Title*

Subtitle

Chenming Zhao

November 24, 2024

abstract

1 Introduction

2 Data

2.1 Overview

The product pricing data for the "No Name" brand sold by Loblaws, used in this analysis, is included in the dataset provided by (Filipp 2024). This dataset compiles detailed product information, including vendors, product names, brands, categories, and unit prices. Additionally, it includes time-series pricing records for each product, capturing timestamps of price changes, current prices, unit prices, and quantities sold. These records are periodically updated on the Hammer platform to reflect the latest pricing trends. This dataset forms the foundation for analyzing price trends, identifying unit variations, and categorizing products into specific types for further analysis. Table 1 previews the information for the target products included in this analysis, starting from June 2024.

To simulate, test, analyze, and clean the polling data, the statistical programming language R was employed (R Core Team 2023). Key libraries that supported the data analysis include `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023), `arrow` (Richardson et al. 2024), `knitr` (Xie 2023), `ggplot2` (Wickham 2016), `kableExtra` (Zhu 2024) and `patchwork` (Pedersen 2024). Additionally, the `rstanarm` (Goodrich et al. 2022) library was utilized to implement multilevel modeling for prediction purposes. Also this analysis inherits the folder structure from Professor Rohan Alexander from University of Toronto (Alexander 2024).

*Code and data are available at: https://github.com/ChenmingZ2000/Loblaws-NoNameBrand-Grocery-Price-Trend-Analysis.git. We gratefully acknowledge Jacob Filipp's groceries dataset used in this analysis. The data is available at: https://jacobfilipp.com/hammer/

2.2 Measurement

In the real world, pricing data for consumer goods is a dynamic reflection of market trends, regional preferences, and competitive strategies. The prices of products fluctuate based on supply chain disruptions, seasonal demand, and retail strategies, making it essential to track and analyze this information systematically. The Hammer dataset serves as a resource in understanding these trends, offering a detailed collection of product pricing data sourced from leading retailers like Loblaws, Walmart, and Metro. This dataset is embedded with complexities arising from the data collection process, competitive pricing strategies, and evolving consumer behavior.

Although the data collection process for (Filipp 2024) data is not explicitly documented, based on the emphasis on "related" content on the Project Hammer website and standard industry methods, it is highly likely that common techniques such as web scraping, API integration, and data standardization were employed. These are typical approaches used to construct structured datasets from online retail platforms. That is, retailers provide public access to their product inventories and prices through their websites, making them prime candidates for automated data extraction. For example, web scraping tools are deployed to extract product details such as names, brands, vendors, categories, unit prices, and quantities, while timestamps capture the exact moment of data retrieval to ensure time-sensitive accuracy. However, even with automated methods, the process is not without its challenges. Retailers often introduce barriers like CAPTCHA systems or dynamic content loading, which require robust solutions such as headless browsers or API integrations to ensure smooth data acquisition.

Sampling and representativeness are considerations in Hammer's data collection process. Since the dataset aims to reflect national pricing trends, it includes a diverse range of vendors and product types to capture regional and competitive variations. However, inherent biases may still exist. For instance, not all vendors update their online product catalogs in real time, which may introduce lag in capturing price changes. Additionally, regional pricing disparities might not always align with the dataset's collection schedule, potentially skewing the data toward more actively monitored regions or vendors.

Once the raw data is collected, it undergoes cleaning and standardization. Product names, for instance, often include variations due to marketing strategies (e.g., "Club Size" or "Family Pack"), which are standardized to ensure consistency across entries. Similarly, data with missing values—such as prices or unit information—are flagged for further investigation.

Finally, the cleaned and standardized data is stored in a structured database, organized by product ID, timestamps, and categorical variables such as vendor, brand, and product category. This structured approach allows researchers to track pricing trends over time, analyze variations across regions or vendors, and categorize products into specific types. Hammer's database not only facilitates trend analysis but also enables broader comparisons and aggregations, allowing researchers to examine shifts in consumer behavior or retailer strategies.

2.3 Analysis Data

This analysis focuses on Loblaws' "No Name" brand products, with the goal of understanding pricing trends. We only selected observations where the vendor is Loblaws and the brand is "No Name" because Loblaws announced a price freeze on more than 1,500 "No Name" products until January 31, 2023, to address the pressures of rising food prices. However, with the price freeze period now over in 2024, we are particularly interested in analyzing the price changes of "No Name" products (see more details in Loblaw Companies Limited 2022). At the same time, we prioritized high-quality data by filtering out records with missing or incomplete price information, as we have plenty of price records for a product in one day and unreliable data could introduce significant bias and reduce the accuracy of our trend analysis. In researching price trends and product variations, we selected specific variables that are impactful for analyzing pricing behavior affected by product categorization. Below is an overview of the data used for this analysis Table 1, along with the selected outcome and predictor variables.

