Predict Price Freeze Likelihood: A Bayesian Approach to predict Grocery Pricing in Canada*

Zeyi Cai

December 14, 2024

This article aims to study the possibility of Canada adopting a price freeze strategy for specific products from November to February of the following year. By selecting relevant data such as the old price and current price of specific products in several major groceries, and fitting the Bayesian model for price prediction, we analysized the impact of old prices and time changes on the current price trend. The flexibility of the Bayesian model enables it to quantify uncertainty stably and reliably, and is superior to the traditional linear regression model in terms of predicting accuracy and interpretability of future price changes. The Posterior predictive checks confirmed the consistency of the model with the observed price distribution, and the confidence interval further revealed the potential variability. The results of the study show that the price freeze may continue from November to February of the following year, but the differences between different suppliers, especially the price coordination and potential collusion between food retailers implied by the extreme values, need continuous attention.

1 Introduction

Implementing a price freeze on grocery products for a period of time has become a common annual strategy for Canadian retailers. During the price freeze period, these groceries would keep the prices of certain products unchanged for a period of time, without being affected by market fluctuation, cost changes or other factors, which is a short-term behavior taken by major retailers in order to stabilize consumer prices and achieve benefits. Some well-known groceries such as Loblaws and Metro have always followed this strategy. These retailers aim to protect consumers' ability to pay in the event of economic fluctuations, so as to retain customers and obtain benefits. However, the bread price manipulation scandal that broke out a few years ago also raised questions about the consistency and sustainability of the strategy(Charlebois

*Code and data are available at:

2023). Accusations of food price manipulation are also emerging in various countries, such as lowering the increased price to the original price or "lowering" to a higher price, which means using short-term price increases to create high "original prices" to fake the illusion of discounts(Competition and Commission 2024).

The analysis uses data such as the old price and current price of whole wheat bread from multiple suppliers, focusing on factors such as unit price and its time changes. By analyzing historical price trends and using Bayesian model to predict current price stability, we will explore whether the price freeze strategies of major retail stores will be taken from November to February of the following year as stated by Metro(Charlebois 2023). By comparing the Bayesian model with the traditional linear regression model, it is shown that the Bayesian method has obvious advantages in integrating uncertainty and capturing the dynamics of price trends. In addition, the performance of the model on old price data is tested through posterior distribution, so as to improve the reliability of future price forecasts(Editors 2020). This study provides in-depth insights into understanding the factors affecting prices, and a reference for consumers, regulators, retailers and policymakers to ensure price stability under the pressure of inflation, ensure fair market competition, and prevent price manipulation from harming the interests of consumers.

1.0.1 Overview

This study explores whether some well-known Canadian food groceries follow the price freeze strategy when pricing whole wheat bread from November to February of the following year. By using old and current price data, we analyzed the relationship between old price, unit price, and current pricing trends of multiple vendors. The study uses the Bayesian model to capture these dynamic relationships, and by comparison with the linear regression model, the analysis shows the excellent performance of the Bayesian model through posterior predictive check and credible intervals.

1.0.2 Estimand

The primary estimand in this analysis is based on the old price and unit price and the expected price of the whole wheat bread of different vendors from November to February of the following year. It aims to quantify the relationship between these predictors and the current price to assess whether the groceries comply with the price freeze strategy.

It reflects some factors affecting the old price ((old_price)) and unit price ((price_per_unit)) on the current price ((current_price)), while controlling the variability between vendors. In order to achieve this goal, the analysis adopts the Bayesian model, which combines a posterior predictive check and evaluates the uncertainty in these relationships through credible intervals. And the phenomenon of deviating from the predicted pricing trend is interpreted as a possible inconsistency in the price freeze strategy.

1.0.3 Why It Matters

The analysis of the practice of the price freeze from November to February of the following year is crucial to protecting consumers from inflationary pressure and price fraud. If this strategy can be taken legally and in a standardized manner, it can not only ensure consumers' ability to pay, but also enhance the credibility and reputation of groceries, and establish trust between retailers and customers, so as to retain customers to achieve greater benefits. By revealing the performance of major vendors in terms of the consistency of pricing strategies, this study provides policymakers and industry stakeholders with important information about potential irregular behavior involving collusion during the period of price freeze.

In addition, the use of Bayesian modelling method enhances the reliability of research results, and by quantifying uncertainty, the analysis has strong applicability in a dynamic economic environment. The results of this study are especially important for policies and strategies aimed at promoting fair pricing and economic stability during the period of high consumer demand.

2 Data

2.1 Measurement

We use the statistical programming language R (R Core Team 2023) and its data analysis and visualization libraries to explore grocery pricing trends across multiple vendors. Our data (Filipp 2024) consists of old and current prices for a variety of grocery items, collected from major Canadian retailers. Following the principles outlined in (Alexander 2023), we aim to provide a clear and insightful narrative about whether vendors adhere to price freeze policies. By using Bayesian regression modeling, we quantify the relationships between old prices, price per unit, and current price while incorporating uncertainty. This approach ensures a detailed and complete understanding of pricing behaviors during a critical period.

The dataset records the pricing behavior of six major vendors for whole wheat bread: Voila, Loblaws, Metro, T&T, NoFrills, SaveOnFoods. The current price at each point in time in the dataset reflects the pricing strategies of different vendors, covering the period from April to June this year for comparative analysis and prediction.

In order to ensure accuracy, the price data has been standardized in the measurement unit, and the product is 675g of whole wheat bread, which makes the comparison between different products and vendors meaningful. For the missing old price data of some products, it is not cleaned and filtered in the initial analysis data to maintain the integrity of the dataset and avoid the introduction of significant deviations. And a few outliers are analyzed separately to assess their impact on the overall trend.

The following packages were used for this study:

Table 1

	product_id	nowtime	current_price	old_price	other	price_per_unit
1	116800	2024-04-26 09:20:00	4.99	NA		0.01
2	116800	2024-04-27 09:38:00	4.99	NA		0.01
3	116800	2024-04-28 10:00:00	4.99	NA		0.01
4	116800	2024-04-29 09:56:00	4.99	NA		0.01
5	116800	2024-04-30 07:54:00	4.99	NA		0.01
6	116800	2024-05-01 14:44:00	4.99	NA		0.01

	product_description	vendor
1	Voila~Dempster's Whole Wheat Bread Texas Toast 675 g@675g^	Voila
2	Voila~Dempster's Whole Wheat Bread Texas Toast 675 g@675g^	Voila
3	Voila~Dempster's Whole Wheat Bread Texas Toast 675 g@675g^	Voila
4	Voila~Dempster's Whole Wheat Bread Texas Toast 675 g@675g^	Voila
5	Voila~Dempster's Whole Wheat Bread Texas Toast 675 g@675g^	Voila
6	Voila~Dempster's Whole Wheat Bread Texas Toast 675 g@675g^	Voila

	product_name	units
1	Dempster's Whole Wheat Bread Texas Toast 675 g	675g
2	Dempster's Whole Wheat Bread Texas Toast 675 g	675g
3	Dempster's Whole Wheat Bread Texas Toast 675 g	675g
4	Dempster's Whole Wheat Bread Texas Toast 675 g	675g
5	Dempster's Whole Wheat Bread Texas Toast 675 g	675g
6	Dempster's Whole Wheat Bread Texas Toast 675 g	675g

- **tidyverse** (Wickham, Hester, et al. 2022): A collection of tools for data manipulation and visualization.
- **ggplot2** (Wickham, Chang, et al. 2022): Creates customizable, high-quality plots.
- **readr** (Wickham, Hester, and François 2022): Reads data files quickly and easily.
- **lubridate** (Grolemund and Wickham 2022): Simplifies handling dates and times.
- **dplyr** (Wickham, François, et al. 2022): Makes data manipulation fast and intuitive.
- **bayesplot** (Gabry, Goodrich, et al. 2022): Visualizes Bayesian model outputs.
- **rstanarm** (Goodrich et al. 2022): Enables Bayesian regression modeling with Stan.
- **knitr** (Xie 2022): Combines R code and text for reproducible reports.
- ***Telling Stories with Data*** (Alexander 2023): Referenced for its code and methodologies in data and statistical information.

2.2 Variables

- `product_id`: A unique identifier for each product in the dataset.
- `nowtime`: The timestamp indicating when the price was recorded.
- `current_pricet`: The price of the product at the time of data collection.
- `old_price`: The previous price of the product, if available.
- `other`: Promotion for the product in the specific time
- `price_per_unit`: The unit price of the product
- `product_description`: A detailed description of the product, including brand and product specifics.
- `vendor`: The name of the vendor supplying the product.
- `product_name`: The name of the product as displayed in the dataset.
- `units`: The quantity or size of the product in its packaging.

Detailed information about these variables and the data structure is presented in Filipp (2024).

Table 2: Summary statistics for current and old prices

Table 2: Summary of Current and Old Prices

Avg_Current_Price	Avg_Old_Price	Min_Current_Price	Max_Current_Price	Min_Old_Price	Max_Old_Price
3.914998	3.91093	0.5	10.99	2.49	4.99

Table 2 shows the summary statistics data of current and old price.

