HW2 - Cereal Cost Prediction

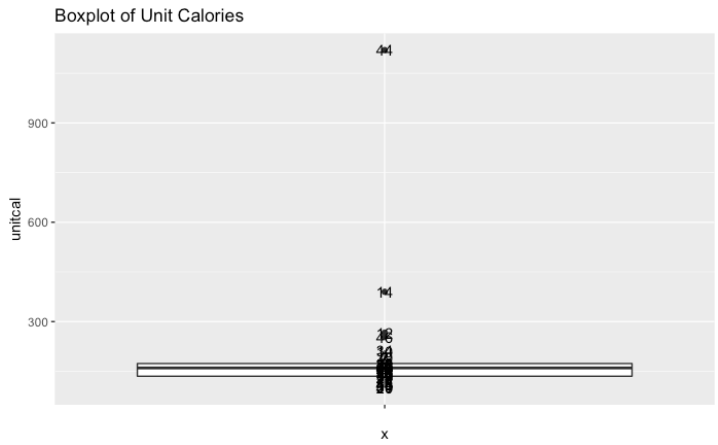
Data Summary

The dataset includes 46 observations. Variables in the data set are the cost of the box of cereal in dollars, name of the cereal, the weight (ounces) of the cereal, the calories per serving, the serving size (cups), the manufacturer's name and a coding of name into the integers 1 to 8 and finally whether the cereal box has a cartoon character on the front (yes = 1, no = 0).

Since serving size changes for each cereal, calories per serving doesn't make much sense. Therefore, the variable calories per serving need the transformation. The new variable is calories per unit (=calories per serving/serving size).

The distribution table has been shown as table 1. In addition, according to the boxplot (Figure 1), there is an outlier which is number 44 (Qi'a Chia, Buckwheat & Hemp Cranberry Vanilla). This outlier was removed in further analysis.

Sample Characteristics Observations(n=46)	
Cost, dollar, mean ± sd	4.48 (±0.97)
Weight, ounces, mean ± sd	16.56(±4.50)
Calories per serving, mean ± sd	136.70(±34.96)
Serving size, cups, mean ± sd	0.85(±0.19)
Manufacturer Code	
1, frequency	1
2, frequency	2
3, frequency	17
4, frequency	2
5, frequency	14
6, frequency	2
7, frequency	2
8, frequency	6
With cartoon character	
Yes, frequency	21
No, frequency	25



The goal is to predict cost according to caloric content, adult/kid target audience and manufacturer.

Random Effects Model

Since we need to consider the correlation between price from the same brand, the random effects model is used to analyze and 8 random effect variables were introduced in the model.

The model is used in this homework is

$$y_{ij} = x_i^t \cdot \alpha + \beta_j + \varepsilon_{ij}$$

where i runs from 1 to ***45***, j runs from 1 to 8, α is overall parameter (fixed effect), β_j is manufacturer level random effects, ε_{ij} is individual random effects.

Priors with t distribution

```
mu[i] <- inprod(x[i, ], alpha[]) + beta[comp[i]]
y[i] ~ dt(mu[i],tau.b)
df1 <- 1/invdf1
invdf1 ~ dunif(0,.5)
df2 <- 1/invdf2
invdf2 ~ dunif(0,.5)
```

Priors

We do not have any reference so I choose to use non-informative prior for α. Therefore, a normal distribution with mean of 0 and a low precision of 0.001 was chosen for α. For the random effect $\beta_j|\tau^2{}_1$, the priors were also chosen as normal distribution with mean of 0. For $\tau^2{}_1$, which is precision of β_j. I choose gamma distribution with moderate mean of $2 = \frac{5}{2.5}$ and standard deviation of $0.8 = \frac{5}{2.5^2}$, which could be in a narrow range. For $\tau^2{}_2$, I choose gamma distribution with mean of $1.07 = \frac{2}{1.86}$, which is the $1/\mathrm{var}(\mathrm{price})$, and standard deviation of $0.57 = \frac{2}{1.86^2}$.

