Pricing Patterns in Organic Markets*

Comparative Analysis of Vendor Strategies and Seasonal Trends in Organic Product Sales

Dingshuo Li

December 2, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: https://github.com/dawsonlll/Organic-Product-Pricing-Analysis.git.

1 Introduction

The organic food market has witnessed a remarkable surge in demand, driven by increasing consumer awareness of health and environmental impacts. This demand has not only heightened the market value of organic products but has also introduced complexities into their pricing strategies across various retail platforms. Notably, organic food typically incurs higher production and certification costs compared to conventional food products, influencing its retail pricing and accessibility. This study aims to dissect these pricing strategies by examining how the prices of organic foods vary across different vendors and over successive months.

This paper explores the intricate dynamics of organic food pricing within a well-defined dataset that records prices from multiple vendors over a specified period. The analysis is structured to address a gap in existing research: the detailed quantification of how prices for organic products are structured and vary in a real-market scenario. Utilizing advanced Bayesian statistical methods, this research assesses the relationship between current and historical prices while considering the influence of temporal factors captured through monthly variations.

Estimand: The primary estimand of this study is the price elasticity of organic foods across different market vendors and months. Specifically, the analysis seeks to estimate how changes in the historical prices (old prices) influence the current pricing strategies, adjusting for variations across months and between vendors. This will allow us to quantify the direct and interactive effects of time (month) and vendor on the current prices of organic products, providing insights into the responsiveness of organic food prices to market dynamics and vendor-specific factors.

The findings reveal distinct pricing patterns that differ by vendor and demonstrate notable fluctuations corresponding to seasonal trends. These insights are crucial for consumers making informed purchasing decisions and for vendors strategizing to capture or expand market share in the competitive organic food sector. The importance of understanding these pricing

mechanisms lies in its potential to influence marketing strategies, consumer spending, and ultimately, the economic sustainability of organic farming practices.

The paper is structured to provide a clear, logical flow from methodology to implications Section 2 outlines our dataset selection and cleaning, ensuring transparency. The **?@sec-model** details the Bayesian model and its rationale, directly linked to our data preparation. The **?@sec-result** displays our findings with visual aids, while the **?@sec-discussion** interprets these findings in the context of broader market dynamics. Finally, **?@sec-limitation** highlights areas for further study, acknowledging any constraints encountered.

2 Data

2.1 Overview

All data analysis was conducted using R R Core Team (2023a), including statistical computing and graphics. And the following R packages were used for conduct code in scripts: tidyverse Wickham et al. (2019), Palmerpenguins Horst, Hill, and Gorman (2020), ggplot2 Wickham (2016), Dplyr Wickham et al. (2023), Knitr Xie (2023), modelsummary Arel-Bundock (2022), arrow Richardson et al. (2024), here Müller (2020), rstanarm Goodrich et al. (2023), sclaes Wickham, Pedersen, and Seidel (2023), splines R Core Team (2023b),broom Robinson, Hayes, and Couch (2023), lubridate Grolemund and Wickham (2011), janitor Firke (2023) and testthat Wickham (2011). Following Alexander (2023a) we conducted data simulate, test simulated data, data cleanning, test analysis data, EDA, data modelling. The R code in scripts were adapted from Alexander (2023b)

The dataset under analysis was sourced from Jacob Filipp's Hammer Project Filipp (2023), which compiles extensive pricing data on organic food products sold across various retail vendors. This dataset is particularly valuable for its comprehensive coverage of both temporal and vendor-specific pricing strategies, offering a unique lens through which seasonal trends and competitive behaviors in the organic food market can be observed. The choice of this dataset was motivated by its granularity and the quality of its data, which includes multiple vendors and a wide temporal range, making it ideal for a nuanced analysis of pricing dynamics.

2.2 Measurement

The dataset used in this study comes from Jacob Filipp's Hammer Project, which provides detailed records of organic food pricing across retailers and months. This data collection effort captures an important economic and social phenomenon: fluctuating pricing strategies in the competitive organic food market. Prices are recorded monthly to reflect the dynamic nature of these market conditions, providing a comprehensive snapshot that facilitates analysis of trends and pricing behavior over time.

The primary step in capturing the phenomenon involves systematically collecting price data directly from retail sources. This data is gathered from grocery merchants. Each price tag observed represents a specific decision by a vendor regarding how much to charge for an organic product at a given time. So the important variables: current price which is the price of an organic product at the time of data collection. It is observed directly from the vendor. And old price which is the previously recorded price for the same product, which provides a historical comparison point. It helps in understanding how prices are adjusted over time.