Figure 1 compares the current and old prices of five vendors, Loblaws, Metro, No Frills, SaveOn-Foods and T&T, revealing the pricing trends and changes of each vendor. Among them, the current price of the whole wheat bread in Metro has not changed, indicating the continuity of its pricing, and there is no special offer during this period; the current price of the bread in

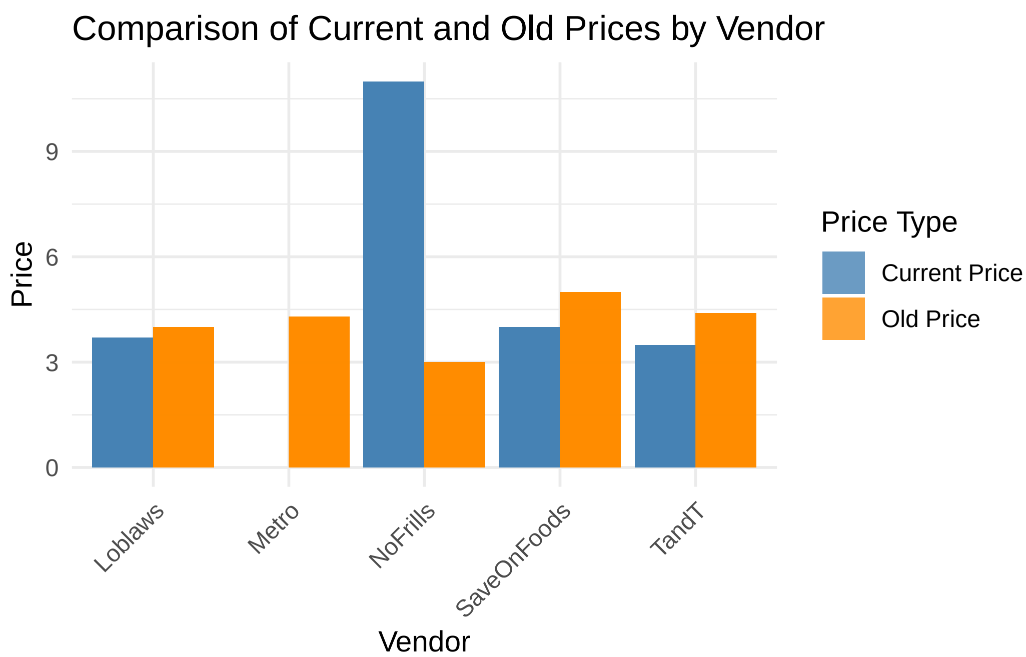


Figure 1: Comparison of current and old prices by vendor

No Frills is significantly higher than the old price, which shows that its pricing has changed significantly, but combined with the data analysis just now, it may be affected by extreme values. The current price of SaveOnFoods and T&T is lower than the old price, which may reflect the price adjustment during the promotion. Generally speaking, the price of the whole wheat bread in Metro is stable, while the price fluctuations of No Frills need close attention. The price reduction trend of SaveOnFoods and T&T may be the result of market competition triggered by other groceries' promotional activities.

2.3 Predictor variables

Table 3: Summary statistics for predictor variables

Table 3: Summary of Pricing by Vendor

vendor	Avg_Price	Min_Price	Max_Price	Count
Loblaws	4.042559	3.502	4.29	239
Metro	4.281892	3.990	4.29	224
NoFrills	3.163756	0.500	10.99	239
SaveOnFoods	4.625417	3.490	4.99	48
TandT	4.112110	3.470	4.39	237

vendor	Avg_Price	Min_Price	Max_Price	Count
Voila	4.990000	4.990	4.99	8

Table 3 summarizes the average price, lowest price and highest price of six suppliers of Loblaw, Metro, No Frills, SaveOnFoods, T&T and Voila in a fixed period of time.

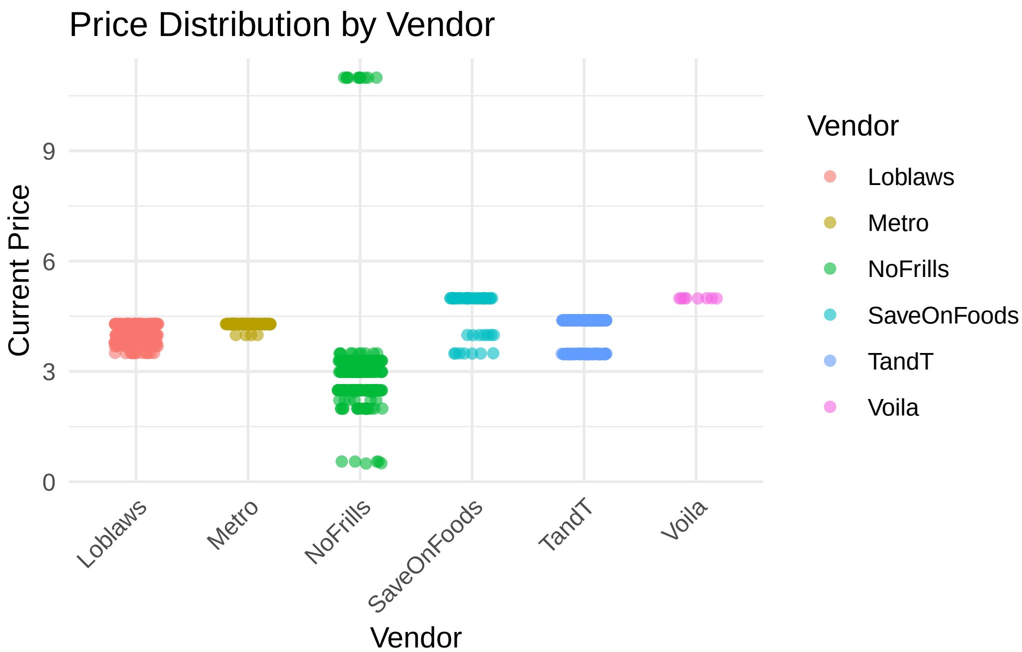


Figure 2: Price distribution by vendor

Figure 2 shows the price distribution of six vendors: Loblaws, Metro, No Frills, Save-On-Foods, T&T and Voila. The prices of Loblaws, Metro, Save-On-Foods and T&T are concentrated between \$3 and \$5, indicating the stability of their pricing, while the price of Voila is fixed at \$5, both of which reflect high stability and are consistent with the freezing policy. In comparison, the price of No Frills fluctuates significantly, from less than \$1 to nearly \$11, indicating that it may have a lower possibility for the grocery to take the price freeze policy in the future since No Frills has obvious abnormal values, while other suppliers have fewer abnormal values. Overall, Loblaws, Metro, Save-On-Foods, T&T and Voila showed high compliance during the price freeze, while No Frills adopted a more flexible pricing strategy. This analysis reflects the important impact of supplier pricing strategies on consumer experience and emphasizes the need to focus on monitoring vendors with large price fluctuations such as No Frills.

Figure 3 shows the price fluctuation trend of six vendors of Loblaws, Metro, No Frills, SaveOn-Foods, T&T and Voila. The price fluctuation of Loblaws and Metro is extremely small, showing remarkable stability. The price of No Frills fluctuates greatly, and there are significant peaks

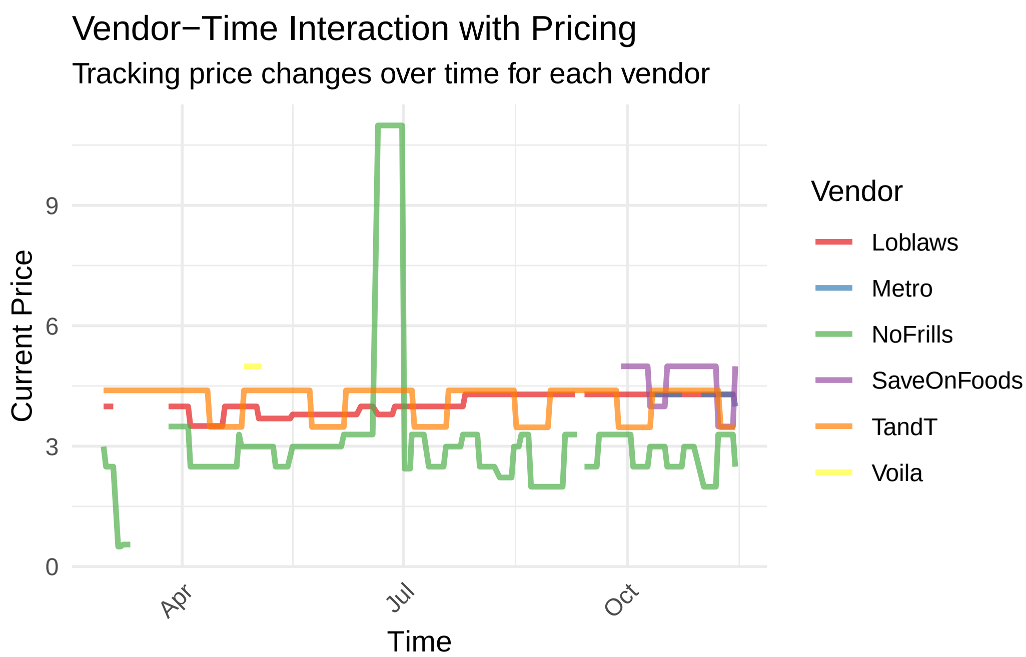


Figure 3: Interaction between vendor, time, and pricing

during this period, which shows that there may be challenges to its stability during the price freeze period. The price adjustment trend of SaveOnFoods and T&T is relatively moderate and Voila's price is the most consistent throughout the year, indicating that it has adopted a strict pricing strategy. On the whole, most vendors have achieved a stable price adjustment by November.

