Statistical Advice on the Effect of Interventions on Beverage Sales

Parham Pishrobat (71097927), Johnson Chen (85784080), Sarah Masri (97415681)

March 29, 2024

Contents

1 Introduction

Concerns around sugar consumption and its health implications have prompted many interventions to encourage consumers to cut down on sugary beverages. The current study investigates the effectiveness of two types of intervention strategies to motivate and incentivize consumers to choose zero-calorie beverages over sugary alternatives. In particular, the research question focuses on the impact of two strategies to inform consumers about calorie content through visual presentations: labels highlighting either the calorie content or the physical activity required to burn calories. Furthermore, the effectiveness of price discounts on behaviour is explored, both independently and in conjunction with explanatory messaging. This analysis aim to determine the suitable model to examine the intervention effects, perform the analysis and identify the significant effects impacting immediate and long-term preferences of the customers, adjust for differences across data collection sites, and validate the underlying assumptions of the chosen model.

2 Data Description and Summaries

The study adopts an interrupted time-series multi-site quasi-experimental design to assess the effectiveness of the five interventions on the purchase patterns of bottled sugary and zero-calorie beverages. The data are recorded from cafeterias and convenience shops at three hospital sites, denoted by A, B, and C.

The experiment, starting on October 27, 2009 and ending 32 weeks later, measure daily sales of bottled sugary and zero-calorie beverages. The study period included a baseline data collection phase, intervention phases to elicit behaviour change, and washout periods to assess the persistence of intervention effects. The dataset contains 631 sales counts for variables zero-calorie, sugary, and all drinks, sale location, day of the week, and the type of intervention applied. The price interventions consist of two periods of 10% discount on zero-calorie beverages, with one phase providing additional explanatory messaging about the discount. The calorie messaging interventions provided information on the caloric content of sugary drinks, the physical activity required to burn off these calories, and a combination of both strategies.

The day of week, site and intervention covariates are each considered categorical data types. Other observations are classified under numerical data types as they measure sales counts. The total sales information is kept in the study to represent the overall patterns of beverage sales irrespective of their sugar content. Missing data is observed over some control periods of the study.

3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an important first step in data analysis in uncovering underlying patterns, relationships, and outliers in the data. To explore the data of the study, time series plots are included in the main report. Additionally, Appendix A provides supplementary visualizations encompassing missing values, histograms. These visualizations provide insight into the distribution of sugary and zero-calorie beverage sales, as well as their temporal trends.

Figure 1 represents the sales time series of zero-calorie and sugary beverages across different sites, colored by the Intervention. Both sugary drink and zero calorie drink sales trends are different across different interventions, indicating that different interventions do have different impacts on sales. The sales trends are different across sites, and the effect of intervention appears to be different across sites as well.