2.3.1 Outcome Variables

In this analysis of Loblaws' "No Name" brand products 2024, the primary outcome variable is the `current_price`, which represents the retail price of a product at the time of the record. Understanding the pricing trends of these products allows us to analyze consumer affordability, retailer pricing strategies, and competitive positioning in the market. By focusing on `current_price`, we aim to identify influences that influence price fluctuations over time and assess how product size, unit and type variations contribute to these trends. This metric serves as the foundation for evaluating the dynamics of product pricing and finding differences across product categories.

2.3.2 Predictor Variables

In this analysis, the predictor variables are the variables that we believe can influence the current price of products in the Loblaws grocery dataset. These variables are chosen based on their potential to explain variations in product pricing. Below are the predictor variables included in our model, along with an explanation of their significance.

Size: Product size is a determinant of pricing, as larger sizes generally cost more due to the increased quantity. However, pricing strategies such as bulk discounts or price-per-unit variations can create non-linear relationships between size and price, making size an important factor to analyze.

Type: Product type (e.g., snacks, seasoning, dairy) reflects category-level differences in pricing strategies. Different types of products may have varying production costs, supply chain logistics, and market demand, all of which contribute to price fluctuations.

Date: Time plays an important role in grocery pricing due to reasons like seasonal demand, inflation, supply chain disruptions, and promotions. Analyzing the impact of time allows us to capture temporal trends and shifts in pricing behavior, such as rising costs during inflationary periods or discounts during specific seasons.

2.3.3 Exclusions and Considerations

Other variables (e.g., vendor or brand-specific promotions) could also influence pricing. However, in this analysis, we focus specifically on the effects of `size`, `type`, and `date`, as they represent the most consistent and accessible predictors in the dataset. Additional variables, while potentially relevant, may not be consistently available or may introduce unnecessary complexity into the model. For example, promotions or regional pricing differences were not included in this analysis due to large amounts of data unavailability. Additionally, while `old_price` might influence pricing strategies, this dataset does not provide enough details. By focusing on the variables above, we aim to construct a reliable and interpretable model for understanding price dynamics within Loblaws' "No Name" brand products.

Table 1: The First Few Lines of Cleaned Grocery Price Analysis Data

product_id	date	current_price	product_name	type	size
1295619	2024-06-11	5.49	Blueberry Waffles	Snacks	Large
1295620	2024-06-11	2.50	Blueberry Waffles	Snacks	Small
1757033	2024-06-11	3.99	Microwave Butter Flavour Popcorn	Snacks	Large
1870035	2024-06-11	5.99	100% Pure Canola Oil	Cooking Oil	Large
1870037	2024-06-11	3.99	100% Pure Canola Oil	Cooking Oil	Small
1870038	2024-06-11	8.49	100% Pure Canola Oil	Cooking Oil	Extra Large
1870043	2024-06-11	17.99	100% Pure Corn Oil	Cooking Oil	Extra Large

2.3.4 Relations Between Variables

The following figures, Figure 1 and Figure 2, present various aspects of data related to Loblaws' "No Name" grocery prices, helping us understand how item prices change with different variables. In the left panel of Figure 1 and both panels in Figure 2, the x-axis is marked with dates categorized by month. The chart analyzes the relationships between product prices, size, type, and date, showing how these variables may interact and influence pricing trends. Product size and type have an impact on pricing trends, and we anticipate that incorporating these variables into the predictive model will improve the accuracy of price predictions under different product and market conditions.

While we observe certain significant fluctuations in the right panel of Figure 2, which may indicate abnormal prices being recorded or the presence of products with extreme prices in specific categories during certain periods, we cannot directly remove this data as it still contains information we need. In other words, the prices of these products can also explain some of the reasons behind price fluctuations. Completely removing these observations would weaken

the model's ability to reflect price trends accurately. Therefore, we choose to retain these observations and mitigate the biases they introduce by combining other variables to build a more robust predictive model.

The boxplot on the right part of Figure 1, shows the pricing differences across product sizes. Compared to "Small" and "Large" products, "Extra Large" products have a higher median price and a wider interquartile range. This reflects inherent pricing strategies where larger quantities are priced higher, but not always linearly (e.g., discounts for bulk purchases). However, the presence of outliers, especially in the "Extra Large" and "Large" categories, indicates that even within the same size group, different product types or brands can significantly impact final pricing. The narrow boxplot for "Small" products suggests a smaller range of price variation, which aligns with the lack of price fluctuations observed in the left panel of Figure 1 for small-sized products.