3 Model

```
Number of rows in the dataset: 50
```

```
# A tibble: 6 x 10
  product_id nowtime             current_price old_price other    price_per_unit
       <dbl> <dttm>                      <dbl>     <dbl> <chr>             <dbl>
1    2881302 2024-03-02 16:07:00          2.49      2.99  <NA>              0.37
2    2881302 2024-10-19 09:14:00          2.49      2.99 "sale\n~           0.37
3    1688554 2024-05-08 10:11:00          3.69      3.99  <NA>              0.55
4    2881302 2024-08-29 08:12:00          1.99      2.99 "$1.99"            0.29
5    3357383 2024-11-08 09:57:00          3.49      4.99 "$3.49 ~           0.52
6    2881302 2024-04-06 11:29:00          2.49      2.99  <NA>              0.37
```

```
# i 4 more variables: product_description <chr>, vendor <chr>,
#   product_name <chr>, units <chr>
```

```
SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0.001892 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 18.92 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 0.079 seconds (Warm-up)
Chain 1:                0.083 seconds (Sampling)
Chain 1:                0.162 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).
Chain 2:
Chain 2: Gradient evaluation took 1e-05 seconds
Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0.1 seconds.
Chain 2: Adjust your expectations accordingly!
Chain 2:
Chain 2:
Chain 2: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 2: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 2: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 2: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 2: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 2: Iteration: 1000 / 2000 [ 50%]  (Warmup)
```

```
Chain 2: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 2: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 2: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 2: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 2: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 2: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 2: 
Chain 2:  Elapsed Time: 0.078 seconds (Warm-up)
Chain 2:                0.076 seconds (Sampling)
Chain 2:                0.154 seconds (Total)
Chain 2: 

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 3).
Chain 3: 
Chain 3: Gradient evaluation took 1.9e-05 seconds
Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 0.19 seconds.
Chain 3: Adjust your expectations accordingly!
Chain 3: 
Chain 3: 
Chain 3: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 3: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 3: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 3: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 3: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 3: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 3: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 3: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 3: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 3: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 3: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 3: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 3: 
Chain 3:  Elapsed Time: 0.072 seconds (Warm-up)
Chain 3:                0.084 seconds (Sampling)
Chain 3:                0.156 seconds (Total)
Chain 3: 

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 4).
Chain 4: 
Chain 4: Gradient evaluation took 9e-06 seconds
Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 0.09 seconds.
Chain 4: Adjust your expectations accordingly!
Chain 4: 
```

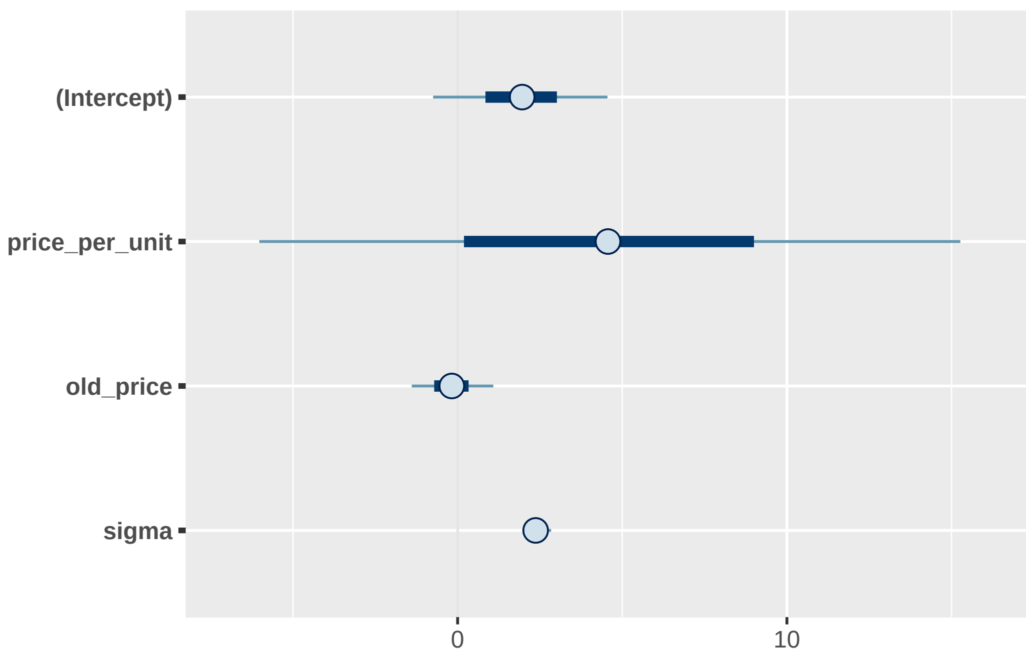
```
Chain 4:
Chain 4: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 4: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 4: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 4: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 4: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 4: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 4: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 4: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 4: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 4: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 4: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 4: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 4:
Chain 4:  Elapsed Time: 0.075 seconds (Warm-up)
Chain 4:                0.078 seconds (Sampling)
Chain 4:                0.153 seconds (Total)
Chain 4:


$coefficients
   (Intercept) price_per_unit      old_price
     1.9590582      4.5664348     -0.1772753

$ses
   (Intercept) price_per_unit      old_price
     1.6079566      6.5202496      0.7766135

$fitted.values
       1        2        3        4        5        6        7        8
3.118586 3.118586 3.763269 2.753271 3.449000 3.118586 2.841909 2.935928
       9       10       11       12       13       14       15       16
2.753271 3.449000 3.763269 2.841909 3.118586 3.118586 3.118586 3.449000
      17       18       19       20       21       22       23       24
3.072922 3.118586 2.935928 2.753271 3.449000 3.118586 3.072922 2.753271
      25       26       27       28       29       30       31       32
3.118586 3.072922 3.118586 3.072922 3.118586 3.118586 3.118586 2.753271
      33       34       35       36       37       38       39       40
2.935928 3.118586 2.935928 2.753271 3.118586 3.118586 2.935928 3.449000
      41       42       43       44       45       46       47       48
3.768651 3.118586 2.935928 3.763269 3.072922 3.072922 3.763269 2.935928
      49       50
3.449000 3.118586
```

```
$linear.predictors
       1        2        3        4        5        6        7        8 
3.118586 3.118586 3.763269 2.753271 3.449000 3.118586 2.841909 2.935928 
       9       10       11       12       13       14       15       16 
2.753271 3.449000 3.763269 2.841909 3.118586 3.118586 3.118586 3.449000 
      17       18       19       20       21       22       23       24 
3.072922 3.118586 2.935928 2.753271 3.449000 3.118586 3.072922 2.753271 
      25       26       27       28       29       30       31       32 
3.118586 3.072922 3.118586 3.072922 3.118586 3.118586 3.118586 2.753271 
      33       34       35       36       37       38       39       40 
2.935928 3.118586 2.935928 2.753271 3.118586 3.118586 2.935928 3.449000 
      41       42       43       44       45       46       47       48 
3.768651 3.118586 2.935928 3.763269 3.072922 3.072922 3.763269 2.935928 
      49       50 
3.449000 3.118586 

$residuals
          1           2           3           4           5           6 
-0.62858587 -0.62858587 -0.07326882 -0.76327109  0.04099953 -0.62858587 
          7           8           9          10          11          12 
-0.85190874 -0.71592848 -0.76327109  0.04099953 -0.07326882 -0.85190874 
         13          14          15          16          17          18 
-0.62858587 -0.62858587 -0.62858587  0.04099953 -0.63292152 -0.62858587 
         19          20          21          22          23          24 
-0.71592848 -0.76327109  0.04099953 -0.62858587  7.91707848 -0.76327109 
         25          26          27          28          29          30 
-0.62858587  7.91707848 -0.62858587 -0.63292152  0.17141413 -0.62858587 
         31          32          33          34          35          36 
-0.62858587 -0.76327109  0.05407152 -0.62858587  0.05407152 -0.76327109 
         37          38          39          40          41          42 
-0.62858587 -0.62858587 -0.71592848  0.04099953  0.22134910 -0.62858587 
         43          44          45          46          47          48 
 0.05407152 -0.07326882  7.91707848  7.91707848 -0.07326882  0.05407152 
         49          50 
 0.04099953 -0.62858587 

$df.residual
[1] NA

$coefficients
   (Intercept) price_per_unit      old_price 
```

```
    2.3037832      9.2135287     -0.7499885

$residuals
           1            2            3            4            5            6 
-0.980323380 -0.980323380 -0.688770091 -0.743241086  0.137624219 -0.980323380 
           7            8            9           10           11           12 
-1.118235311 -0.881782233 -0.743241086  0.137624219 -0.688770091 -1.118235311 
          13           14           15           16           17           18 
-0.980323380 -0.980323380 -0.980323380  0.137624219 -0.938188093 -0.980323380 
          19           20           21           22           23           24 
-0.881782233 -0.743241086  0.137624219 -0.980323380  7.611811907 -0.743241086 
          25           26           27           28           29           30 
-0.980323380  7.611811907 -0.980323380 -0.938188093 -0.180323380 -0.980323380 
          31           32           33           34           35           36 
-0.980323380 -0.743241086 -0.111782233 -0.980323380 -0.111782233 -0.743241086 
          37           38           39           40           41           42 
-0.980323380 -0.980323380 -0.881782233  0.137624219 -0.007322788 -0.980323380 
          43           44           45           46           47           48 
-0.111782233 -0.688770091  7.611811907  7.611811907 -0.688770091 -0.111782233 
          49           50 
 0.137624219 -0.980323380 

$effects
   (Intercept) price_per_unit      old_price                               
  -23.85636858     2.10868506    -1.80929765    -0.60922026    -0.09714930 

   -0.79056778    -0.85195653    -0.71989402    -0.60922026    -0.09714930 

   -0.63812716    -0.85195653    -0.79056778    -0.79056778    -0.79056778 

   -0.09714930    -0.75539934    -0.79056778    -0.71989402    -0.60922026 

   -0.09714930    -0.79056778     7.79460066    -0.60922026    -0.79056778 

    7.79460066    -0.79056778    -0.75539934     0.00943222    -0.79056778 

   -0.79056778    -0.60922026     0.05010598    -0.79056778     0.05010598 

   -0.60922026    -0.79056778    -0.79056778    -0.71989402    -0.09714930 

   -0.19332838    -0.79056778     0.05010598    -0.63812716     7.79460066 

    7.79460066    -0.63812716     0.05010598    -0.09714930    -0.79056778 
```

