Why is our food so caloric?

Bayesian Linear Regression of Food Nutrition

Thy Luong, Evgeniia Selezneva, Jasmine Wang

Introduction

Statement of the Problem

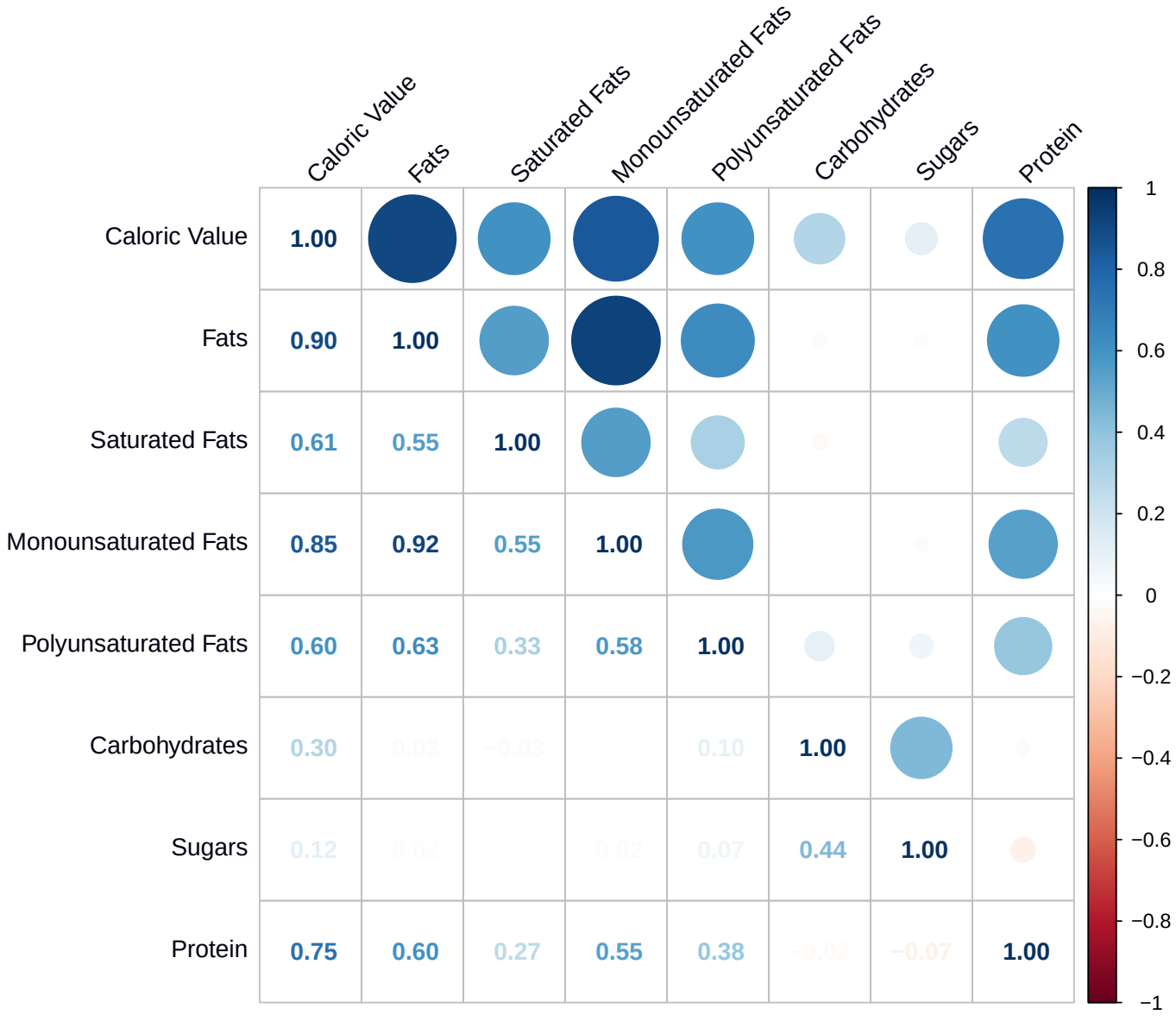
- How different nutrients affect the caloric value of food?
- What nutrients have and don't have caloric value?