Each recorded price becomes an entry in the dataset, associated with metadata that includes the vendor, the time of collection, and the product details. This structured approach ensures that each entry in the dataset is a reliable and accurate representation of the real-world pricing phenomena observed. These entries are then ready for analytical processes that can derive meaningful insights about organic food pricing dynamics.

2.3 Data Cleanning

The analysis begins with the consolidation of raw data from two distinct sources: product transaction records and product details. These datasets were sourced from the Hammer Project, containing over 12 million rows of diverse product data, including non-organic items and a variety of vendors beyond the scope of this study. The transaction records (raw_data) include timestamps, vendor details, and product identifiers, while the product details (product_data) provide comprehensive information such as product names, brands, pricing history, and unit measurements.

To conduct analysis, the two datasets were merged using an inner join on the product_id, which serves as a unique identifier linking transaction records to specific product details. This joining process ensured that each transaction record was enriched with relevant product information, creating a comprehensive view of each sale's context.

Post-merger, the dataset still contained an extensive range of products and vendors, many of which were irrelevant to the focus of this study—organic food pricing dynamics among major vendors. To refine the dataset first did vendor selection. The dataset was filtered to include only transactions from major retailers known for their significant market presence and influence in organic food sales, specifically "Metro," "NoFrills," "Walmart," "Voila," "TandT," and "Galleria." This selection was based on the objective to analyze pricing strategies within the most competitive segments of the retail market. A further filtration was applied to isolate organic products. This was achieved by searching for the keyword "organic" within the product_name field, using a case-insensitive search to ensure comprehensive inclusion of all relevant items. This step significantly reduced the dataset size and increased its relevance to the study's aims.

The current_price and old_price were parsed to ensure that they were represented as numerical values, free from any non-numeric characters that could interfere with quantitative analysis. This parsing was crucial for accurate statistical calculations and comparisons. The nowtime

field, representing the timestamp of each transaction, was used to extract the month component. This transformation was essential for analyzing seasonal trends and month-to-month pricing variations. Lastly, entries with missing or zero values in current_price were removed to maintain the integrity of pricing analysis. Additionally, any remaining rows with incomplete data were dropped, ensuring that the dataset used in subsequent analyses was entirely robust and clean.

2.4 Outcome variables

2.4.1 Predictor variables

The old price variable, representing the product's price from the preceding month, serves as a critical metric for understanding the evolution of market trends and pricing strategies over time. The summary Table 1 reveal a broad range of "Old Price" from a minimum of $1.69 to a maximum of $38.56, with a median price of $5.49, suggesting a median market position. The mean, slightly higher at $7.66, indicates a positive skew in the data, with the first quartile at $3.99 and the third at $10.49, reflecting substantial variability in historical pricing. This variability underscores the diverse pricing strategies across different vendors and product types, and provides a foundation for analyzing how prices adjust in response to market dynamics, consumer demand, and possibly supply chain fluctuations.

Table 1: Summary of old price

Statistic	Value
Min	1.69000
1st Qu.	3.99000
Median	5.49000
Mean	7.65968
3rd Qu.	10.49000
Max	38.56000

The vendor variable acts as a categorical identifier for retailers within the dataset, providing crucial insights into the diversity of pricing strategies employed across different retail environments. This variable is essential for comparing and contrasting market behaviours as it delineates the dataset into distinct groups based on the retailer. The distribution of data Table 2 shows that vendors varies significantly, with Metro being the most frequently occurring vendor with 1,981 entries, followed by Voila with 1,513 entries. Walmart, Galleria, TandT, and NoFrills appear with decreasing frequency, suggesting a wide range in the scale and possibly the geographic reach of these retailers. This spread indicates that Metro and Voila dominate this dataset, which could imply a greater influence or market share in the organic products sector compared to the other vendors. Understanding the pricing strategies of these vendors

is key to analyzing market trends, as each vendor may react differently to market pressures, consumer preferences, and seasonal changes. Analyzing this variable can reveal patterns that are specific to certain types of retailers and provide valuable insights into the competitive landscape of the organic food market.

Table 2: Preview of Breakdown of Organic Product Listings by Vendor

Vendor	Frequency
Metro	1981
Voila	1513
Walmart	588
Galleria	173
TandT	15
NoFrills	5

The "Month" variable in the dataset, recorded as a numerical value, serves as a crucial indicator for analyzing seasonal impacts on organic food pricing. It reflects when the data was collected, revealing essential patterns of how prices fluctuate through different months, likely influenced by consumer demand, supply variations, and promotional periods. The distribution of observations shown in Table 3, the peaks are in summer season suggests significant market activity during the summer months, potentially driven by an increased availability of fresh produce or higher consumer buying rates. In contrast, the notably lower entries in spring and winter indicate reduced activity, which could be due to offseason market slowness or less frequent data collection. This variable is integral to understanding how temporal factors influence pricing strategies, providing valuable insights for vendors, consumers, and analysts studying market dynamics in the organic food sector.