Figure 2 shows how average prices vary over time, segmented by product `type` (left) and `size` (right). In the left panel, we observe distinct pricing patterns across product types. For instance, "Seasoning" demonstrates higher volatility compared to "Flour/Rice," likely due to differences in demand elasticity. The "Milk" category shows occasional price dips, possibly influenced by the introduction of new products or promotional events.

The right panel shows price trends segmented by size. "Extra Large" products generally have higher average prices but exhibit more unstable trends over time. In contrast, "Small" products show minimal price variation, possibly due to their inherently lower pricing or because their unit price is relatively higher compared to larger sizes.

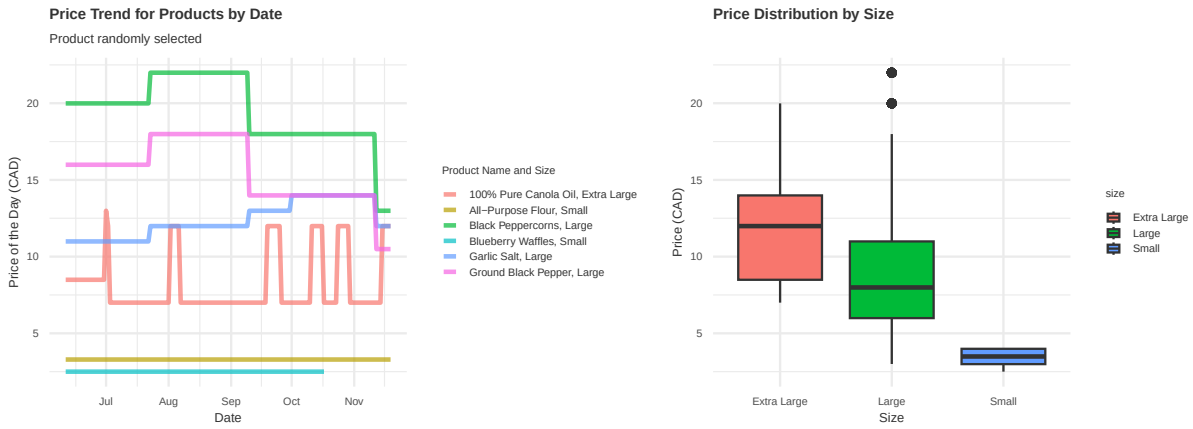

Figure 1: Price Trend for Six of the Loblaws "No Name" Groceries Over Time (left) and Price Distribution by Size (right).

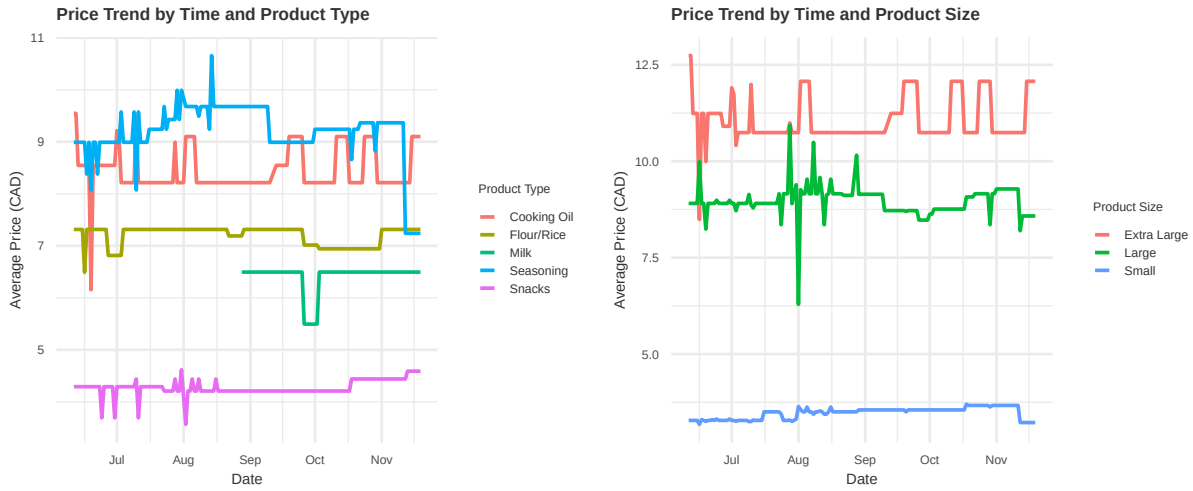

Figure 2: Average Support Rate for Donald Trump from 2024-07-21 to 2024-10-23 reported by different pollsters