Food Nutrition Dataset

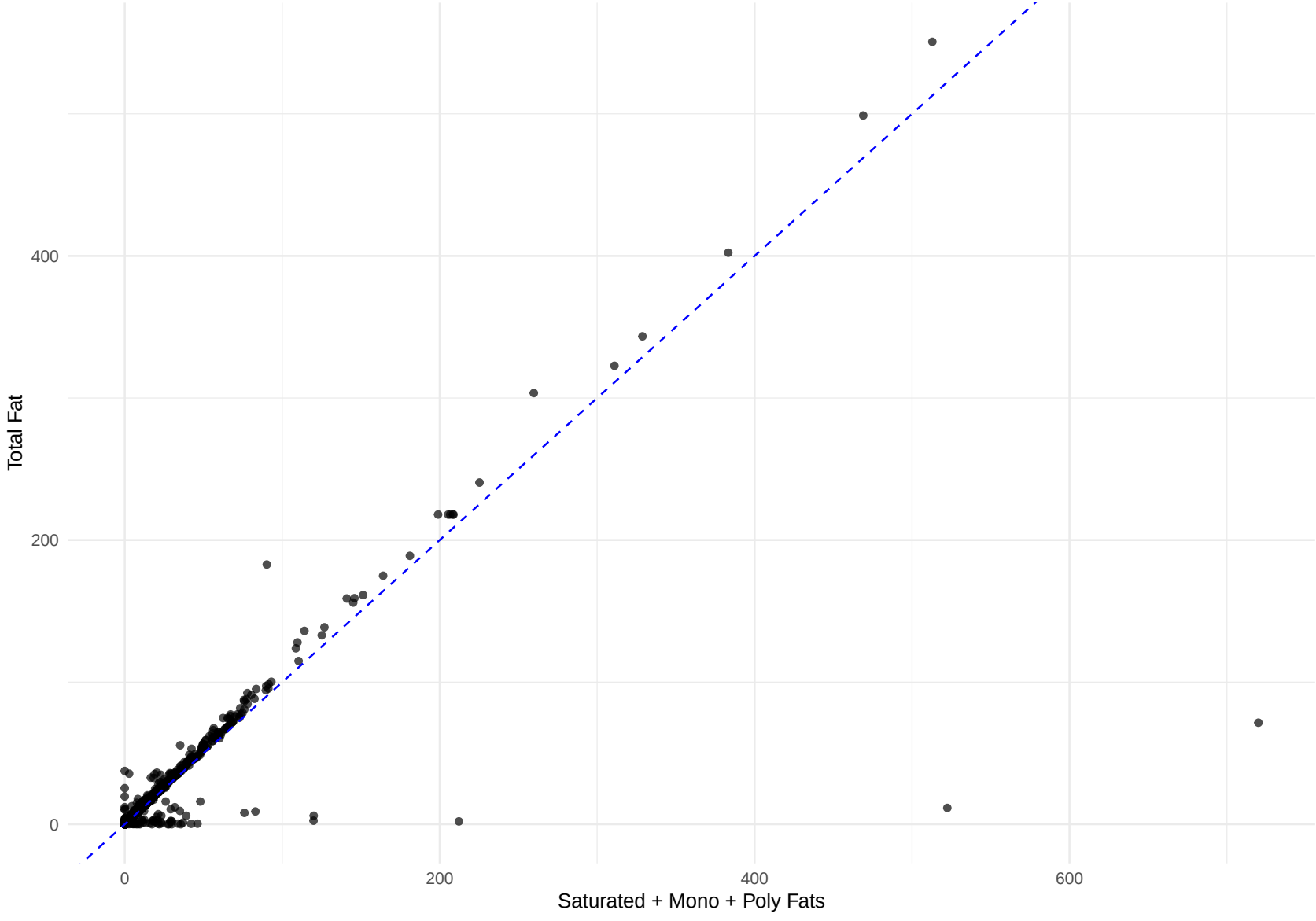
- 2,395 rows (one row - one product)
- 5 groups of products
- 35 columns
- 32 different nutrients
- food name, caloric value and nutrition density

Source: Kaggle.com

EDA: Correlation Analysis



EDA: Fat Multicollinearity Check



Bayesian Analysis

Bayesian Lasso Regression: Theory

$$\beta_j \mid \tau_j \sim N(0,\ \sigma^2\tau_j), \qquad \tau_j \sim \text{Exp}(\lambda^2/2)$$

Each coefficient has a **local shrinkage parameter** $\tau_j \rightarrow$ implements the Laplace prior through a Normal–Exponential mixture.

- **Small** $\tau_j \rightarrow \sigma^2\tau_j$ is tiny $\rightarrow \beta_j$ is strongly shrunk toward 0 $\rightarrow$ indicates an unimportant variable
- **Large** $\tau_j \rightarrow$ larger variance $\rightarrow \beta_j$ is less shrunk $\rightarrow$ indicates an important variable

Variable selection: Use the posterior of β_j; a common rule is to check whether its **credible interval excludes 0**.

Bayesian Lasso Regression: Selected Variables

All Regressors:

- Fat, Saturated Fats, Monounsaturated Fats, Polyunsaturated Fats, *Carbohydrates*, *Protein*, Water, *Vitamin A*, Magnesium, Phosphorus

All Regressors without Fat:

- Saturated Fats, Monounsaturated Fats, Polyunsaturated Fats, *Carbohydrates*, *Protein*, Cholesterol, *Vitamin A*, Vitamin E, Magnesium, Phosphorus

All Regressors without Fat Types:

- Fat, *Carbohydrates*, *Protein*, *Vitamin A*

Spike-and-Slab Prior: Theory

- For each coefficient β_j, introduce an **inclusion indicator** $\gamma_j \in \{0, 1\}$:

$$\gamma_j \sim \text{Bernoulli}(\pi)$$

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j)\, \delta_0 \; + \; \gamma_j\, N(0, \tau^2)$$