Table 3: Preview of Breakdown of Organic Product Listings by Month

Month	Count
2	4
3	44
4	31
5	8
6	999
7	1253
8	804
9	691
10	353
11	88

2.4.2 Response Variable

The response variable current price represents the most recent price at which organic food products were sold across various vendors. The summary statistics Table 4 for this variable provide a snapshot of its distribution within the dataset. The minimum price recorded is 0.97, indicating a lower-end pricing point for some products, while the maximum price peaks at 29.99, reflecting premium offerings within the organic sector. The first quartile is 3.22, suggesting that at least 25% of the products are priced at or below this level, which could be indicative of basic or smaller-sized organic goods. The median price of 4.49 and the mean of 5.69 suggest a central tendency towards moderately priced items, with the mean being slightly higher due to the influence of more expensive products pulling the average up. The third quartile at 7.49 shows that 75% of the products fall below this price point, highlighting a pricing structure that might cater more to a mid-range consumer segment. This comprehensive look at current price helps in understanding the economic landscape of organic food pricing and its variability, which is crucial for analyzing market strategies and consumer behaviour.

Table 4: Preview of Summary of Current Price

Statistic	Value
Min	0.970000
1st Qu.	3.220000
Median	4.490000
Mean	5.695207
3rd Qu.	7.490000
Max	29.990000

3 Model

3.1 Model Overview

In our study, we are particularly interested in understanding how the current price of organic products is influenced by several key factors: the price of the product in the previous month (old price), the vendor from which the product is purchased, and the time of year (month). To systematically examine these relationships, we developed a Bayesian regression model. This model incorporates these variables to elucidate their impacts on current pricing strategies. By leveraging historical pricing data alongside vendor-specific information and temporal dynamics, our model aims to uncover meaningful patterns that can inform both theoretical understanding and practical applications in market pricing strategies.

3.2 Model set-up

$$\text{Current Price}_i = \beta_0 + \beta_1 \times \text{Month}_i + \beta_2 \times \text{Old Price}_i + \sum(\beta_{\text{vendor}} \times \text{Vendor}_i) + \epsilon_i$$

Where:

- Current Price_i is the dependent variable representing the price of organic products at time i.
- Month_i is a continuous independent variable reflecting the month of the year to account for seasonal effects on pricing.
- Old Price_i is a continuous independent variable representing the price of the product in the previous month, allowing for analysis of price evolution.
- Vendor_i is a categorical variable with levels for each vendor included in the study, capturing vendor-specific pricing strategies.
- β_0 is the intercept, $\beta_0 \sim \text{Normal}(0, 2.5)$.
- β_1, the coefficient for Month, $\beta_1 \sim \text{Normal}(0, 2.5)$.
- β_2, the coefficient for Old Price, $\beta_2 \sim \text{Normal}(0, 2.5)$.
- β_{vendor}, coefficients for each vendor level, $\beta_{\text{vendor}} \sim \text{Normal}(0, 2.5)$ for each vendor level.
- ϵ_i is the error term, assumed to be normally distributed.

We conducted the Bayesian analysis in R (R Core Team 2023a), utilizing the `rstanarm` package (Goodrich et al. 2023). To maintain consistency and reliability in our modeling approach, we employed the default priors provided by `rstanarm`. These priors are well-suited for a wide range of data types and ensure robustness in the inference process, particularly in complex models such as ours that involve multiple predictors and hierarchical structures.

Normal priors with a mean of 0 and a standard deviation of 2.5 are assigned to the coefficients and the intercept. These priors were chosen to impose a slight regularization effect, reducing the risk of overfitting by moderating the influence of extreme values or outliers in the data.

3.3 Model justification

The Bayesian linear regression model was selected for this analysis to explore how various factors such as month, old price, and vendor influence the current price of organic products. This model type was chosen due to its flexibility in incorporating prior knowledge and handling uncertainty in estimates, which is particularly beneficial in markets where data may exhibit variability and non-standard distributions. Bayesian methods allow for the integration of prior beliefs or empirical evidence into the analysis, which can be particularly useful when dealing with organic product pricing where past market trends can inform current expectations. The

primary assumptions include the linearity between predictors and the outcome, independent and identically distributed residuals, and normally distributed error terms. These assumptions are typical for linear regression but must be validated through diagnostic checks to confirm no significant deviations occur.