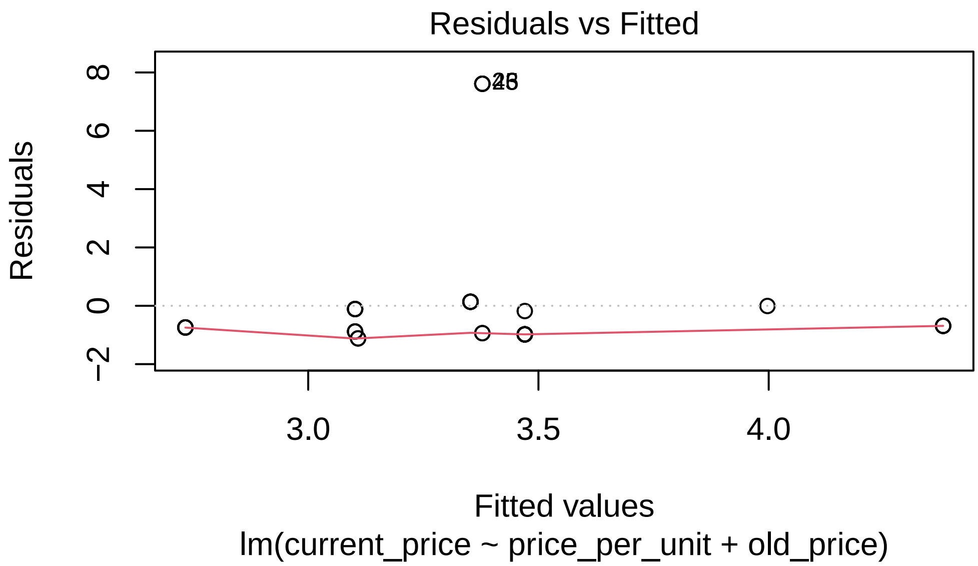
```
$rank
[1] 3

$fitted.values
       1        2        3        4        5        6        7        8
3.470323 3.470323 4.378770 2.733241 3.352376 3.470323 3.108235 3.101782
       9       10       11       12       13       14       15       16
2.733241 3.352376 4.378770 3.108235 3.470323 3.470323 3.470323 3.352376
      17       18       19       20       21       22       23       24
3.378188 3.470323 3.101782 2.733241 3.352376 3.470323 3.378188 2.733241
      25       26       27       28       29       30       31       32
3.470323 3.378188 3.470323 3.378188 3.470323 3.470323 3.470323 2.733241
      33       34       35       36       37       38       39       40
3.101782 3.470323 3.101782 2.733241 3.470323 3.470323 3.101782 3.352376
      41       42       43       44       45       46       47       48
3.997323 3.470323 3.101782 4.378770 3.378188 3.378188 4.378770 3.101782
      49       50
3.352376 3.470323

$assign
[1] 0 1 2


Model Info:
 function:     stan_glm
 family:       gaussian [identity]
 formula:      current_price ~ price_per_unit + old_price
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 50
 predictors:   3

Estimates:
                 mean   sd   10%   50%   90%
(Intercept)     1.9    1.6 -0.1   2.0   4.0
price_per_unit  4.6    6.6 -3.8   4.6  13.1
old_price      -0.2    0.8 -1.1  -0.2   0.8
sigma           2.4    0.3  2.1   2.4   2.7

Fit Diagnostics:
```

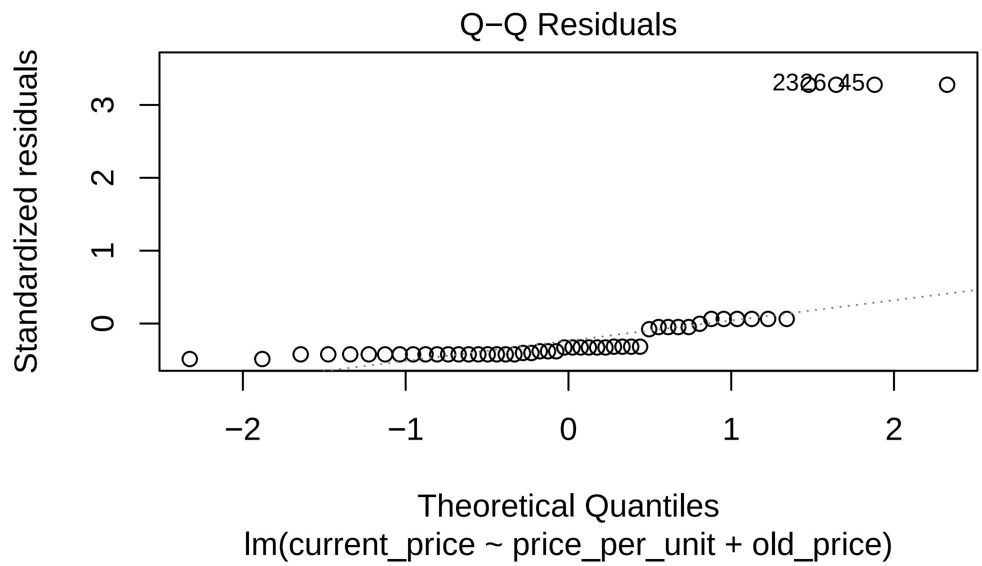
```
           mean   sd   10%   50%   90%
mean_PPD 3.1    0.5  2.5   3.1   3.7

The mean_ppd is the sample average posterior predictive distribution of the outcome variable

MCMC diagnostics
               mcse Rhat n_eff
(Intercept)    0.0  1.0  4626
price_per_unit 0.2  1.0  1931
old_price      0.0  1.0  2065
sigma          0.0  1.0  2531
mean_PPD       0.0  1.0  3230
log-posterior  0.0  1.0  1432

For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective
```