Make three priors with different precision

- Prior A:

$$\alpha \sim \mathrm{N}(0, 0.001)$$

$$\beta_j|\tau^2{}_1 \sim N(0, \tau^2{}_1);\ \ \tau^2{}_1 \sim Gamma(5, 2.5)$$

$$\varepsilon_{i,j}|\tau^2{}_2 \sim N(0, \tau^2{}_2);\ \ \tau^2{}_2 \sim Gamma(2, 1.86)$$

- Prior B:

$$\alpha \sim \mathrm{N}(0, 0.1)$$

$$\beta_j|\tau^2{}_1 \sim N(0, \tau^2{}_1);\ \ \tau^2{}_1 \sim Gamma(5, 2.5)$$

$$\varepsilon_{i,j}|\tau^2{}_2 \sim N(0, \tau^2{}_2);\ \ \tau^2{}_2 \sim Gamma(2, 1.86)$$

- Prior C:

$$\alpha \sim \mathrm{N}(0, 0.0001)$$

$$\beta_j|\tau^2{}_1 \sim N(0, \tau^2{}_1);\ \ \tau^2{}_1 \sim Gamma(5, 2.5)$$

$$\varepsilon_{i,j}|\tau^2{}_2 \sim N(0, \tau^2{}_2);\ \ \tau^2{}_2 \sim Gamma(2, 1.86)$$

Result

The table 1.a showed that posterior of manufacturer level random effects. According to the result, cereal from Nature's Path has highest price and cereal from Quaker has lowest price. The Figure 2 showed the boxplots of the estimate betas, there is slight difference between brands. The table 1.b showed that posterior of fixed effect, indicated that weight had positive association with price but calories per serving and serving size had negative association. Cereal with adult target audience had higher price than cereal with kids target audience. In addition, trace plots showed as Figure 2.a and Figure 2.b. All trace plots showed all parameters converge well.

The DIC result showed as Table 3 overall predictive power of the model. Lower DIC indicated better fit. Therefore, t-distribution will be the better model.

Prior Distribution	Normal	T
DIC	128.7	131.5

Table 3. DIC values

Appendix

Sample Characteristics Observations(n=46)	
Cost, dollar, mean ± sd	4.48 (±0.97)
Weight, ounces, mean ± sd	16.56(±4.50)
Calories per serving, mean ± sd	136.70(±34.96)
Serving size, cups, mean ± sd	0.85(±0.19)
Manufacturer Code	
1, frequency	1
2, frequency	2
3, frequency	17
4, frequency	2
5, frequency	14
6, frequency	2
7, frequency	2
8, frequency	6
With cartoon character	
Yes, frequency	21
No, frequency	25

	Parameter	Mean	S.D.
Beta1	365 everyday value	0.1251	0.5757
Beta2	Cascadian Farm	0.2916	0.5164
Beta3	General Mills	-0.22	0.3736
Beta4	Kashi	-0.3	0.5016
Beta5	Kelloggs	0.3058	0.385
Beta6	Nature's Path	0.4793	0.5424
Beta7	Post	-0.332	0.5059
Beta8	Quaker	-0.3527	0.066

Table 2a. Posterior of Betas (Random Effects)

	Parameter	Mean	S.D.
Alpha1	intercept	5.4675	1.1418
Alpha2	weight	0.0665	0.0343
Alpha3	unitcalories	-0.0035	0.0046
Alpha4	size	-1.5767	0.8797
Alpha5	cartoon	-0.5437	0.3014

Table 2b. Posterior of Alphas (Fixed Effects)

