My Sweave Report

Thy Luong, Evgeniia Selezneva, Jasmine Wang

November 25, 2025

1 Problem Statement

Everybody needs to eat food, but it is often complicated to understand what exactly we are putting into our bodies and stay informed about the quality of our diet. Our project aims to get a better understanding of food nutrition to help demystify this process. We wanted to explore how different types of nutrition affected caloric value. We found...

2 Introduction

Our plan to was to analyze the relationship between different types of nutrition and caloric value using data from a food nutrition dataset. We wanted to use Bayesian linear regression for Normal-Inverse-Gamma conjugate prior-posterior to see if any of the nutrition types are useful for predicting caloric value. Using Bayesian variable selection methods such as Bayesian lasso and spike-and-slab regression, we came up with several models to analyze the relationship between nutrition types and calories. We then compared how well these models predicted our test data.

3 Data Collection

We used food nutrition data from Kaggle compiled from scrapping the internet. It contains nutritional data for a total of 2,395 various rows (food) and 35 columns (food, 32 different nutrients, caloric value, and nutrition density) meaning we worked with up to 32 predictors for caloric value. During exploratory data analysis, we decided to investigate nutrients that we believed were more relevant to caloric value based on prior common knowledge we had. These nutrients were fat (all types), carbohydrates, and protein. We also looked at caloric value itself and in relation to these nutrients.