3 Model

Our model references the analytical approach proposed by Chen, Aravkin, and Martin (2018), adopting a hierarchical Bayesian framework to model the price trends of Loblaws' "No Name" grocery products. Specifically, the model is based on a Gamma regression framework with a log-link function, designed to handle the positively skewed distribution of product prices. By using the Gamma distribution, we can capture the inherent variability in prices while adhering to the non-negative nature of price data. The Gamma distribution is particularly well-suited for modeling continuous, non-negative, and positively right-skewed data, which are common in economic contexts such as insurance claims and survival data (Chen, Aravkin, and Martin 2018).

The goal of our modeling strategy is to build a pricing model aimed at predicting changes in `current_price` over time while considering key variables that influence price fluctuations. The structure of the model combines fixed effects and random effects and includes multiple variables, such as `product_id`, product `size`, and product `type`. This modeling approach is designed to balance the complexity of the data while ensuring robust predictions. The hierarchical structure of the model allows us to capture variations at the level of individual products while identifying overall trends across different product sizes and types. It also enables us to forecast future price trends of Loblaws "No Name" groceries, observing whether the retailer's pricing behavior aligns with their commitments.

Here, we provide a brief overview of the Bayesian analytical model used to investigate the price trends of Loblaws "No Name" products. Background details, model assumptions, and diagnostics are included in Appendix A.

3.1 Model set-up

This model is a generalized linear mixed model (GLMM) within a Bayesian framework, specifically designed to handle positively skewed data:

$y_i \sim \text{Gamma}(\mu_i, \phi)$

Define y_i as the current price of the product in the i-th observation, assumed to follow a Gamma distribution with mean μ_i and shape parameter ϕ. This model accounts for variations across different product sizes, types, and temporal trends, with random intercepts for each product ID.

The linear predictor for the mean (μ_i) of the Gamma distribution is defined as follows:

$log(\mu_i) = \alpha + \beta_1 \cdot \text{size}_i + \beta_2 \cdot \text{type}_i + \beta_3 \cdot \text{date}_i + u_{\text{productID}[i]}$

where:

- y_i is the observed current price for the i-th product.
- μ_i is the predicted mean price for the i-th product, modeled on a log scale.
- $\alpha \sim \text{Normal}(2, 5)$ is the global intercept representing the baseline price.
- $\beta_1 \sim \text{Normal}(0, 5)$ captures the effect of product size on pricing (e.g., Small, Large, Extra Large).
- $\beta_2 \sim \text{Normal}(0, 5)$ represents the effect of product type (e.g., Snacks, Flour/Rice, Cooking Oil) on pricing.
- $\beta_3 \sim \text{Normal}(0, 5)$ accounts for the temporal trend in prices over time.

Random Effects: $u_{\text{productID}} \sim \text{Normal}(0, \sigma_{\text{productID}})$

where:

$u_{\text{productID}[i]}$ is the random intercept for each product ID, capturing product-specific variations in pricing.

3.2 Model justification

3.2.1 Model Prior Selection, Assumptions, and Validation

Our prior selection was informed by the research in gamma modeling Veazie et al. (2023), we aim to provide sufficient flexibility for the model to adapt to the data. For example, the prior for the intercept was set as $\text{Normal}(2, 5)$, reflecting the assumption that the baseline log-transformed price of Loblaws "No Name" products fluctuates around 2 while allowing for some uncertainty. For the effect coefficients of variables such as `size`, `type`, and `date`, we chose

priors of Normal$(0, 5)$, representing the expectation of moderate impacts of these variables on pricing. These priors enable the model to adjust based on observed data.

The model assumes that price variations at the product level are non-negative, right-skewed, and that the price differences associated with size and type are random and normally distributed. While these assumptions simplify computation and enhance interpretability, they may not fully capture the impact of market conditions, regional competition, or seasonal trends on pricing dynamics. For instance, the model assumes that the effect of size is consistent across all product types, but in reality, certain product types may be more significantly influenced by supply chain issues or promotional activities. Additionally, the model incorporates random effects for individual products to capture specific variations for each product, assuming these effects remain consistent over time.

- **TODO**

3.2.2 Expectation of relationships between variables

For the variable `type`, we anticipate that different product types will influence the estimation of current prices. While we cannot precisely determine which product categories incur higher costs due to supply chain or consumer demand, it is reasonable to assume that categories such as "Snacks" or "Flour/Rice" may exhibit distinct pricing dynamics. For instance, "Snacks" might experience greater price volatility due to internet trends or elastic demand, whereas "Flour/Rice" could reflect a more stable trend. These category-specific variables introduce variability into the overall pricing estimates.

