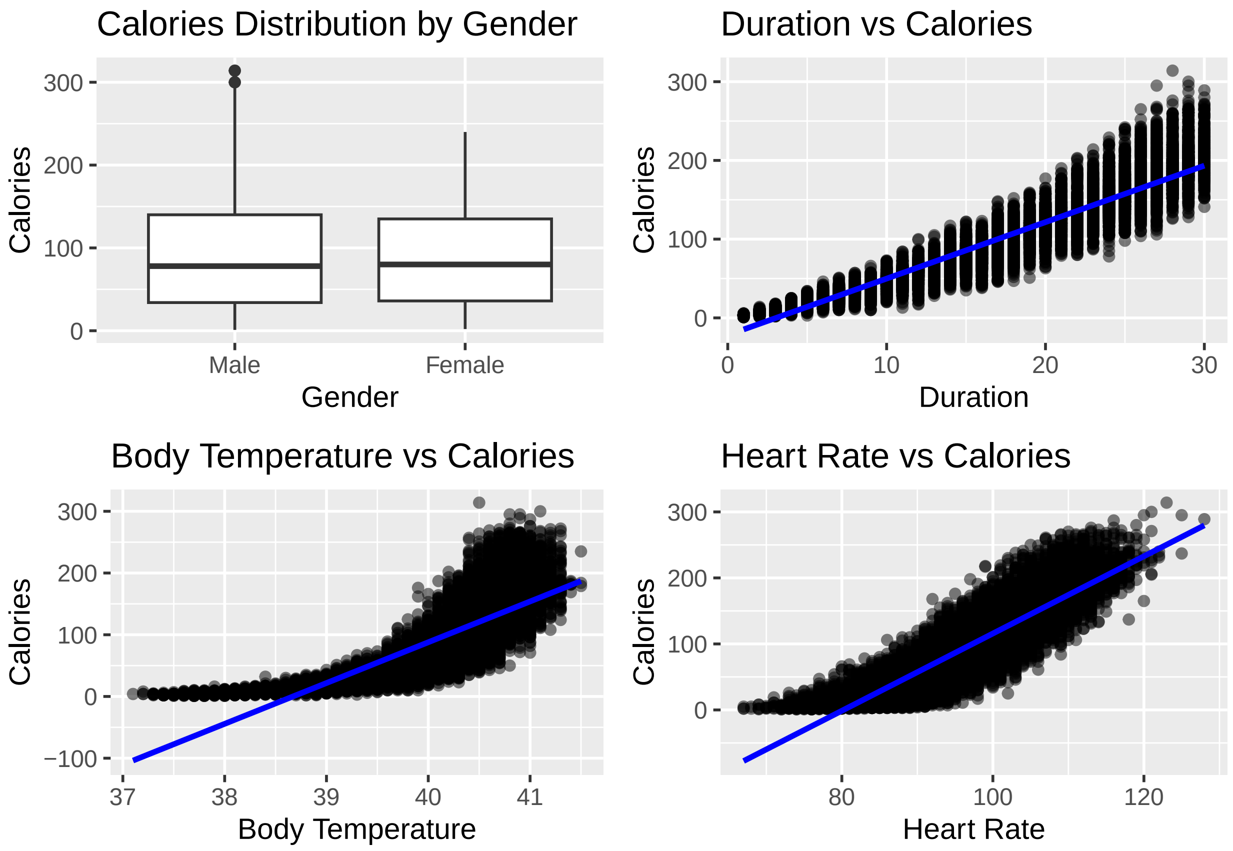
2023-12-10

1 Introduction

The proposed project aims to establish a predictive relationship between physical exercise attributes and calories output.The reason why I choose this project is that:nowadays,the health industry's standard exercise and nutrition advice doesn't fit everyone's unique body responses. Personalized plans are needed for better health outcomes, which requires understanding how individual traits and exercise reactions affect calorie burning.Besides,The study will analyze two datasets: 'exercise.csv' containing variables such as user demographics and post-exercise vitals, and 'calories.csv' detailing corresponding caloric expenditure.

1.1 Visualization

For this part, I will show some visualizations of the data.



2 Method and analysis

2.1 Bayesian linear regression model

2.1.1 Method

Inference was conducted using Bayesian inference via Markov Chain Monte Carlo (MCMC) sampling, utilizing the No-U-Turn Sampler (NUTS).