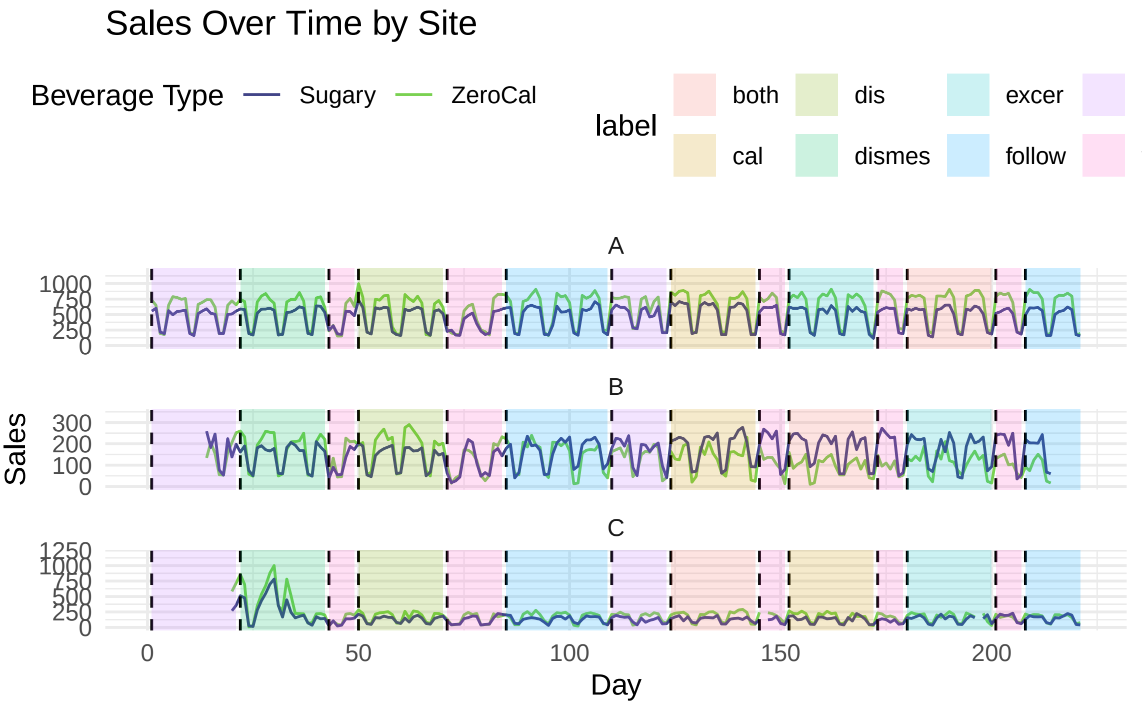

Figure 1: *Time Series plot illustrating the daily sales volumes of zero-calorie (in green) and sugary (in blue) beverages across three hospital sites over 30 weeks. The dashed and their corresponding shorthand labels mark the interventions to allow for visual assessment of their impact on beverage sales trends.*

A boxplot is a method for graphically demonstrating the property of statistical distribution of numerical data. Figure 2 in the appendix shows the distribution of sugary and zero-calorie beverage sales across different intervention strategies. Each boxplot captures the sales data variability with the central line denoting the median, the edges of the box indicating the interquartile range, and the whiskers extending to the furthest points that are not considered outliers. Note that bimodality of sales exists in the data. Specifically, there are considerably more beverage sales on weekdays than on the weekend. To better represent the data, the weekday and weekend sales are separated into two plots. It is observable that the maximum and median number of sales during the week varies with intervention.

3.1 Missing Values and Data Imbalance

The dataset contains some missing values, as obseved in Figure 5. It is important to identify what kind of missing data exists within a dataset to better understand how to handle missingness during formal analysis. There are 9 days of unrecorded zero-calorie and sugary beverage sales, 7 of which represent the last week of the study at site B. This report removed the last week of study at site B, treating it as not recorded entirely. The rest of missing values are considered Missing At Random (MAR) due to the small amount.

It is also important to note whether of not the dataset reflects a roughly equal number of observations between the sites and interventions respectively. If one class in either the site or interventions are represented disproportionately, then models used may become biased towards the most frequently seen class. Balance between sites appears to be reasonable, where some imbalance is present between interventions. Namely, some of the no-intervention levels are imbalanced when compared to the other interventions, this does not have a significant impact on the analysis of the interventions.

4 Formal Analysis

This study employs Interrupted Time Series Analysis (ITSA) with Linear Mixed Models (LMMs) to assess the impact of labelling interventions on beverage selection. ITSA is a statistical technique to characterize the temporal changes before and after interventions. This approach is especially pertinent to the current analysis as the study employs an interrupted time-series multi-site quasi-experimental design. Although ITSA can be carried out using a variety of linear models, LMMs are preferred to allow for the inclusion a random effect to accommodate for variability across collection sites. LMMs manage complex data, handle missing values, and capture both the fixed effects of interventions and the random effect of variability across sites. The logarithm of the ratio between zero-calorie and sugary beverage sales is considered as outcome of interest because it symmetrically quantifies relative preference shifts, where a higher log ratio indicates a preference for healthier options and vice versa. Correlation structure is defined based on the temporal spacing between observations, reflecting the notion that as time between sales data points increases, their correlation decreases in a manner that may not follow a simple pattern.