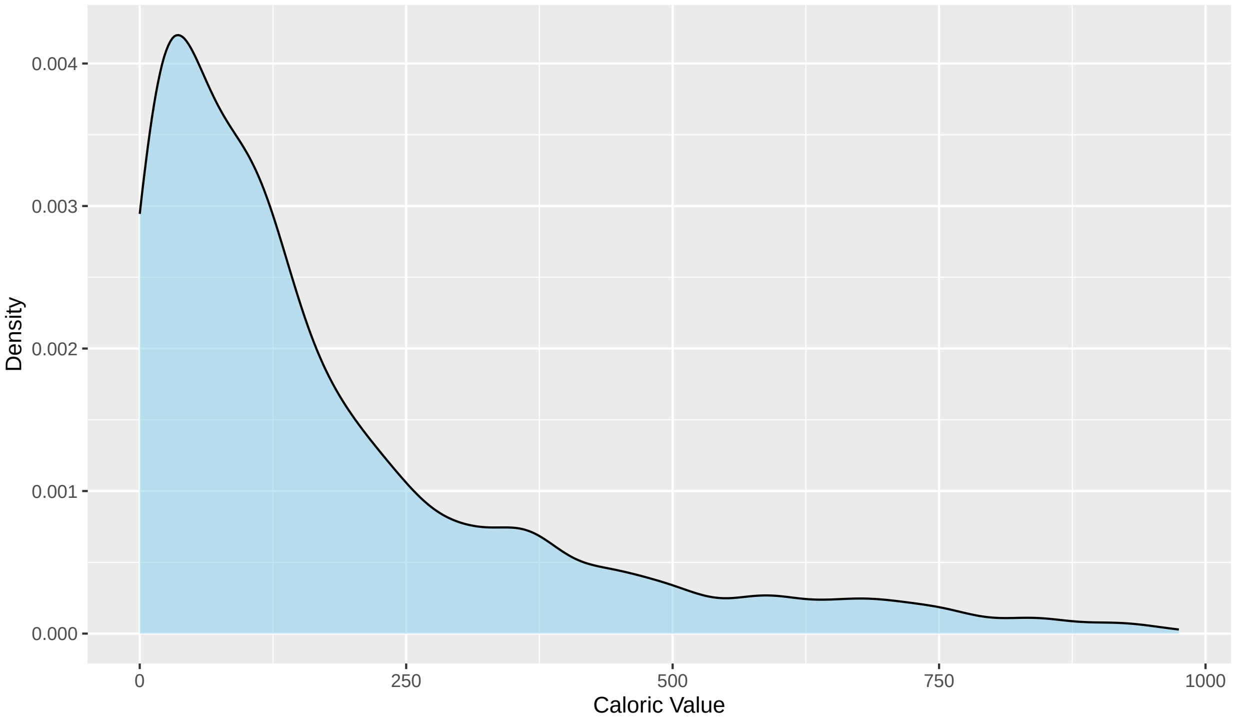

Figure 1: Density Plot of the Caloric Value

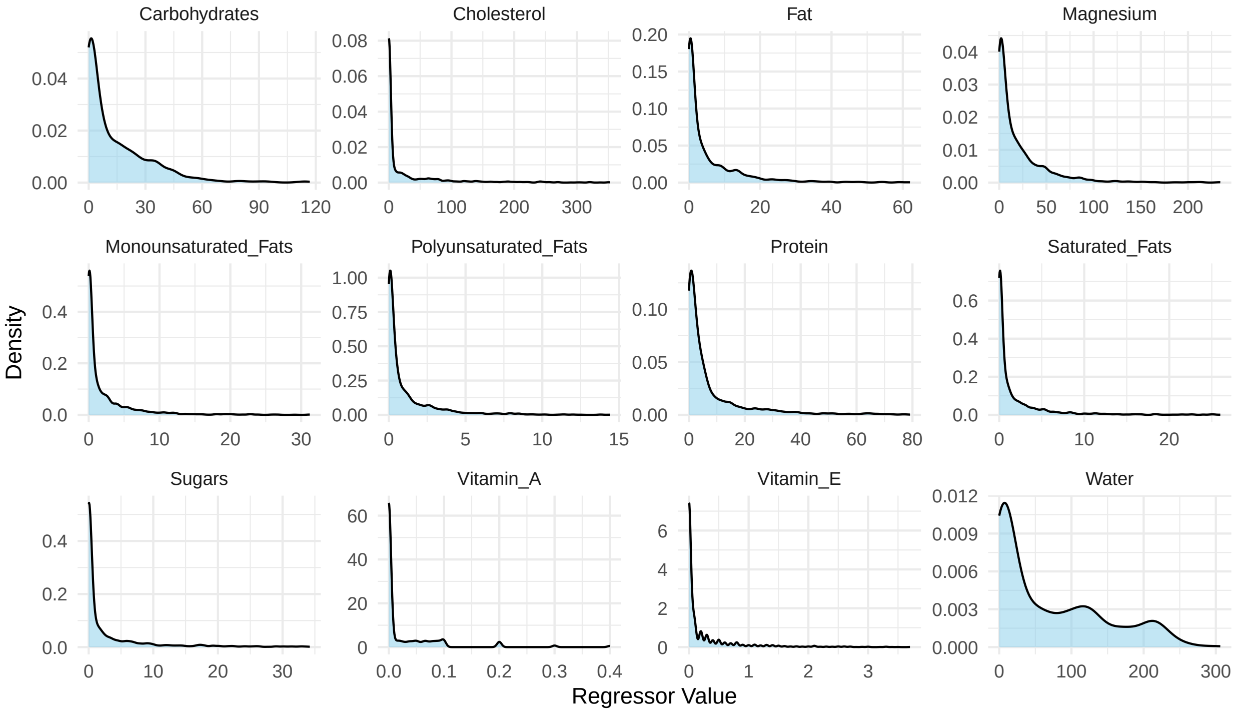

Figure 2: Density Plot of Some Regressors

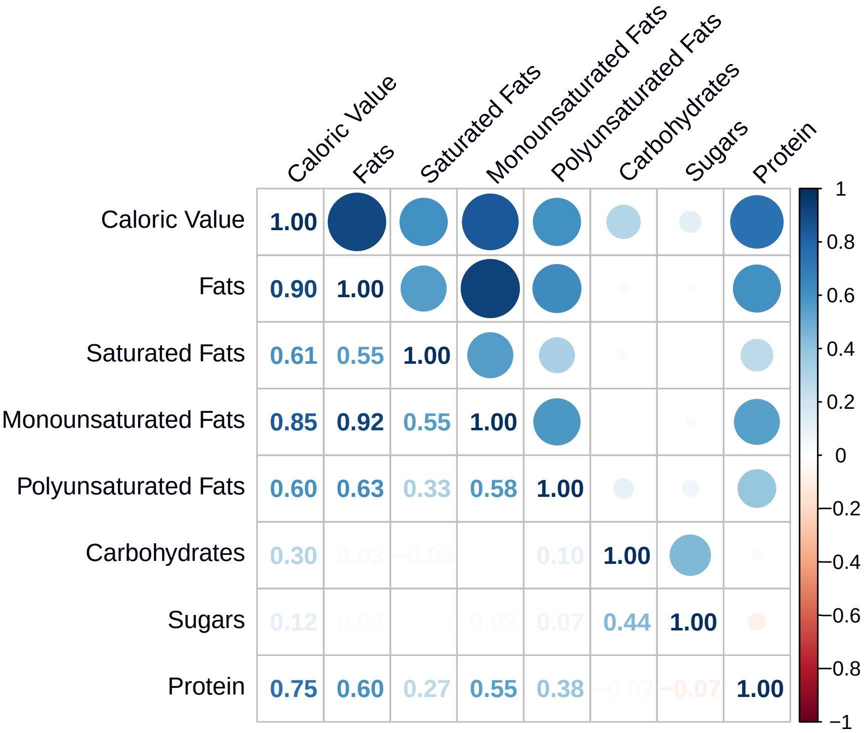

Figure 3: Correlation Plot of Some Regressors

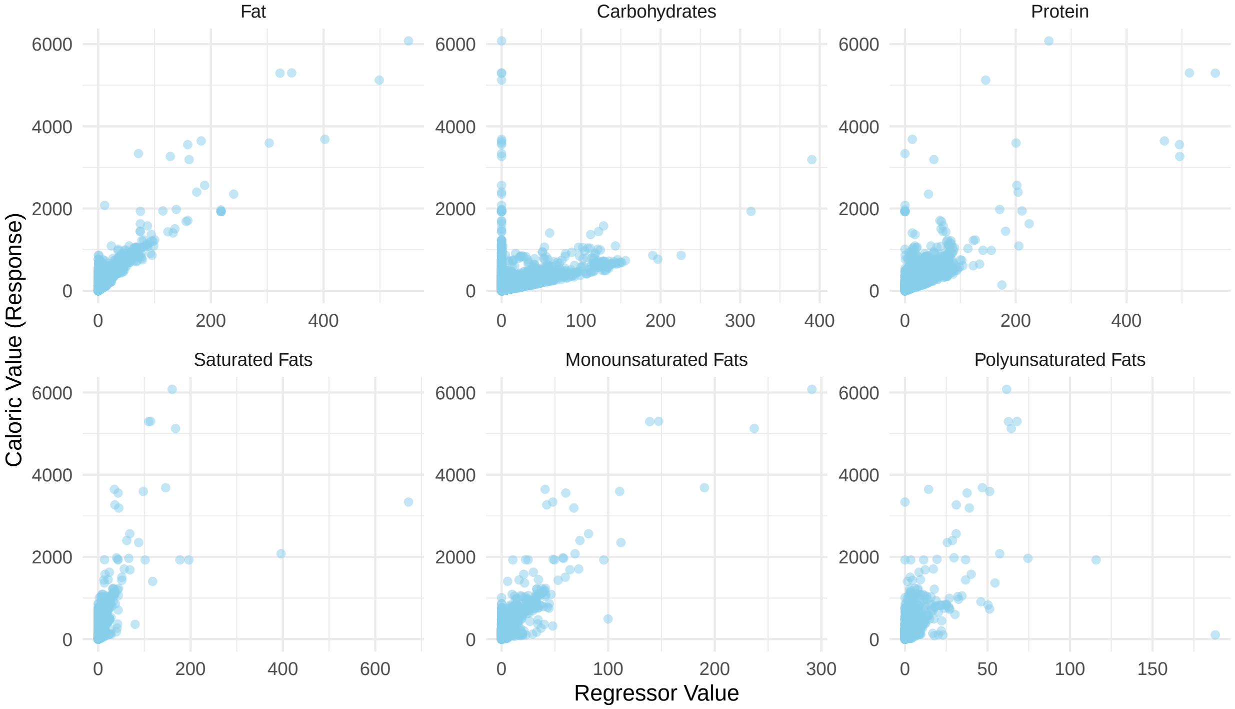

Figure 4: Caloric Value vs. Regressors Plot

4 Data Analysis and Results

4.1 Bayesian Linear Regression

For our regression model, we used $Y = X\beta + \epsilon$ where $Y \in \mathbb{R}$, X is $n \times p$, $\beta \in \mathbb{R}^p$, and $\epsilon \sim N_n(0, \Sigma)$. In particular, Y is our response variable (caloric value data), X is the design matrix of independent observations of the predictors (nutrients data), β are the regression coefficients which corresponds to our predictors, and ϵ is the random error. We will assume the errors are independent n is the number of rows in our data (2,395) and p is the number of predictors use (at most 32).

We decided to use a multivariate normal distribution for the prior on β and an inverse-gamma distribution on σ_2.

4.2 Bayesian Variable Selection

Not all the predictors we have in our dataset are necessarily useful for predicting our response variable, caloric value. There are various Bayesian variable selection methods to help with this issue. For this analysis, we chose to use Bayesian lasso and spike-and-slab.

