My title*

My subtitle if needed

Cher Ning-Li

December 3, 2024

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

1 Introduction

#TODO: Overview paragraph

The estimand that is targetted within this paper's analysis is the price of a dozen eggs across the 8 top Canadian grocers.

#TODO: Results paragraph

#TODO: Why it matters paragraph

#TODO: Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

2 Data

2.1 Overview

The dataset used for this paper's analysis is titled "Canadian Grocery Price Data" and was obtained from Project Hammer (Filipp 2024), which compiles grocery store price listings across the 8 major Canadian Grocers — Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods. The dataset contains data starting from February 28, 2024 and is updated regularly with the addition of new entries. Only the entries up until November 25, 2024 are included within this paper's analysis.

*Code and data are available at: (https://github.com/cher-ning/cndgrocers)[https://github.com/cher-ning/cndgrocers].

The dataset was downloaded as two separate csv files titled `hammer-4-product.csv` and `hammer-4-raw.csv`. The `product` table includes information about specific products that are available across the different grocers, with unique product ID numbers to identify each listing at each grocer. The `raw` table includes the price and more time-specific information regarding each product's listing. The two tables were joined by matching the product IDs to create the full dataset, a comprehensive table where each entry holds all the necessary product and listing information.

Under the guidance of Alexander (2023), the R programming language (R Core Team 2023) was used for analysis of this dataset. The package `tidyverse` (Wickham et al. 2019) was used to simulate data and test the simulated data before analysis. Packages `tidyverse` (Wickham et al. 2019) and `arrow` (Richardson et al. 2024) were utilized to clean the full raw dataset, which were then also used alongside packages `testthat` (Wickham 2011) and `here` (Müller 2020) to test the cleaned dataset. Lastly, packages `tidyverse` (Wickham et al. 2019), `rstanarm` (Goodrich et al. 2022), `arrow` (Richardson et al. 2024), and `here` (Müller 2020) were used to build predictive models using the cleaned dataset.

2.2 Measurement

The observations in the dataset are scraped from the grocer's official website's interface, which means it only contains the information that is publically listed and available. The recorded prices for each item are the listed price for the "in store pick up" option, with the target pickup neighbourhood being a neighbourhood in Toronto. Certain values and listings may be missed, since the internal APIs that power the grocer's websites are not accessed and specific extracts may error for certain vendors on particular days for unforeseeable reasons. Starting from July 11, the targetted variety of grocery items for data collection was greatly increased, meaning that there may be products which are missing pricing information before then.

2.3 Data-Cleaning

After combining the two tables, `raw` and `product`, by matching each entry's product ID, the full dataset with $12,753,964$ entries was created. Then, entries with "egg" or "eggs", and "dozen" or "12" appearing in their product name or units were selected for. All February entries were omitted from analysis due to data collection starting on 28 February, 2024, which means there is a lack of sufficient data for that month to make any reliable conclusions from. The relevant variables of interest to us are: `nowtime`, `vendor`, `current_price`, `old_price`, and `product_name`, as explained below:

- **nowtime**: Date and time of when the data was collected
- **vendor**: One of the 8 Canadian grocery vendors
- **current_price**: Price at time of extract
- **old_price**: An 'old' struck-out price, indicating the item's regular pre-sale price

- **product_name**: Product name, may include brand and/or units

For ease of analysis, some additional variables were derived.

- **month**: The month of when a data entry was collected; extracted from `nowtime`
- **prev_month_avg**: The average price of a dozen eggs across all vendors during the previous month

All entries with empty values in any of these key variables were dropped. It is important to note that this would filter out any unannounced price changes, as only price decreases for an explicitly advertised sale event would include both `current_price` and `old_price` variables. The first 5 rows from the cleaned dataset along with these derived variables is shown in Figure 1.

current_price	old_price	vendor	product_name	month	prev_month_avg
4.99	5.49	Loblaws	Naturegg Nest Laid White Eggs, Large	May	6.018973
7.49	7.99	Loblaws	Naturegg Omega Plus Solar Free Range Eggs, Large	May	6.018973
4.28	5.99	TandT	Naturegg Large Omega-3 Brown Eggs (12s)	May	6.018973
5.99	6.89	Metro	Large White Free Run Omega-3 Eggs	May	6.018973
4.99	5.49	Loblaws	Naturegg Nest Laid White Eggs, Large	May	6.018973

Figure 1: Sample of Analysis Data

2.4 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Talk way more about it.

