Bayesian final project

Yingmai Chen

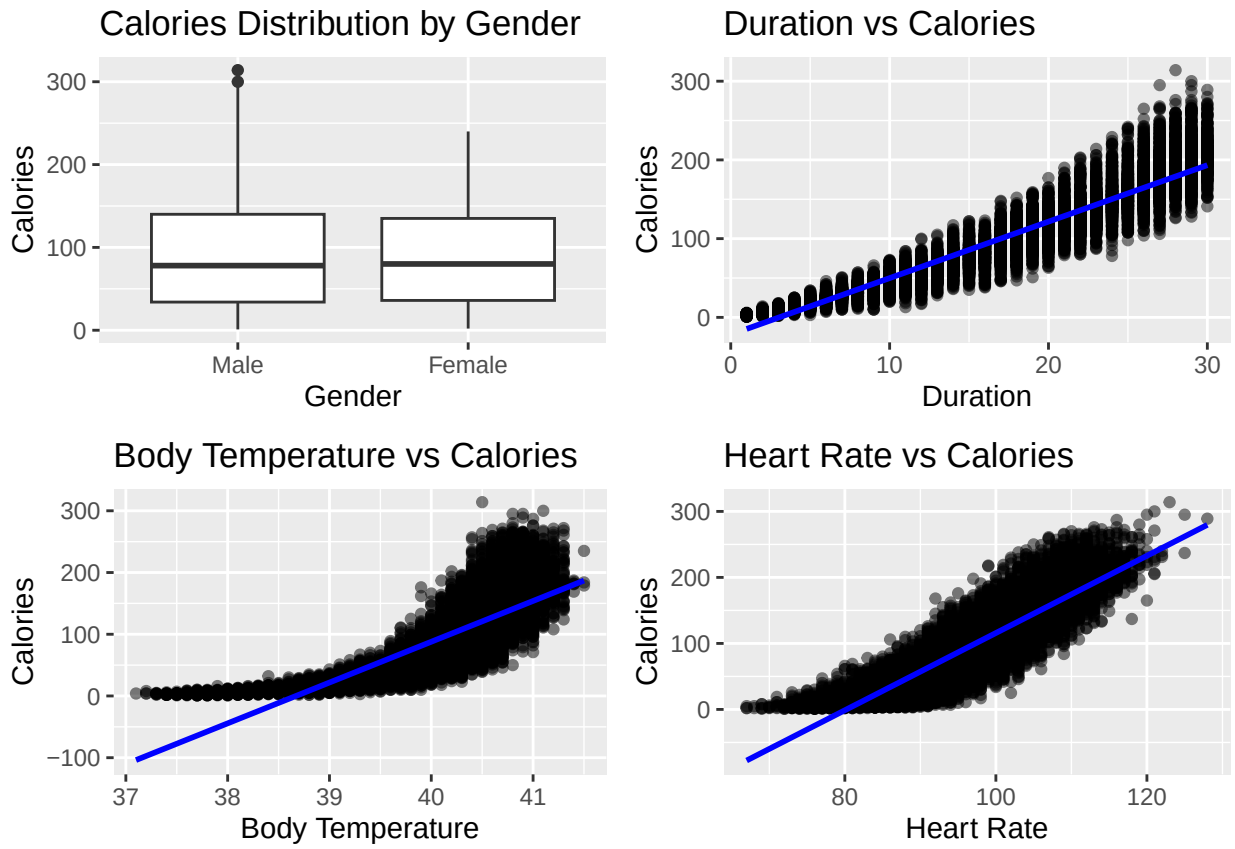
2023-12-10

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 visualization of the data.



2 Method and analysis

2.1 Bayesian linear regression model

2.1.1 method

The approximation method used for inference is Bayesian inference through Markov Chain Monte Carlo (MCMC) sampling, specifically employing the No-U-Turn Sampler (NUTS), which is an extension of Hamiltonian Monte Carlo (HMC).