For the variable `date`, we hypothesize that the prices of Loblaws "No Name" products fluctuate over time, reflecting seasonal trends, market dynamics, and promotional activities. This hypothesis is based on historical observations, where prices often vary due to promotions or stock-outs. Long-term trends, on the other hand, may reflect broader economic conditions or changes in production costs. The variable size captures the impact of product quantity on pricing. We expect larger-sized products, such as "Extra Large," to generally have higher prices due to increased quantities. However, economies of scale may result in lower unit prices for larger sizes, while smaller sizes, despite lower overall costs, may have higher unit prices. This pricing strategy is common in retail to ensure accessibility for different consumer groups while optimizing profit margins.

To account for variations among individual products, we introduced a random intercept `(1 | product_id)` for each product. This intercept captures product-specific price differences, such as brand effects, unique demand characteristics, or production costs.

By modeling the interaction between size and type `(size * type)`, the model allows for a more detailed understanding of relationships, such as pricing differences for "Extra Large" products between "Cooking Oil" and "Milk." This interaction term helps identify combined effects that cannot be fully explained by individual variables. Together, these variables enable

the model to capture the dynamic pricing structure of Loblaws "No Name" products, reflecting time trends and structural differences across product categories and sizes.

4 Results

Our results are summarized in Table **??**.

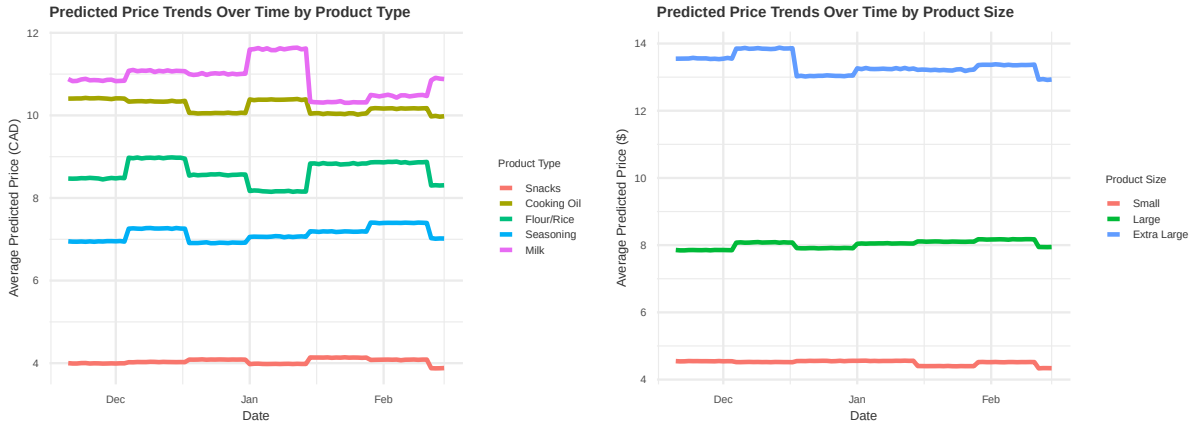

Figure 3: Prediction of average price for Loblaws' No Name Grocery by type and by size

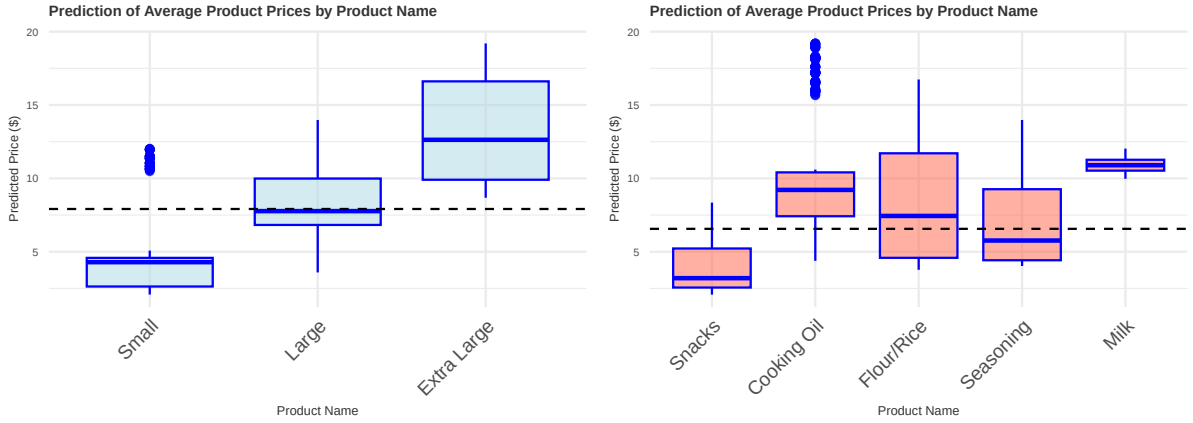

Figure 4: Prediction of average price for Loblaws' No Name Grocery by type and by size

Appendix

A Model details

R package `bayesplot` (Gabry J 2024) is used for posteriors.