- **Spike** (δ_0): strong prior mass at $0 \Rightarrow$ variable excluded
- **Slab** $(N(0, \tau^2))$: diffuse prior $\Rightarrow$ variable can be nonzero
- **Joint posterior**: $p(\beta, \gamma \mid y, X) \propto p(y \mid \beta)\, p(\beta \mid \gamma)\, p(\gamma)$
- **Posterior inclusion probability (PIP)** for variable j: $\text{PIP}_j = P(\gamma_j = 1 \mid y, X) = \sum_{\gamma_{-j}} \int p(\beta, \gamma \mid y, X)\, d\beta$
- **Selection rule (threshold = 0.5)**: select variable j if $\text{PIP}_j > 0.5$.

Spike-and-Slab Prior: Selected Variables

All Regressors:

- Magnesium, *Vitamin A*, *Protein*, *Carbohydrates*, Polyunsaturated Fats, Monounsaturated Fats, Saturated Fats, Fat, Phosphorus, Water, Vitamin B6, Manganese, Calcium

All Regressors without Fat:

- Phosphorus, Vitamin E, *Vitamin A*, *Protein*, *Carbohydrates*, Polyunsaturated Fats, Monounsaturated Fats, Saturated Fats, Magnesium, Vitamin B6, Cholesterol

All Regressors without Fat Types:

- Fat, *Carbohydrates*, *Protein*, *Vitamin A*

Bayesian Linear Regression: Theory

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0,\, \sigma^2 I)$$

$$\beta \,|\, \sigma^2 \sim \mathcal{N}_p(\xi,\, \sigma^2 \Omega), \quad \sigma^2 \sim \text{Inverse-Gamma}(a, b)$$

$$\xi = 0, \quad \Omega = 10^3 I_p, \quad a = 0.01, \quad b = 0.01$$

Choose regressors, for which 95% credible interval of β does not include 0.

Bayesian Linear Regression: Full Model

Regressor	Mean	2.5% quantile	97.5% quantile
Fat	171.47	143.50	199.17
Protein	120.25	88.05	152.96
Carbohydrates	116.22	103.17	129.01
Saturated Fats	77.86	66.46	89.21
Monounsaturated Fats	41.84	17.01	66.39
Vitamin A	12.23	2.42	22.15

Same list of regressors is chosen by the Lasso All Regressors model

Bayesian Linear Regression: Lasso Without Fat

Regressor	Mean	2.5% quantile	97.5% quantile
Monounsaturated Fats	168.68	154.46	182.96
Protein	144.11	123.95	164.49
Carbohydrates	119.36	108.93	129.71
Saturated Fats	88.00	77.11	99.17
Polyunsaturated Fats	27.95	16.31	39.65
Vitamin A	11.63	2.21	20.90
Vitamin E	12.45	2.65	22.29

Bayesian Linear Regression: Lasso/Spike-and-Slab Without Fat Types

Regressor	Mean	2.5% quantile	97.5% quantile
Fat	265.52	254.02	277.24
Protein	131.08	119.23	143.08
Carbohydrates	110.11	100.81	119.62
Vitamin A	11.40	1.82	21.04

Bayesian Linear Regression: Spike-and-Slab All Regressors

Regressor	Mean	2.5% quantile	97.5% quantile
Fat	170.67	144.74	196.32
Protein	125.36	99.90	150.64
Carbohydrates	115.79	105.62	126.15
Saturated Fats	77.73	66.74	88.52
Monounsaturated Fats	42.67	18.73	66.50
Vitamin A	11.77	2.98	20.84

Bayesian Linear Regression: Spike-and-Slab Without Fat

Regressor	Mean	2.5% quantile	97.5% quantile
Monounsaturated Fats	168.39	153.91	182.71
Protein	132.80	107.06	158.31
Carbohydrates	118.82	108.33	129.29
Saturated Fats	88.12	77.14	99.23
Phosphorus	28.38	2.82	53.61
Polyunsaturated Fats	27.89	15.98	39.75
Vitamin E	12.39	2.62	21.97
Vitamin A	11.50	2.32	20.87

Model Comparison

Model	R2	MSE	N Predictors
1: Full	0.69322	50789	32
2: BLasso	0.69257	50896	12
3: BLasso (no Fat)	0.65694	56795	9
4: BLasso (no Fat Types)	0.69577	50366	4
5: Spike&Slab	0.69257	50896	12
6: Spike&Slab (no Fat)	0.66008	56276	10
7: Spike&Slab (no Fat Types)	0.69575	50370	4
8: Fat+Carbs+Protein only	0.69718	50133	3

Train-test split: 80%-20%

Conclusions

- The simple model with fats, proteins and carbohydrates appeared to have the best regression metrics
- Multicollinearity might be the reason for poorer performance of bigger models
- Even so, models chosen by variable selection techniques provide additional information on other nutrients
- For example, notice that all models name Vitamin A as an important predictor

Thank you!