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.

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

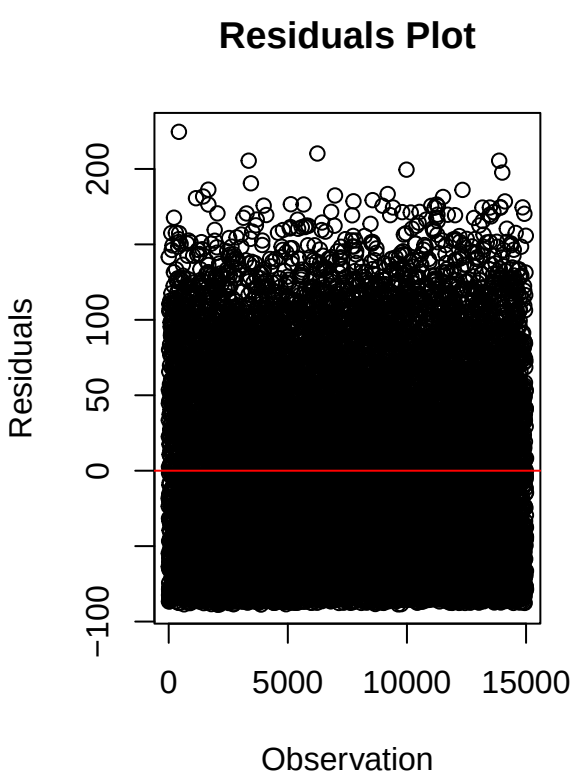
In this formula,L is the loss function.β is the vector of regression coefficients.σ is the standard deviation of the normal distribution. y_i is the observed value of the response variable for the i-th observation.X_i is the vector of predictor values for the i-th observation.N is the total number of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

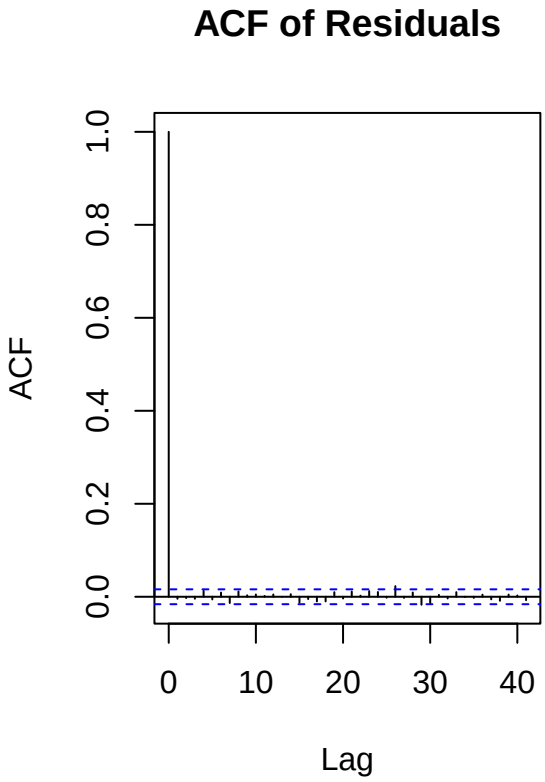
2.1.2.4 predictors The predictors are the variables Duration, Heart Rate, and Body Temperature, which are hypothesized to be associated with the response variable Calories.

2.1.3 sensitive analysis to the prior

According to 95% credible intervals, it shows a credible effect of the predictors on the calories burned and a precise estimate of the standard deviation of the residuals. According to the standard errors associated with these mean estimates are relatively small, which implies that the parameter estimates are precise. According to the convergence diagnostics,the model has converged well. In summary, the model appears to be well-specified and the results are reliable.