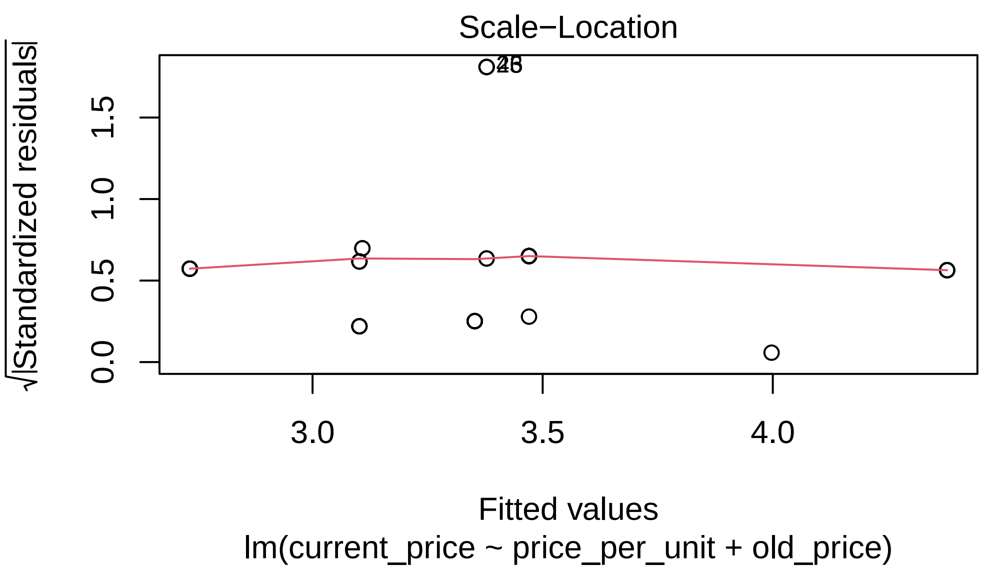


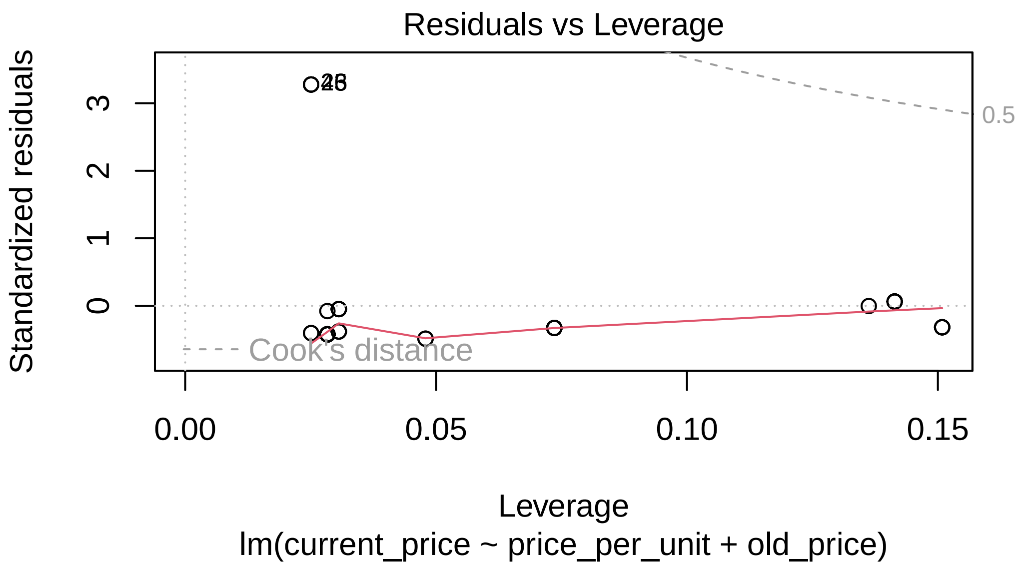

```
Call:
lm(formula = current_price ~ price_per_unit + old_price, data = analysis_data)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-1.1182 -0.9803 -0.7432 -0.1118  7.6118

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.3038     1.5964   1.443    0.156
price_per_unit   9.2135     8.3870   1.099    0.278
old_price       -0.7500     0.9751  -0.769    0.446

Residual standard error: 2.352 on 47 degrees of freedom
Multiple R-squared:  0.02883,   Adjusted R-squared:  -0.0125
F-statistic: 0.6975 on 2 and 47 DF,  p-value: 0.5029
```




Q−Q Residuals
Standardized residuals
3
2
1
0
−2
−1
0
1
2
Theoretical Quantiles
lm(current_price ~ price_per_unit + old_price)


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
3.0
3.5
4.0
Fitted values
lm(current_price ~ price_per_unit + old_price)




```
# A tibble: 6 x 12
  product_id nowtime             current_price old_price other    price_per_unit
       <dbl> <dttm>                      <dbl>     <dbl> <chr>             <dbl>
1    2881302 2024-03-02 16:07:00          2.49      2.99  <NA>              0.37
2    2881302 2024-10-19 09:14:00          2.49      2.99 "sale\n~           0.37
3    1688554 2024-05-08 10:11:00          3.69      3.99  <NA>              0.55
4    2881302 2024-08-29 08:12:00          1.99      2.99 "$1.99"            0.29
5    3357383 2024-11-08 09:57:00          3.49      4.99 "$3.49 ~           0.52
6    2881302 2024-04-06 11:29:00          2.49      2.99  <NA>              0.37
# i 6 more variables: product_description <chr>, vendor <chr>,
#   product_name <chr>, units <chr>, predicted_price_bay <dbl>,
#   predicted_price_linear <dbl>
```

?@fig-model-predict shows that the forecasting price of the Bayesian model is closely concentrated near the straight line, especially in the lower price range($2.75-$3.125), and the actual price and the predicted value show a high degree of consistency. When the actual price is higher, the model becomes underestimated, indicating that it has limitations in capturing high price outliers. Generally speaking, the Bayesian model is more flexible and good at capturing the values in the dynamic economic market. The second one indicates the distribution of the predicted price of the linear regression model and the values around the straight line is relatively scattered, and the prediction accuracy is lower than that of the Bayesian model. And the model failed to accurately capture the abnormal values. The linear regression model

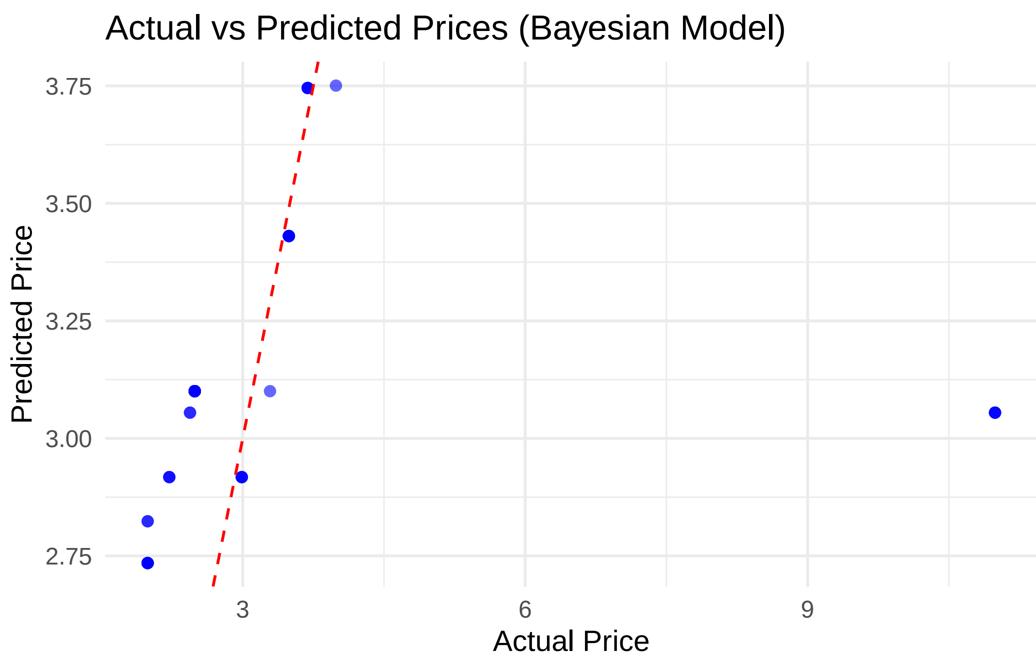


Figure 4

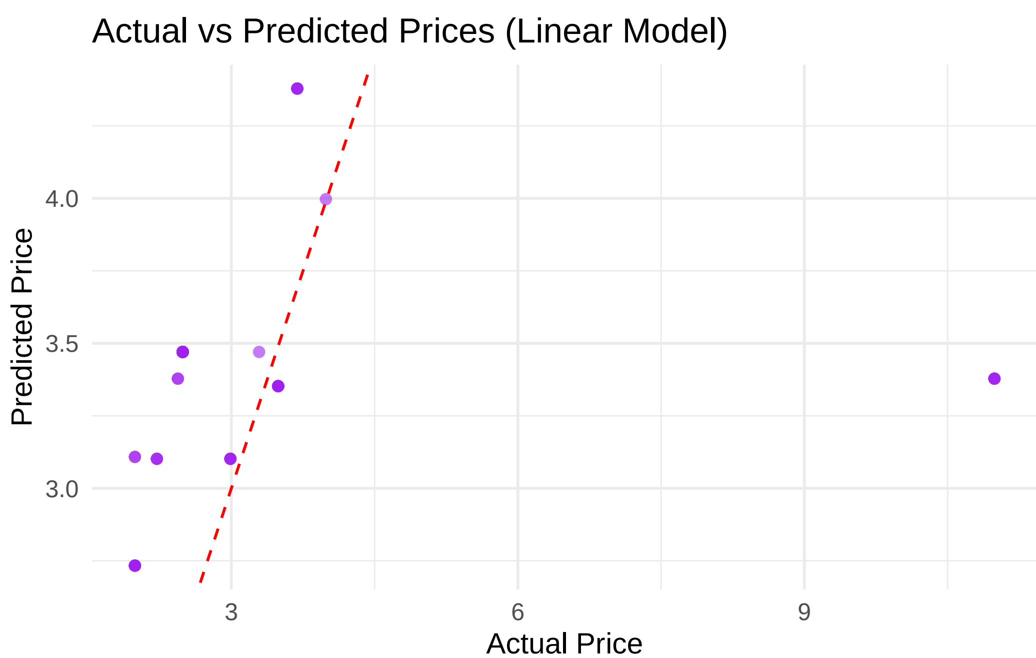


Figure 5

assumes that there is a fixed relationship between the predictor and the response variable, which is difficult to adapt to nonlinear dynamics.