Due to space limitations, we have chosen to display only the posterior and prior comparisons for a subset of variables in the posterior analysis. If we were to display all variables included in the model, the chart would become overly complex and difficult to interpret. Although the effects of multiple variables differ, focusing on this subset may give the impression that the model's sensitivity to variations across different factors is limited. Thus, we have selected representative variables, including the intercept, size categories ("sizeLarge" and "sizeSmall"), and product types ("typeSnacks," "typeFlour/Rice," and "typeSeasoning"). The output results for other variables not shown are generally consistent with the trends observed here. Additionally, in regression models (such as stan_glmer), categorical variables (factors) are typically automatically converted into dummy variables. One category is set as the reference category (baseline), and its effect is included in the intercept, while the coefficients for other categories represent their offsets relative to the baseline. For example, in our model, if `Extra Large` is set as the baseline (default), it will not appear as a separate coefficient. The coefficients for `sizeLarge` and `sizeSmall` would then represent their deviations relative to `Extra Large`.

A.1 Posterior predictive check

Figure 5 displays a comparison between the distribution of observed data y and the distribution of simulated data y_{rep} generated by our model. The dark solid line represents the observed data's actual distribution, while the multiple light lines correspond to distributions from various simulated datasets. If the distribution of simulated data y_{rep} closely resembles that of the observed data y, it suggests that the model effectively captures the underlying characteristics of the observed data.

In Figure 5, we observe that the simulated data distributions (light lines) align closely with the real data distribution (dark line) across the entire range, particularly at the peaks and troughs of the observed data. This indicates that the model has captured the distributional features of the observed data, including its skewness and multimodal structure. The ability to replicate these features demonstrates the model's strong predictive capacity and its reliability in estimating posterior distributions. This result builds confidence in the model's suitability for further inference and prediction.

Figure 6 displays the MCMC sampling trace plots for six key parameters in our model, including the `(Intercept)`, size variables `sizeLarge` and `sizeSmall`, and type variables `typeSnacks`, `typeFlour/Rice` and `typeSeasoning`. Each plot includes four chains from Chain 1 to Chain 4, representing independent MCMC sampling processes. In a well-converged model, our

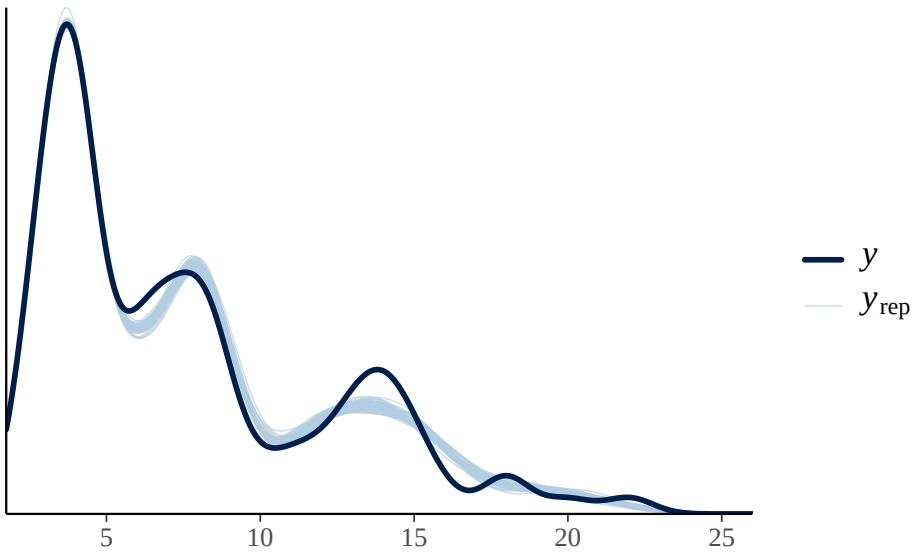

Figure 5: PP Check for Model

expectation is that each chain to stabilize and fluctuate around a central value without showing a clear trend (i.e., the mean of the samples converges over time). Additionally, the chains for a parameter should exhibit consistent behavior, indicating agreement among the chains.