2.1.2 Statistical analysis

2.1.2.1 Estimators The model uses Bayesian estimation, which means the estimators are the posterior distributions of the parameters beta and sigma.

2.1.2.2 Prior The prior for β is a normal distribution with mean 0 and standard deviation 10:

$$\beta \sim \mathcal{N}(0, 10)$$

The prior for σ, the standard deviation of the normal distribution for the likelihood, is an inverse gamma distribution with both shape and scale parameters set to 0.01:

$$\sigma \sim \text{Inv-Gamma}(0.01, 0.01)$$

2.1.2.3 Loss function The loss function in the context of the Bayesian regression model is the negative log-likelihood of the data given the parameters.

$$L(\beta, \sigma) = \sum_{i=1}^{N} \frac{1}{2\sigma^2}(y_i - X_i\beta)^2 + \frac{N}{2}\log(2\pi\sigma^2)$$

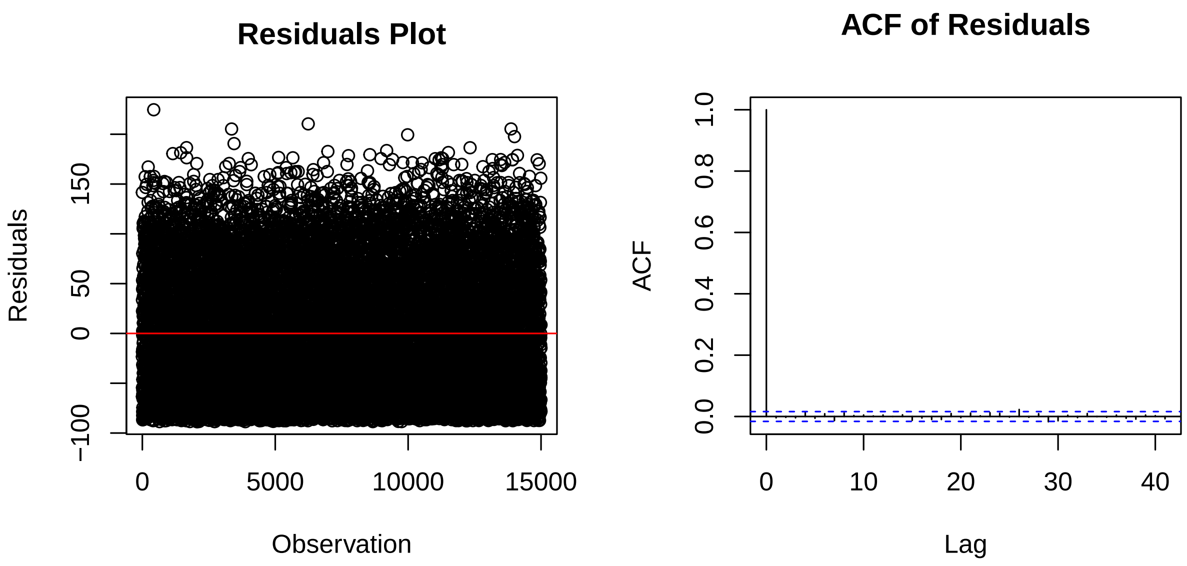
In this formula,L is the loss function.β is the vector of regression coefficients.σ is the standard deviation of the normal distribution. y_i is the observed value of the response variable for the i-th observation.X_i is the vector of predictor values for the i-th observation.N is the total number of observations.

2.1.2.4 Predictors The predictors are the variables Duration, Heart Rate, and Body Temperature, which are hypothesized to be associated with the response variable Calories.

2.1.3 Sensitive analysis to the prior

The 95% credible intervals suggest a credible impact of predictors on calories burned and precise estimates of residual standard deviation. Small standard errors indicate high precision in parameter estimates. Convergence diagnostics confirm good model convergence, indicating overall reliability and well-specification of the model.

2.1.4 Model checking



The model is free of bias and autocorrelation in residuals but has outliers, indicating some prediction inaccuracies possibly due to extreme values or anomalies.