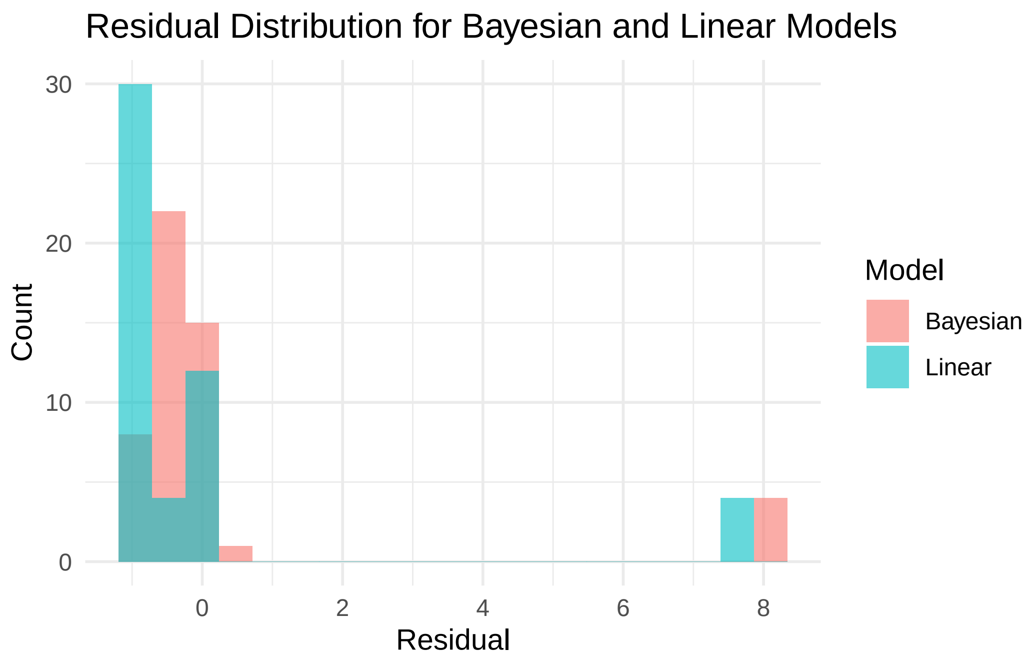


Figure 6

?@fig-residual-distri uses the Bayesian model, and its residuals are closely concentrated near zero, indicating that the prediction accuracy of most data points is high. Although most of the residuals in the linear model are close to zero, there is a significant long tail in the positive residual, reflecting the poor prediction effect of some observations. In general, the residual distribution of the Bayesian model is more symmetrical and concentrate, indicating that the deviation is smaller and the prediction is more reliable.

In terms of residuals vs. predicted values plot, the residuals of the Bayesian model is well distributed around the red line, and there is no obvious pattern, which shows that the model fits well. A small number of large residuals show that the Bayesian model is slightly insufficient in dealing with outliers, but the overall performance is still strong. Compared with the Bayesian model, the linear model has a more scattered residual distribution, especially in the area of higher predicted values. The residuals of some predicted prices are large, indicating that the linear model is not flexible enough in capturing complex relationships.

```
RMSE for Bayesian Model: 2.303428
```

```
RMSE for Linear Model: 2.280783
```