Bayesian lasso is a regularization method used in regression to decrease variance at the cost of increasing bias. Instead of using an estimation of ordinary least squares, it uses L_1-constrained least squares by adding an L_1 penalty. As opposed to some other regularization methods, Bayesian lasso can reduce irrelevant coefficients to zero by penalty.

$$\min_{\beta}\left(\frac{1}{n}||Y - X\beta||_2^2 + \lambda||\beta||_1\right)$$

Spike-and-slab regression uses a mixture prior that sorts coefficients into a "spike" distribution (mass around zero), or a wide, diffused "slab" distribution. This prior is conditional on $\gamma_i \sim \mathit{Bernoulli}$ which represents whether β_i is included.

$$\beta_i \,|\, \gamma_i \sim (1 - \gamma_i)\,\mathcal{N}(0,\, \tau_o^{-1}) + \gamma_i\,\mathcal{N}(0,\, \tau_1^{-1}),$$

This results in a posterior that can give the probability of each β_i given the data. From there we can use MCMC methods to get the posterior inclusion probability (PIP) of γ_i to decide whether to include β_i in the model. This PIP is representative of the proportion of samples in MCMC that resulted in a model with β_i. Our threshold for including β_i was 0.5.

5 Summary and Discussion

6 Appendix

7 References

8 Self-Reflection (Individual)