2.5 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

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. - maybe should try to specify the prev_month_avg down to also selection by vendor

Appendix

A Additional data details

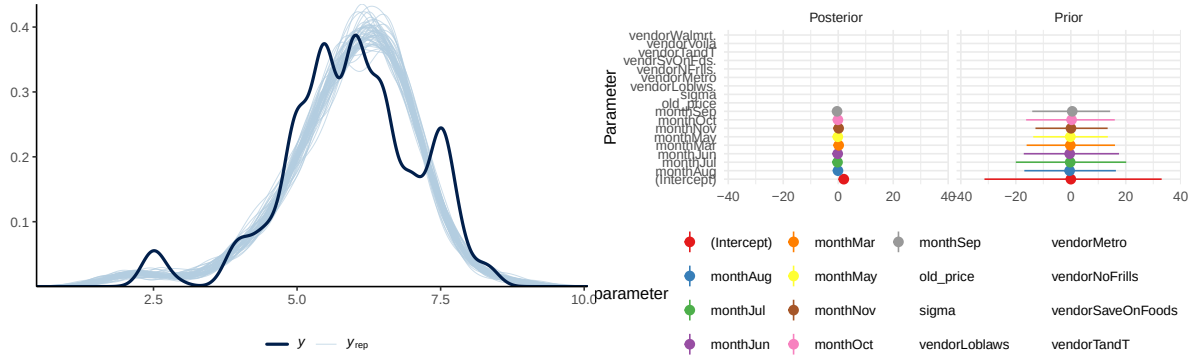
!!! ADD FIRST 5 ROWS DATA HERE

B Model details

B.1 Posterior predictive check

In Figure 2a we implement a posterior predictive check. This shows...

In Figure 2b we compare the posterior with the prior. This shows...



(a) Posterior prediction check

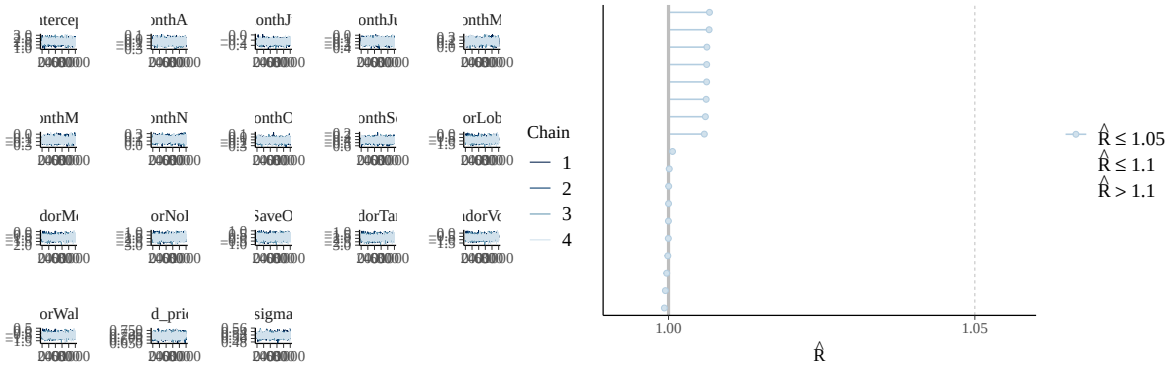
(b) Comparing the posterior with the prior

Figure 2: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 3a is a trace plot. It shows... This suggests...

Figure 3b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 3: Checking the convergence of the MCMC algorithm

References

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

Filipp, Jacob. 2024. *Canadian Grocery Price Data*. https://jacobfilipp.com/hammer/#FAQ.

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

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

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://CRAN.R-project.org/package=arrow.

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.

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.