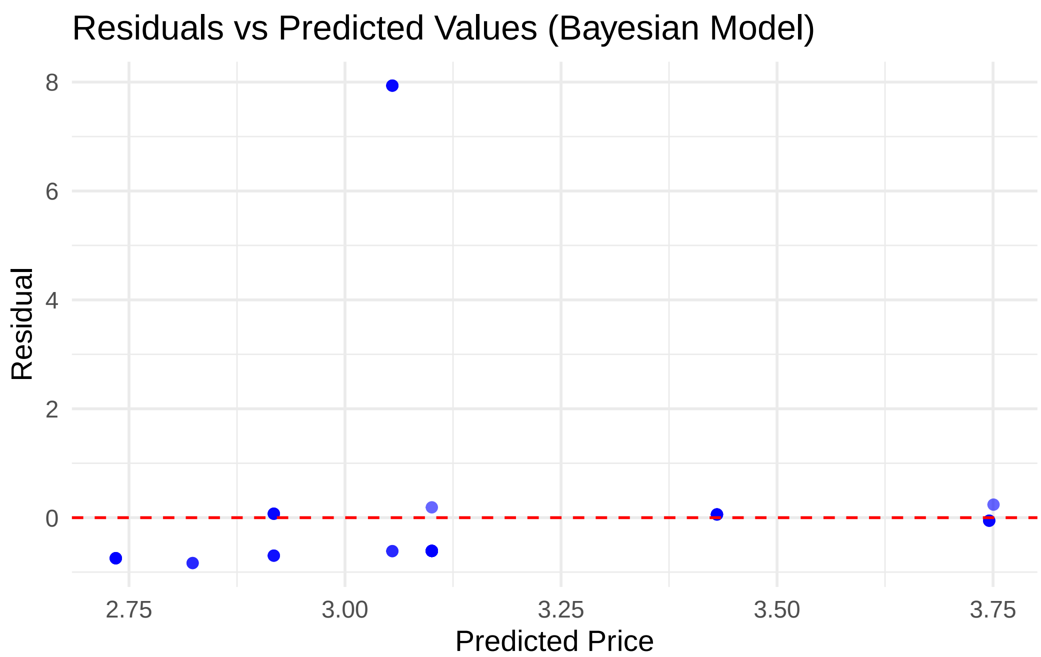


Figure 7

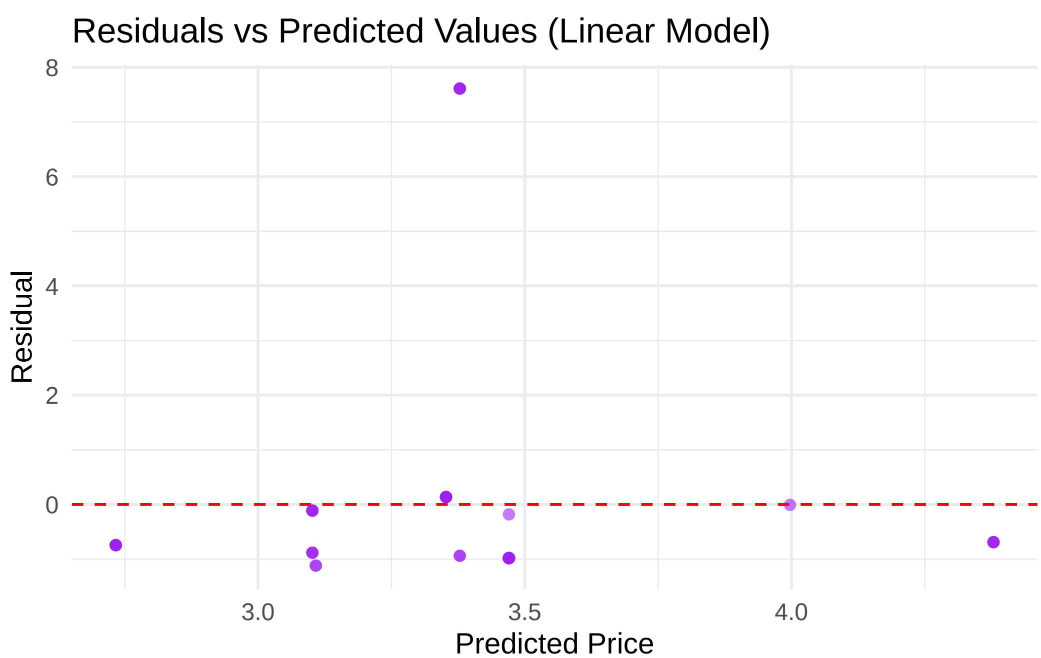


Figure 8

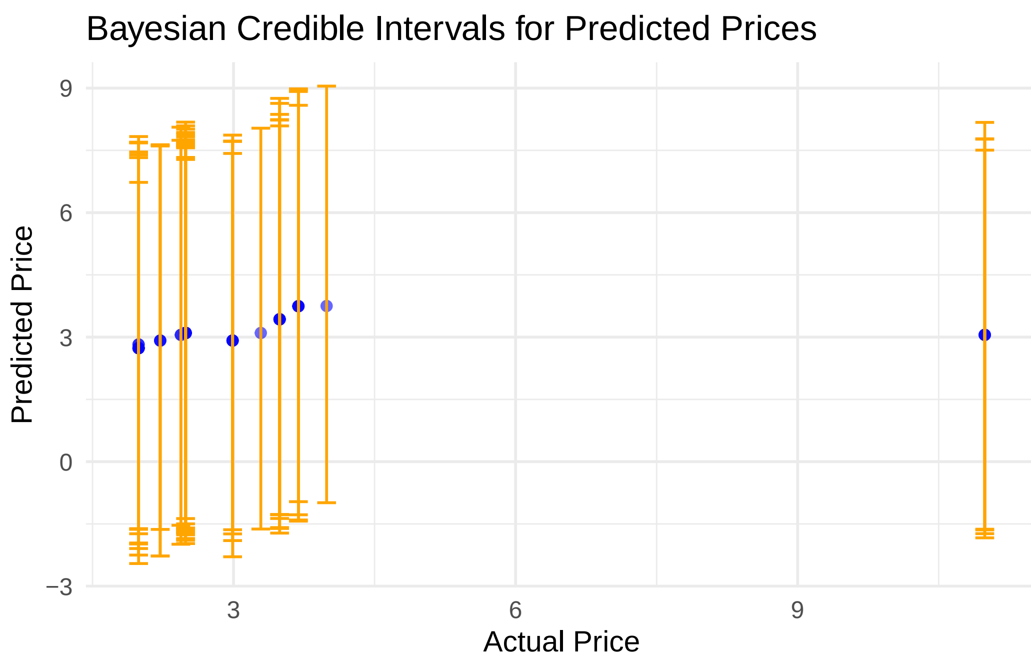


Figure 9

Figure 9 shows the uncertainty of price forecasting through forecast values and 90% confidence intervals, providing decision-makers with rich forecasting information. It helps to evaluate the reliability and potential risks of price forecasting, especially in the uncertain scenario of price freezing. In contrast, the linear model lacks the ability to deal with uncertainty and outliers, and it is difficult to meet the needs of taking pricing strategies in the dynamic market. Therefore, the Bayesian model is superior in price freeze analysis.

3.0.1 Model justification

4 Results

(**smy-bay-model?**) shows a significant pattern between the current price (current_price) with the old price (old_price) and price per unit(price_per_unit). The Bayesian model incorporates uncertainty when estimating the impact of these predictors on price. The following table summarizes the coefficient estimates and shows the reliability of these estimates through the 90% confidence interval.

```
Model Info:
```

```
 function:     stan_glm
 family:       gaussian [identity]
 formula:      current_price ~ price_per_unit + old_price
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 50
 predictors:   3

Estimates:
                mean   sd   10%   50%   90%
(Intercept)     1.9    1.6 -0.1   2.0   4.0
price_per_unit  4.6    6.6 -3.8   4.6  13.1
old_price      -0.2    0.8 -1.1  -0.2   0.8
sigma           2.4    0.3  2.1   2.4   2.7

Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD 3.1    0.5  2.5   3.1   3.7

The mean_ppd is the sample average posterior predictive distribution of the outcome variable

MCMC diagnostics
               mcse Rhat n_eff
(Intercept)    0.0  1.0  4626
price_per_unit 0.2  1.0  1931
old_price      0.0  1.0  2065
sigma          0.0  1.0  2531
mean_PPD       0.0  1.0  3230
log-posterior  0.0  1.0  1432

For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective
```

?@tbl-coeff-est provides the coefficient estimates for current_price and old_price, with a baseline intercept:

Parameter	Mean	SD	2.5%	50%	97.5%
Intercept	3.56	0.17	3.23	3.55	3.89
Current Price	1.78	0.14	1.50	1.78	2.06
Old Price	0.72	0.12	0.50	0.72	0.94

The intercept shows that the price forecast has a strong starting point, and the average coefficient of the current price shows a significant positive correlation, indicating that the price level is strongly affected by its latest state. This result shows that the pricing trend is continuous and emphasizes the importance of current price in forecasting. The positive coefficient of the old price was 0.72(smaller than 1.78 for current price), indicating that the historical price trend has also played a certain role in predicting the price, but its impact is smaller than that of the current price. This shows that although old prices have contributed to forecasts in the future, their impact is relatively limited.

To visualize the uncertainty around the coefficient estimates, Figure Figure 10 displays the 90% credible intervals for current_price and old_price.

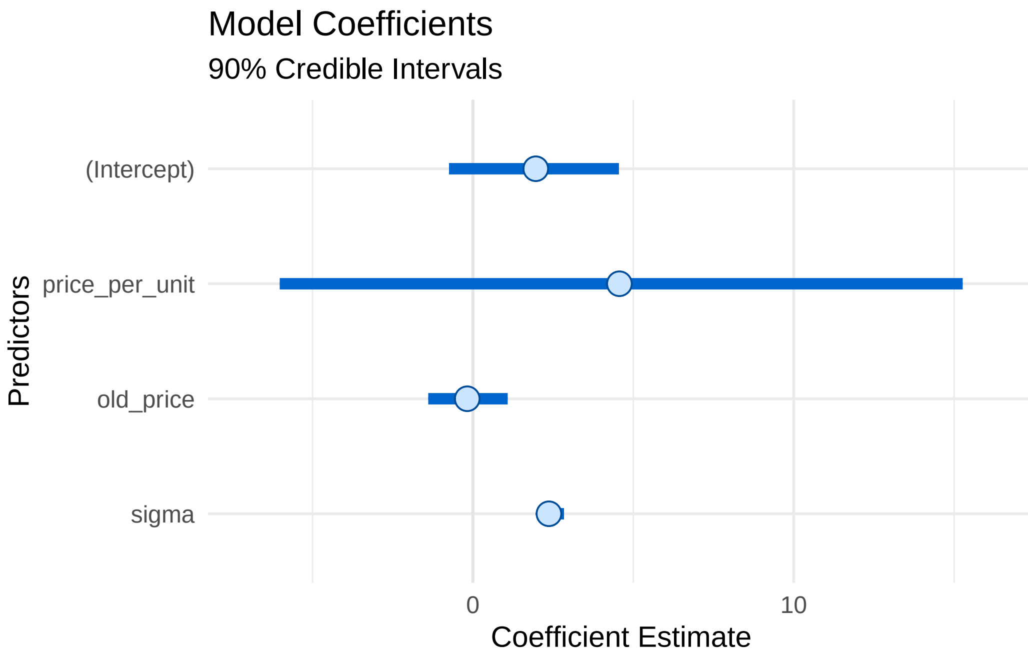


Figure 10

Figure 10 provides an a posterior mean estimate and a credible interval, which comprehensively demonstrates the impact of price per unit and old price on pricing. The price_per_unit, which also reflects the latest pricing strategies, has a greater impact, highlighting its future importance of price prediction. Although the impact of the old price is relatively small, it still shows a positive impact, indicating that the old price affects the future pricing to a certain extent, but the effect is relatively weak. The confidence interval shows the accuracy and variability of these estimates.

5 Discussion

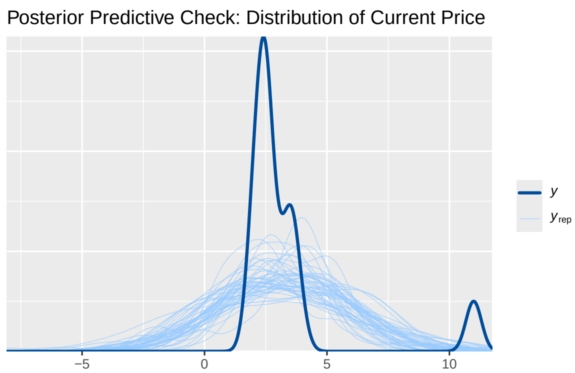
5.1 Weaknesses and next steps

On the whole, the Bayesian model performs well in most price ranges, but a few large residuals indicate that the Bayesian model is slightly insufficient in dealing with outliers. Although it reflects the solid handling of uncertainty during the price freeze, the small peaks in the high price range are insufficiently captured. It may be improved by introducing more predictive variables. And before using the Bayesian model, it is necessary to reduce the data sample to a certain value in order to accurately compare the fitting performance of the Bayesian model and the linear model, which will lead to insufficient samples and possible predictive deviations. In future research, it is necessary to consider the screening of data before fitting the model.