2.1.4 model checking



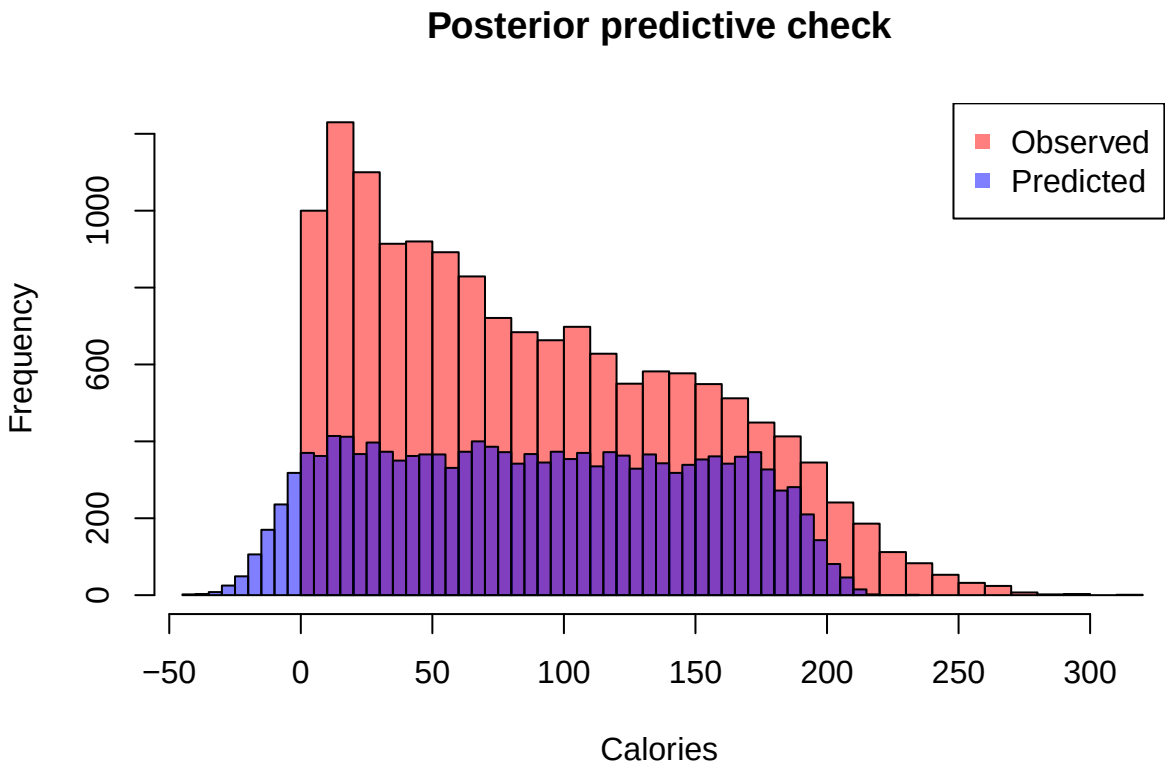


Overall, based on these two diagnostic plots, the model seems to be performing adequately. The residuals do not show any clear patterns, and the ACF plot suggests there is no significant autocorrelation.

3.result and conclusions

3.1 result of Bayesian linear regression model

3.1.1 posterior predictive checks



3.1.1.1 description For Overlap of Distributions, the most common values, the model's predictions align well with what is observed.For Shape of Distributions,The similar shapes of the observed and predicted distributions around the peak of the histogram demonstrate that the model is generally successful at capturing the core behavior of the data. However, discrepancies in the tails, especially the right tail, suggest the model may be overestimating the frequency of higher values.For Coverage of Extremes,the model's predictions do not align as closely with the observed data at the extreme low end and the presence of outliers that have not been sufficiently accounted for in the model.For the frequency of Occurrence,Disparities in the frequency of occurrence within certain bins of the histogram, notably where observed frequencies in the lower range exceed predicted frequencies, indicate that the model may not be accurately estimating the likelihood of lower value occurrences.In summary, while the model seems to do a reasonable job of capturing the central tendency of the data, it might not be as accurate in the tails of the distribution.