Figure 1. Boxplot of Unit Calories

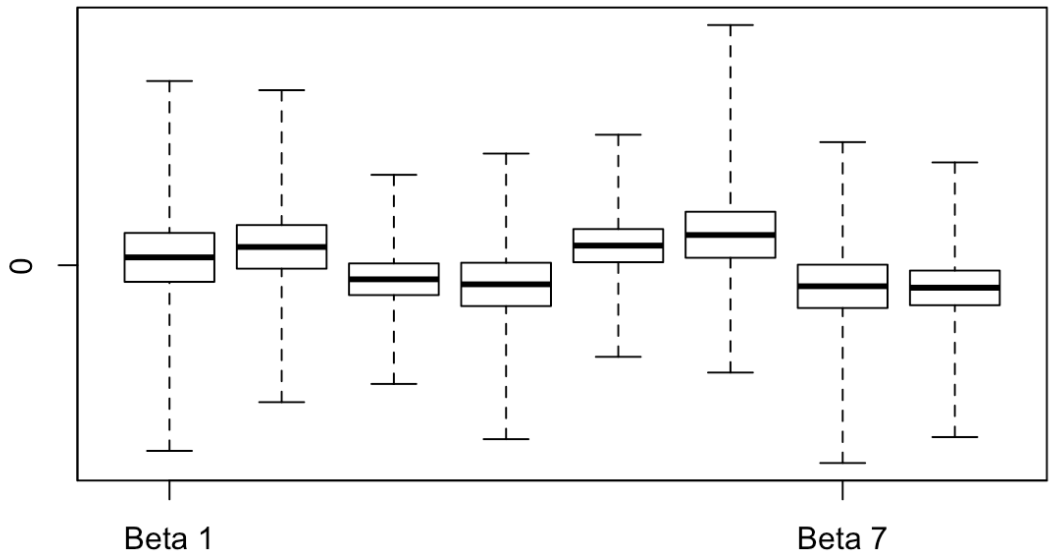


Figure 2. Boxplot of Betas

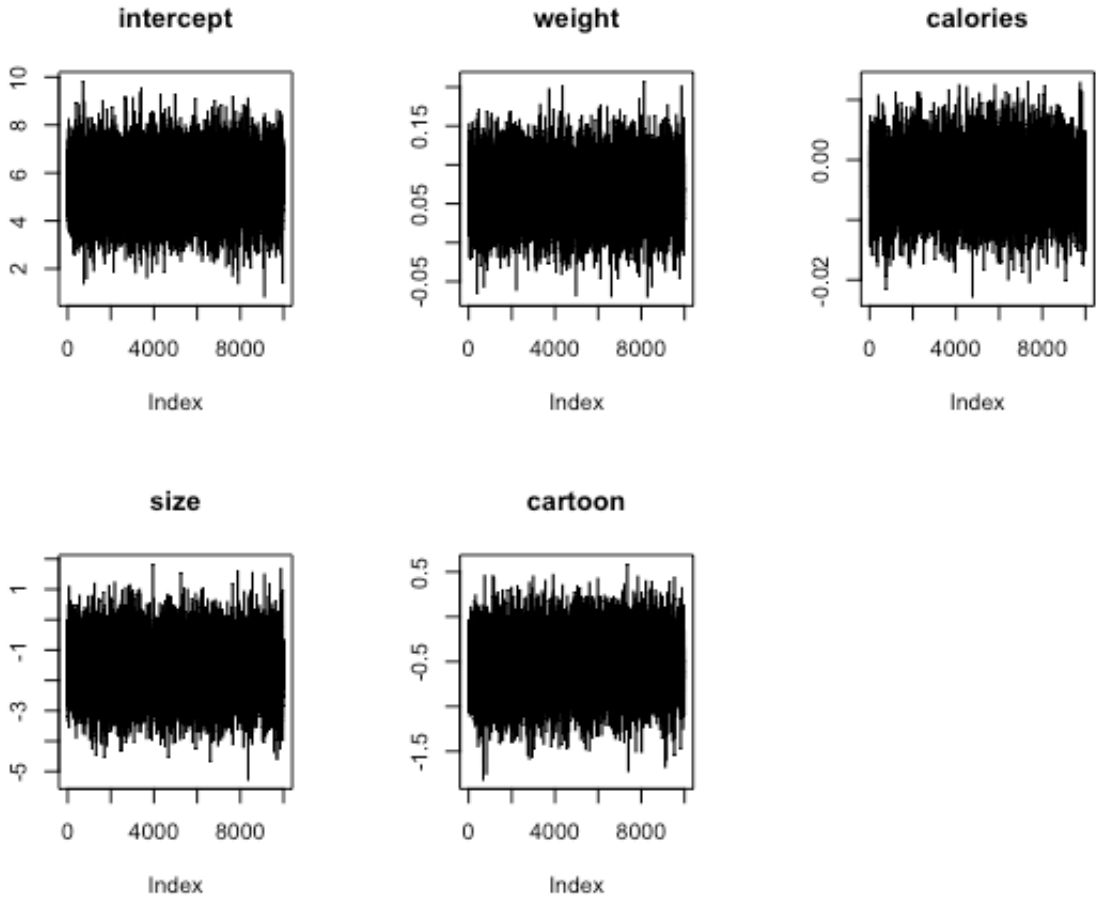