Unlike frequentist regression, Bayesian regression provides a probabilistic approach to inference, offering a full posterior distribution of the parameters, which helps in understanding the uncertainty around the estimated effects. This model is robust to various types of data and does not strictly require the normality assumption for residuals that a classical linear regression would. This is particularly beneficial given the often skewed or heterogeneous nature of pricing data. Each predictor's influence is quantified by respective coefficients, adjusted for vendor-specific variations and temporal trends. The use of normal priors for these coefficients is justified by the need to regularize estimates, thus preventing overfitting and ensuring stable predictions even with potentially collinear or sparse data.

3.3.1 Comparison with Alternative Models

Although a standard multiple linear regression could model continuous outcomes effectively, it lacks the ability to incorporate prior distributions and assess uncertainty comprehensively. odels such as logistic regression or Poisson regression were considered; however, these models are more suited to categorical or count data, respectively. Given that our outcome variable, current price is continuous, a linear approach is more appropriate. These were also evaluated, especially for capturing any non-linear dynamics between the predictors and the current price. While potentially offering a closer fit to certain complex patterns, the increased model complexity could hinder interpretation and require more extensive data to validate effectively.

3.4 Model Vialidation

Model diagnostics included checks for multicollinearity, heteroscedasticity, and normality of residuals. Posterior predictive checks were performed to ensure the model adequately captures the observed data patterns. These steps help in verifying model assumptions and the appropriateness of the model for the given data. The background details and diagnostics are included in Appendix B.

4 Results

Our results are summarized in Table 6.

Table 5: Summary of Bayesian Model

	Bayesian_model
(Intercept)	3.53
	(0.22)
month	−0.42
	(0.02)
old_price	0.62
	(0.01)
vendorMetro	0.35
	(0.17)
vendorNoFrills	0.09
	(0.97)
vendorTandT	−0.02
	(0.57)
vendorVoila	1.17
	(0.17)
vendorWalmart	0.19
	(0.19)
Num.Obs.	4275
R2	0.703
R2 Adj.	0.702
Log.Lik.	−9245.456
ELPD	−9253.9
ELPD s.e.	94.2
LOOIC	18 507.7
LOOIC s.e.	188.4
WAIC	18 507.7
RMSE	2.10

Table 6: Explanatory models of flight time based on wing width and wing length

	Bayesian_model
(Intercept)	3.53
	(0.22)
month	−0.42
	(0.02)
old_price	0.62
	(0.01)
vendorMetro	0.35
	(0.17)
vendorNoFrills	0.09
	(0.97)
vendorTandT	−0.02
	(0.57)
vendorVoila	1.17
	(0.17)
vendorWalmart	0.19
	(0.19)
Num.Obs.	4275
R2	0.703
R2 Adj.	0.702
Log.Lik.	−9245.456
ELPD	−9253.9
ELPD s.e.	94.2
LOOIC	18 507.7
LOOIC s.e.	188.4
WAIC	18 507.7
RMSE	2.10

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

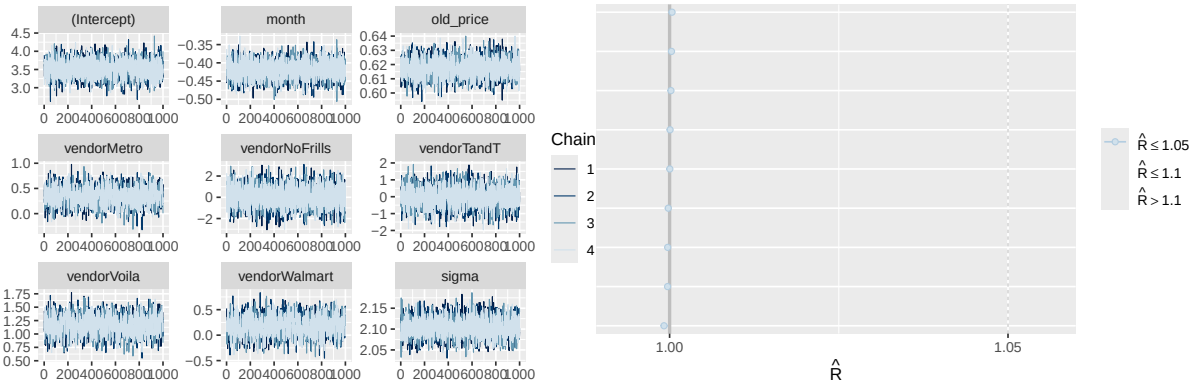
In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

Figure 1a is a trace plot. It shows... This suggests...

Figure 1b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 1: Checking the convergence of the MCMC algorithm

References

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

———. 2023b. *Telling Stories with Data: With Applications in r*. Chapman; Hall/CRC.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Filipp, Jacob. 2023. *Hammer Project: Comprehensive Dataset on Organic Food Pricing*. https://jacobfilipp.com/hammer/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Samuel Brilleman. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

———. 2023b. *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://CRAN.R-project.org/package=arrow.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 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://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.