2.2 Hierarchical model based on gender

```
## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## https://mc-stan.org/misc/warnings.html#bulk-ess
```

2.2.1 Method

This is a Bayesian hierarchical model for layered data analysis, using Markov Chain Monte Carlo (MCMC) sampling and posterior analysis via Stan, accommodating complex hierarchical data structures.

2.2.2 Statistical analysis

2.2.2.1 Estimators The estimates include Beta Coefficients (beta[J]),Standard Deviation of Residuals (sigma),standard Deviation of Beta Coefficients Across Groups (sigma_beta),Mean Value of Beta Coefficients Across Groups (mu_beta).

2.2.2.2 Prior

1. Mean value of beta coefficients across groups (`mu_beta`):$\mu_\beta \sim \text{t}(\text{degrees of freedom} = 3, \text{mean} = 0, \text{scale} = 10)$
2. Standard deviation for the distribution of beta across groups (`sigma_beta`):$\sigma_\beta \sim \text{Exponential}(\text{rate} = 1)$
3. Coefficients for predictors, for each gender group (`beta[j]`):$\beta_j \sim \text{t}(\text{degrees of freedom} = 3, \mu_\beta, \sigma_\beta)$
4. Standard deviation of the residuals (`sigma`):$\sigma \sim \text{Exponential}(\text{rate} = 1)$

2.2.2.3 Loss function

$$\mathcal{L}(\boldsymbol{\beta},\sigma) = -\sum_{i=1}^{N} \log\left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}\sigma}\left(1+\frac{1}{\nu}\left(\frac{y_i - \mathbf{X}_i\boldsymbol{\beta}_{\text{gender}[i]}}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}\right)$$

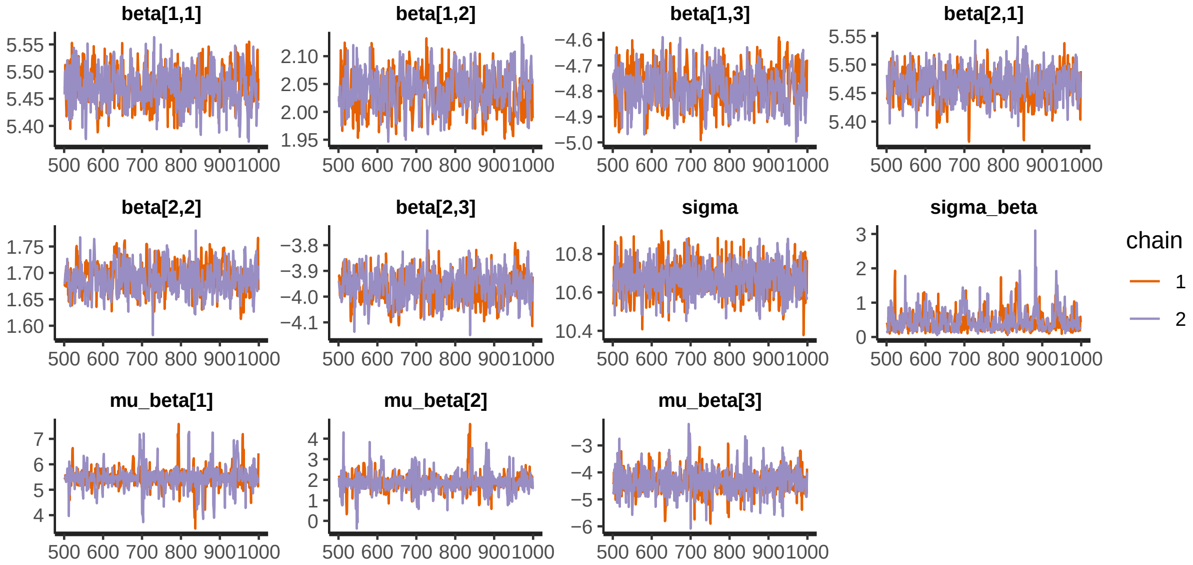
From this formula,N is the number of observations.$\mathbf{X}_i$ is the vector of predictor values for the i-th observation.$\boldsymbol{\beta}_{\text{gender}[i]}$ is the vector of coefficients corresponding to the gender group of the i-th observation.y_i is the actual value of the response variable for the i-th observation.σ is the scale parameter (standard deviation of the residuals).ν is the degrees of freedom of the Student-t distribution (set to 4 in the model).Γ is the gamma function.