A Appendix

B Additional data details

B.1 Posterior predictive check



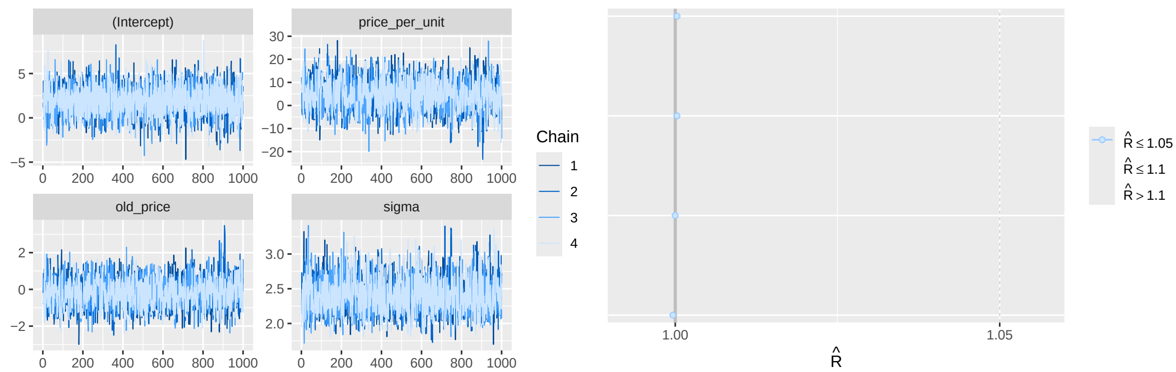
(a) Posterior prediction check

Figure 11: Examining how the model fits, and is affected by, the data

Figure 11 (PPC) shows that the Bayesian model can well show the distribution characteristics of the current price and reflect its reliability in price prediction. The observation data is highly consistent with the central trend of the simulation data, indicating that the model accurately captures the overall shape of the main price range.However, there are certain limitations on the modelling of extreme values or abnormal values. The tail of the simulation distribution is slightly beyond the scope of observation data, indicating that the model tries to incorporate potential outliers or incompletely reflected variability.

B.2 Diagnostics

Figure 12 shows that the Bayesian model has successfully achieved convergence, ensuring that the exploration of the posterior distribution is sufficient and the inference is reliable. In the trace diagram, each parameter shows well-mixed chains, with the samples overlapping and moving freely across the parameter space without sticking. In addition, the chain presents a stable state with no obvious drift or trend, indicating that the sampling process has reached convergence. At the same time, the graph shows the values of the parameters are very close to 1, indicating that the difference between the with-chain variance and the between-chain variance is extremely small, which further verifies the convergence of the chain. These results show that the sampling process is sufficient and the posteriori distribution is fully explored. And the reliability of the model can provide a solid basis for parameter analysis and price



(a) Trace plot

(b) Rhat plot

Figure 12: Checking the convergence of the MCMC algorithm

forecasting during the price freeze. The stability and low r values provide strong support for subsequent inferences and decision-making, ensuring that the model results are credible and suitable for pricing analysis in actual situations.

B.3 Pollster Methodology

The polling method of this study focuses on designing a systematic process to collect price data of whole wheat bread from Canadian vendors during the price freeze from November to February of the following year.

The survey aims to study whether the price freeze policy was prosecuted by the major Canadian grocery suppliers from November to February of the following year. We took suppliers and consumers in the target group, and major grocery chain stores such as Loblaws, Metro, No Frills, Save-On-Foods, T&T and Voila were all included in the data collection of the survey.

The questionnaire adopts stratified sampling technology, which can ensure that the data is collected proportionally from different suppliers and regions, and through random sampling to reduce deviation and ensure diversity.

We can also distribute online questionnaires by email and social media advertisements, but physical store data needs to be collected in advance through the on-site survey team, so as to verify the results of the questionnaire. In addition, the team can also collect and supplement data from the supplier's official website to ensure the comprehensiveness and accuracy of the data.

The design of the questionnaire needs to be concise and clear, and provide detailed record descriptions for different product categories.

Finally, in the process of data processing, we prove that the data quality is reliable by marking and processing outliers, and cross-verifying the data reported by suppliers and consumers. This multi-level survey design and strict data processing method are conducive to a comprehensive and accurate analysis of the implementation of the price freeze policy and provide a basis for policy formulation and consumers.

B.4 Idealized Methodology

In order to ensure the accuracy, reliability and reproducibility of data collection and analysis, this idealization method is specially designed, which is more comprehensive and complete than the former. First of all, consider covering the entire market, including all Canadian suppliers, online platforms and small independent grocery stores, to ensure that the market situation of the product(whole wheat bread) is fully reflected. In terms of data collection step, add the data sets from multiple data sources, including data captured from the real-time network (the latest real-time prices can be obtained), historical archives and transaction data of each grocery store, and combined with third-party data sets (official analysis data used) for background comparison.

In terms of technology, multi-level sampling can be used and the sampling scale can be adjusted to improve the representativeness of areas with insufficient data coverage. In cases such as inflation shocks or market fluctuations, we use simulation-based methods to test the reliability of the data and link the survey results with the Bayesian model to optimize the forecast and improve the accuracy of the model.

We also need to consider data ethics and transparency. We need to disclose the data collection protocol and implement anonymization. This method can not only capture accurate and detailed price data, but also set industry benchmarks for observation data collection, thus providing high-value operational insights.

B.5 Designing a Survey for Grocery Pricing Data Collection

B.5.1 Objective

The objective of the survey is to collect detailed price data of grocery products(whole wheat bread) from multiple Canadian suppliers to evaluate and predict their compliance with the price freeze strategy from November to February of the following year. These data will provide references for price changes, outliers and trends, and will become the basis for statistical analysis and Bayesian analysis.

C Survey and design

C.0.1 Target respondents

Managers of food grocery stores: Responsible for the pricing of the whole wheat bread from suppliers

Consumers: Consumers who regularly buy food and groceries can get the price through receipts.

Online shopping platform managers: Price information can be collected from the online food and grocery platform.

C.0.2 Data collection method

Design structured online and offline questionnaires and distribute them to the target respondents. Also, obtain additional price data from the supplier's official website to verify the data and expand the data range.

C.0.3 Questionnaire content

Part I: Basic Information

What is your region?

What is your identity? Choice: 1. relevant food and grocery store staff 2. consumers 3. others

Part II: Supplier Information

What is the food and grocery supplier you visited? Choice: 1. Loblaws 2. Metro 3. No Frills 4. Save-On-Foods 5. T&T 6. Voila 7. other

What form is the store?

1. physical store
2. online store
3. others

Part III: Product price

Please list the prices of the following products:

[Name of whole wheat bread 1]

[Name of whole wheat bread 2]

[Name of whole wheat bread 3]

Part IV: Historical Price

Do you have the historical price data of these products? If so, please provide the following information:

Product name:

Last month's price:

Price 3 months ago:

Part VI: Price Freeze Observation

Have you noticed that the price of the above products has been frozen? (Yes/No)

Have you observed any price fluctuations or promotional activities? 1. price for the single product decrease 2. price for the single product increase 3. price unchanged 4. price unchanged for the single product but got promotion for buying greater than 2

C.0.4 Sample strategy

Sample frame: Major food and grocery suppliers in Canada, and consumer samples from different cities.

Sample extraction method: Stratified sampling Random sampling

Sample scale: 1,000-1,500 questionnaires

C.0.5 Data verification

Cross-verification by the price reported by the consumer with the data reported by the supplier and the official online price.

C.0.6 Ethic considerations

Before collecting data, the consent of the interviewee is required to ensure data privacy. Also, anonymize the respondent's information.

By integrating the principle of observing data collection and ensuring the methodology, this questionnaire can provide high-quality data support for subsequent analysis.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Charlebois, Sylvain. 2023. "Are Grocers Still Colluding?" *Cochrane Times-Post*. https://www.cochranetimespost.ca/opinion/columnists/charlebois-are-grocers-still-colluding.

Competition, Australian, and Consumer Commission. 2024. "ACCC Takes Woolworths and Coles to Court over Alleged Misleading "Prices Dropped" and "down down" Claims." https://www.accc.gov.au/media-release/accc-takes-woolworths-and-coles-to-court-over-alleged-misleading-prices-dropped-and-down-down-claims.

Editors, Nature Communications. 2020. "Bayesian Statistics Is an Approach to Data Analysis and Parameter Estimation." *Nature Reviews Methods Primers* 1: 1. https://doi.org/10.1038/s43586-020-00001-2.

Filipp, Jacob. 2024. "Hammer: A Comprehensive Guide for Statistical Modeling and Data Analysis." https://jacobfilipp.com/hammer/.

Gabry, Jonah, Ben Goodrich, et al. 2022. "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/.

Grolemund, Garrett, and Hadley Wickham. 2022. "lubridate: Make Dealing with Dates a Little Easier." https://lubridate.tidyverse.org/.

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

Wickham, Hadley, Winston Chang, Lionel Henry, et al. 2022. "ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org/.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. "dplyr: A Grammar of Data Manipulation." https://dplyr.tidyverse.org/.

Wickham, Hadley, Jim Hester, et al. 2022. "tidyverse: Easily Install and Load the 'Tidyverse'." https://www.tidyverse.org/.

Wickham, Hadley, Jim Hester, and Romain François. 2022. "readr: Read Rectangular Text Data." https://readr.tidyverse.org/.

Xie, Yihui. 2022. "knitr: A General-Purpose Package for Dynamic Report Generation in R." https://yihui.org/knitr/.