Figure 2a. Trace plot of alphas

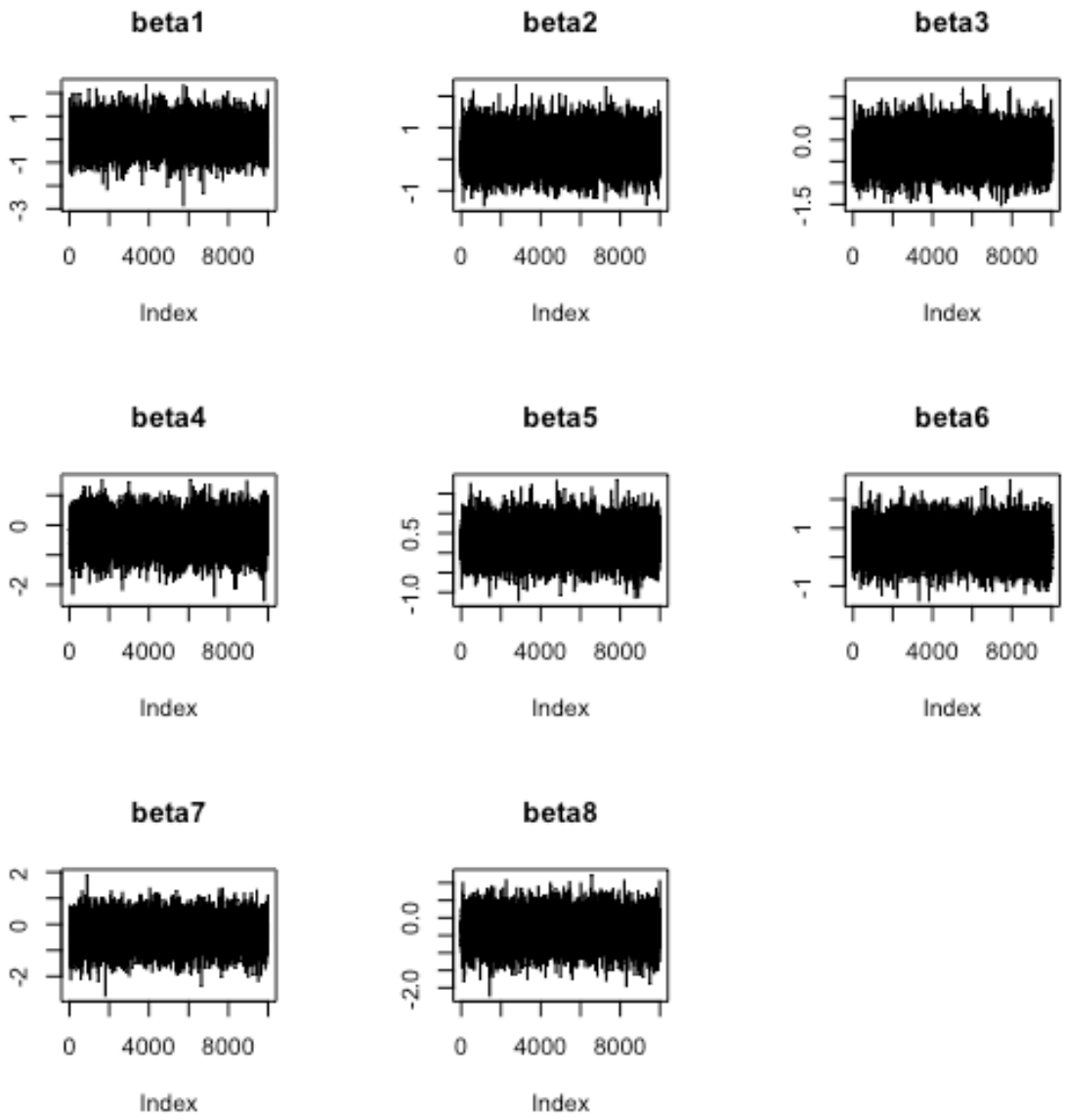


Figure 2b. Trace plot of betas