From Figure 6, we observe that the sampling chains for all six parameters show a steady state, with fluctuations around a central value and no observable trends over the iterations. Moreover, the chains for each parameter are closely aligned, demonstrating consistency between the chains. This isuggests that the model has reached convergence and the posterior distributions for these parameters are stable. These trace plots support the reliability of our MCMC sampling process and provide confidence in the validity of the posterior estimates.

Figure 7 shows the $\hat{R}$ diagnostic statistic for our model, which is used to assess the convergence of MCMC sampling. Ideally, is that the $\hat{R}$ value would be close to value 1, which indicates consistent variance across different MCMC chains, signifying that the model has reached a stable state. The horizontal axis represents the $\hat{R}$ values, where values closer to 1 indicate better convergence. Our expectation is that, all parameters would have $\hat{R}$ values below 1.05, which means minimal variance differences between chains and consistency across them. This plot shows that all $\hat{R}$ values are concentrated around 1, and all $\hat{R}$ values are lower than the threshold of 1.05 (indicated by the vertical dashed line on the right), suggesting that the sampling chains have converged well, providing thw reliability for our model results.

Figure 8 shows the posterior distributions of certain parameters in our model, including `(Intercept)`, size variables `sizeLarge` and `sizeSmall`, and type variables `typeSnacks`, `typeFlour/Rice`, and `typeSeasoning`. The shape and location of these distributions help us understand the extent to which the observed data supports these parameter estimates and the degree of uncertainty associated with them.