2.2.2.4 Predictors The predictors are the variables gender, Duration, Heart Rate, and Body Temperature, which are hypothesized to be associated with the response variable Calories

2.2.3 Sensitive analysis to the prior

In summary, the model appears to be capturing distinct effects for different gender groups, indicated by the coefficients. However, the relatively large value suggests there is substantial variability not accounted for by the model. It's also important to compare these results with those from a non-hierarchical model to understand the impact of introducing the hierarchical structure.

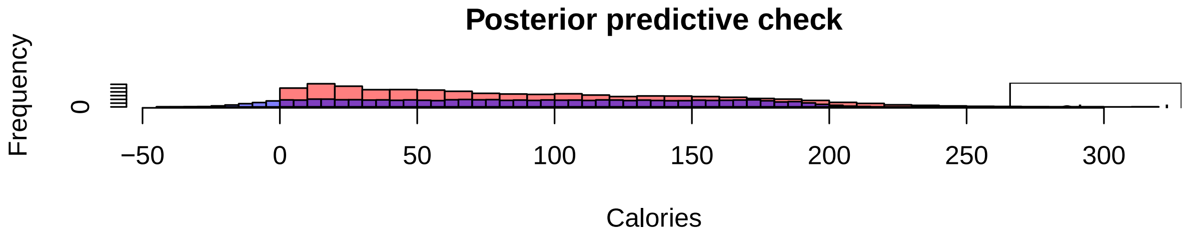
2.2.4 Model checking



It seems that the chains for each parameter are mixing reasonably well. There are no apparent trends or drifts in the chains, which is a good sign of convergence.

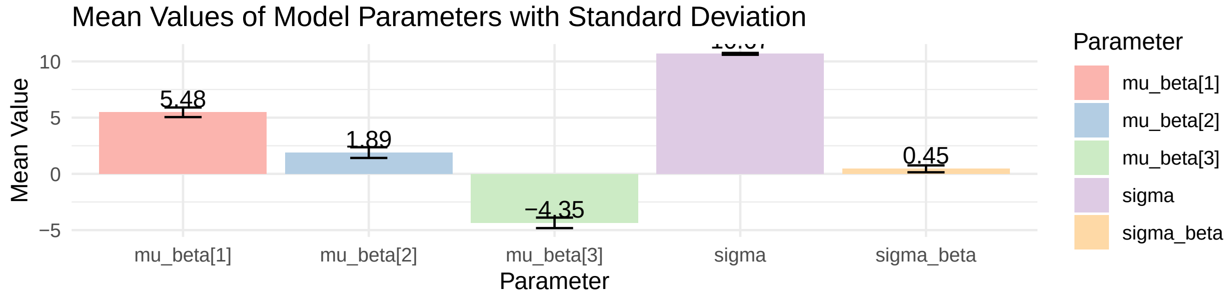
3.Results and Conclusions

3.1 Result of Bayesian linear regression model(Posterior predictive checks)



The model aligns well with observed data for common values and captures the central tendency effectively, as shown by the overlap of distributions and the shape around the histogram's peak. However, there are discrepancies in the tails, particularly the right tail, indicating a possible overestimation of higher values. The model also struggles with extremes and outliers, particularly at the lower end, and may not accurately estimate the likelihood of lower value occurrences. Overall, while the model is generally successful at capturing core data behavior, its accuracy diminishes at the distribution's extremes.

3.2 Result of Hierarchical model based on gender(Point estimate)



The point estimate for mu_beta[1] is precise, evidenced by its small standard deviation. However, mu_beta[2] and especially mu_beta[3] show larger standard deviations, reflecting greater uncertainty in these estimates. The sigma parameter's small point estimate and standard deviation suggest a stable estimation of model residuals. Conversely, the large standard deviation of sigma_beta's point estimate indicates significant uncertainty in estimating group coefficient variability.

3.3 Conclusions

In this project, I examined Bayesian regression and hierarchical modeling, emphasizing their efficacy in uncertainty management and prior knowledge integration. These approaches were instrumental in parameter estimation and uncertainty analysis, especially for group variations like gender.This work enhanced my understanding of Bayesian inference, underscoring its relevance in complex data analysis.