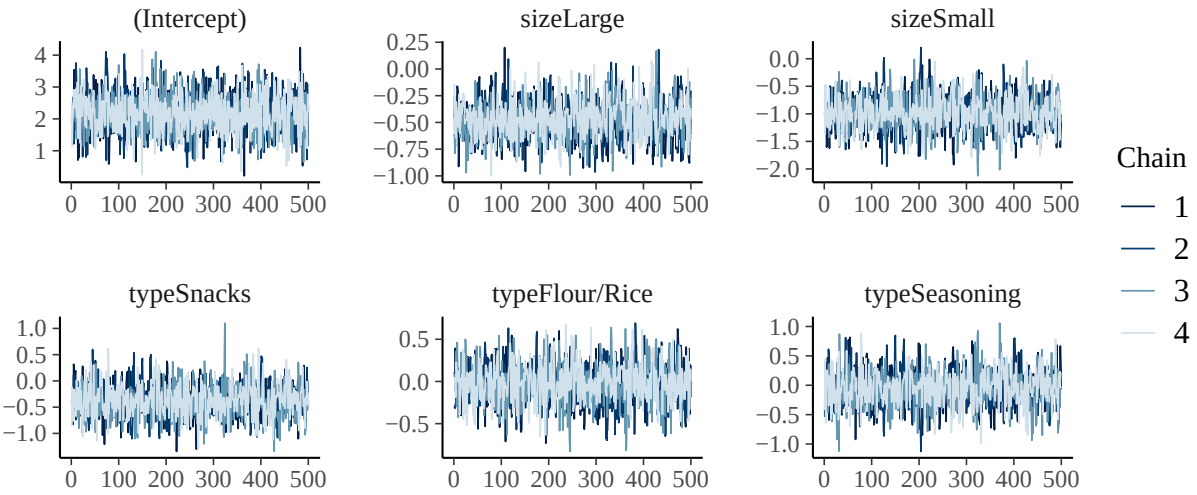

Figure 6: MCMC sampling trace plots for model

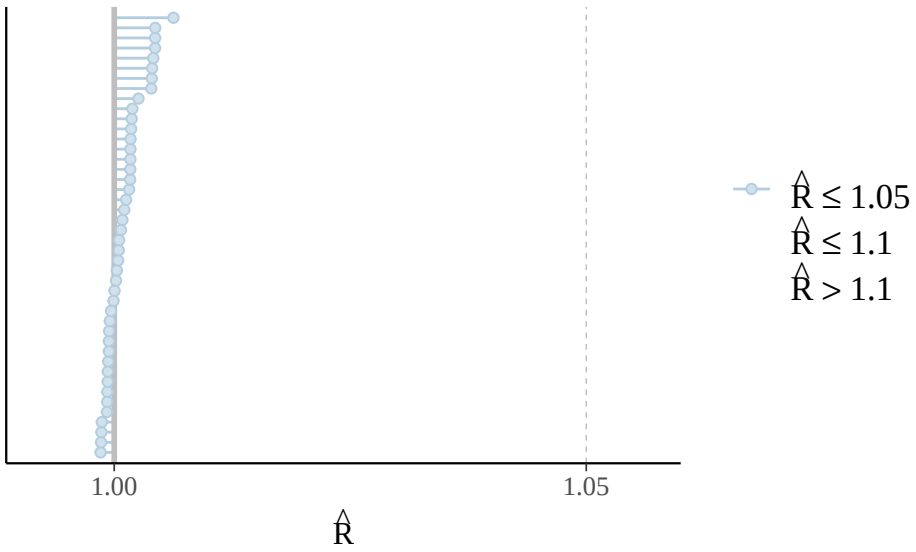

Figure 7: R hat Diagnostic statistic for model

The `(Intercept)` is centered nearly around 2.5, reflecting the baseline level for product prices in our model when all other factors are held constant. The narrow and peaked distribution suggests strong support for this estimate with relatively low uncertainty. For `sizeLarge` and `sizeSmall`, the distributions are shifted slightly to the left, indicating that these sizes are generally associated with lower prices compared to the baseline category `ExtraLarge`. Type variables `typeSnacks`, `typeFlour/Rice`, and `typeSeasoning` exhibit posterior distributions centered close to zero, implying that their influence on product pricing is minimal or uncertain, as the data does not strongly support a directional impact.

Overall, the results suggest that the model successfully captures the baseline pricing level and the relative influence of product sizes, but it struggles to discern clear effects for product types, possibly due to overlapping pricing patterns or insufficient variation within these categories.

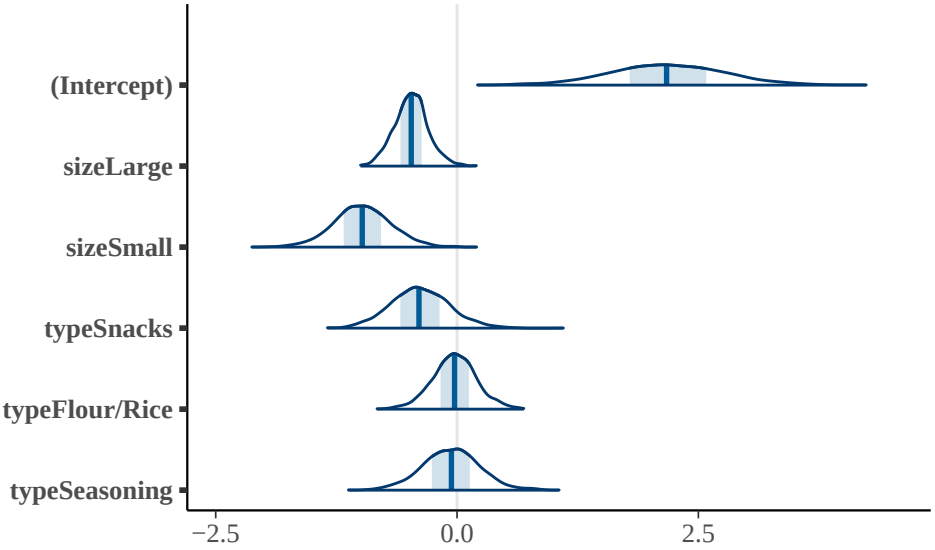

Figure 8: posterior distributions of four parameters in model

References

Alexander, Rohan. 2024. "Starter Folder." https://github.com/RohanAlexander/starter_folder.git.

Chen, Xin, Aleksandr Y. Aravkin, and R. Douglas Martin. 2018. "Generalized Linear Model for Gamma Distributed Variables via Elastic Net Regularization." https://arxiv.org/pdf/1804.07780.

Filipp, Jacob. 2024. "Hammer Canadian Grocery Price Data." https://jacobfilipp.com/hammer/.

Gabry J, Mahr T. 2024. *Bayesplot: Plotting for Bayesian Models.* https://mc-stan.org/bayesplot/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Loblaw Companies Limited. 2022. "Loblaw Hits the Brakes on Food Inflation by Freezing Prices on 1,500 No Name Products." https://www.loblaw.ca/en/loblaw-hits-the-brakes-on-food-inflation-by-freezing-prices-on-1500-no-name-products/.

Pedersen, Thomas Lin. 2024. *patchwork: The Composer of Plots*. https://patchwork.data-imaginist.com.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Veazie, Peter, Orna Intrator, Bruce Kinosian, and Ciaran S. Phibbs. 2023. "Better Performance for Right-Skewed Data Using an Alternative Gamma Model." *BMC Medical Research Methodology* 23 (298). https://doi.org/10.1186/s12874-023-02113-1.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://cran.r-project.org/web/packages/knitr/index.html.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/.