Upon initial analysis, which focuses solely on the direct before-and-after impacts of interventions, the study does not identify any statistically significant intervention effects. This outcome highlights the inadequacy of the method in capturing the dynamic and evolving nature of consumer responses over time. Subsequently, the analysis is refined to incorporate interactions between interventions and time, providing a richer understanding of how consumer preferences for beverages evolve in response to interventions. This nuanced approach reveals that exercise-focused messaging, in particular, has a significant and increasingly positive influence on consumer preferences towards healthier beverage options over time. Furthermore, while discount strategies initially appear effective, their impacts do not sustain over time. Moreover, the analysis points to the significance of the day of the week, with consumer behaviour exhibiting notable shifts during weekends, suggesting either variations in beverage preferences between weekdays and weekends or change in the population during weekends.

These insights draw attention to the challenges of influencing dietary choices and the variable success rates of different intervention strategies. Specifically, the effectiveness of visually linking calorie consumption with physical activity emerged as a significant factor in promoting zero-calorie beverage choices. This variation in intervention effectiveness, coupled with the importance of intervention timing demonstrate the strategic value of timing in intervention planning. While the methodology employed provides valuable perspectives on intervention impacts, it operates under certain assumptions, such as linear temporal changes and uniform intervention effects across sites, which may simplify the complex dynamics of actual consumer behaviour. This complexity points to the challenges involved in designing universally effective health promotion interventions, underscoring the need for targeted, contextually aware strategies to encourage healthier consumer choices effectively.

5 Conclusions

In conclusion, our analysis demonstrates the value of integrating ITSA with LMM for evaluating public health interventions in shifting consumer behaviour. While our findings highlight the potential of certain interventions to positively influence consumer behaviour, they also point to the importance of considering temporal patterns in the design and implementation of health promotion strategies. The refined analysis, which incorporates interactions between interventions and time, reveal the progressive influence of exercise-related messaging of the calorie content on shifting consumer preferences towards zero-calorie beverages. Furthermore, the discounting strategies show a significant impact after commencing the intervention but fail to sustain the impact. Variations in preference between weekdays and weekends further highlighted the importance of timing and pointing to different target populations.

6 References

Columbia University Mailman School of Public Health. (n.d.). Repeated Measures Analysis. Columbia University Mailman School of Public Health. https://www.publichealth.columbia.edu/research/population-health-methods/repeated-measures-analysis

UCLA Statistical Consulting Group. (n.d.). Introduction to Linear Mixed Models. Retrieved March 1, 2024, from https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/

University of Virginia Library. (n.d.). Getting Started with Generalized Estimating Equations. Retrieved March 1, 2024, from https://library.virginia.edu/data/articles/getting-started-with-generalized-estimating-equations

A Appendix : Supplementary Visualizations

The following sub section contains additional figures and their analysis results.

A.1 Boxplot

The Figure 2 hows the sales distribution data from weekdays and weekends across sites.

A.2 Histogram Plots

The Figure 3 (histogram plots) show the frequency distribution of sales for Sugary (in purple), Zero-Calorie (ZeroCal, in teal), and Total (in yellow) beverages. The x-axis of each histogram represents the sales volume, while the y-axis indicates the count of observations within each sales range. The pattern in all histograms is similar: most sales numbers cluster at the lower end of the scale, suggesting a higher frequency of days with fewer sales; however, the sales histograms exhibit a second weaker mode, indicating two common sales volumes across the observed period.

A.3 Scatter Plot

The figure 4 (scatter plot) depicts the relationship between zero-calorie and sugary beverage sales at three different hospital sites: A or chop (purple), B or HF (blue), and C or NS (yellow). The x-axis represents zero-calorie beverage sales, and the y-axis represents sugary beverage sales. A dashed line, suggesting the line of equality, indicates where the sales for both types would be equal. Points above the line indicate higher sugary beverage sales when compared to zero-calorie ones, and points below the line indicate the opposite. The clustering of points towards the upper right suggests that for higher sales volumes, sugary beverages tend to sell as much as or more than zero-calorie options, particularly in site A (chop). The plot reveals variability in the sales patterns across sites, with the HF site having a more direct correlation between ZeroCal and Sugary sales when compared to other sites.

A.4 Missing Values

The figure 5 visualizes the missing values.

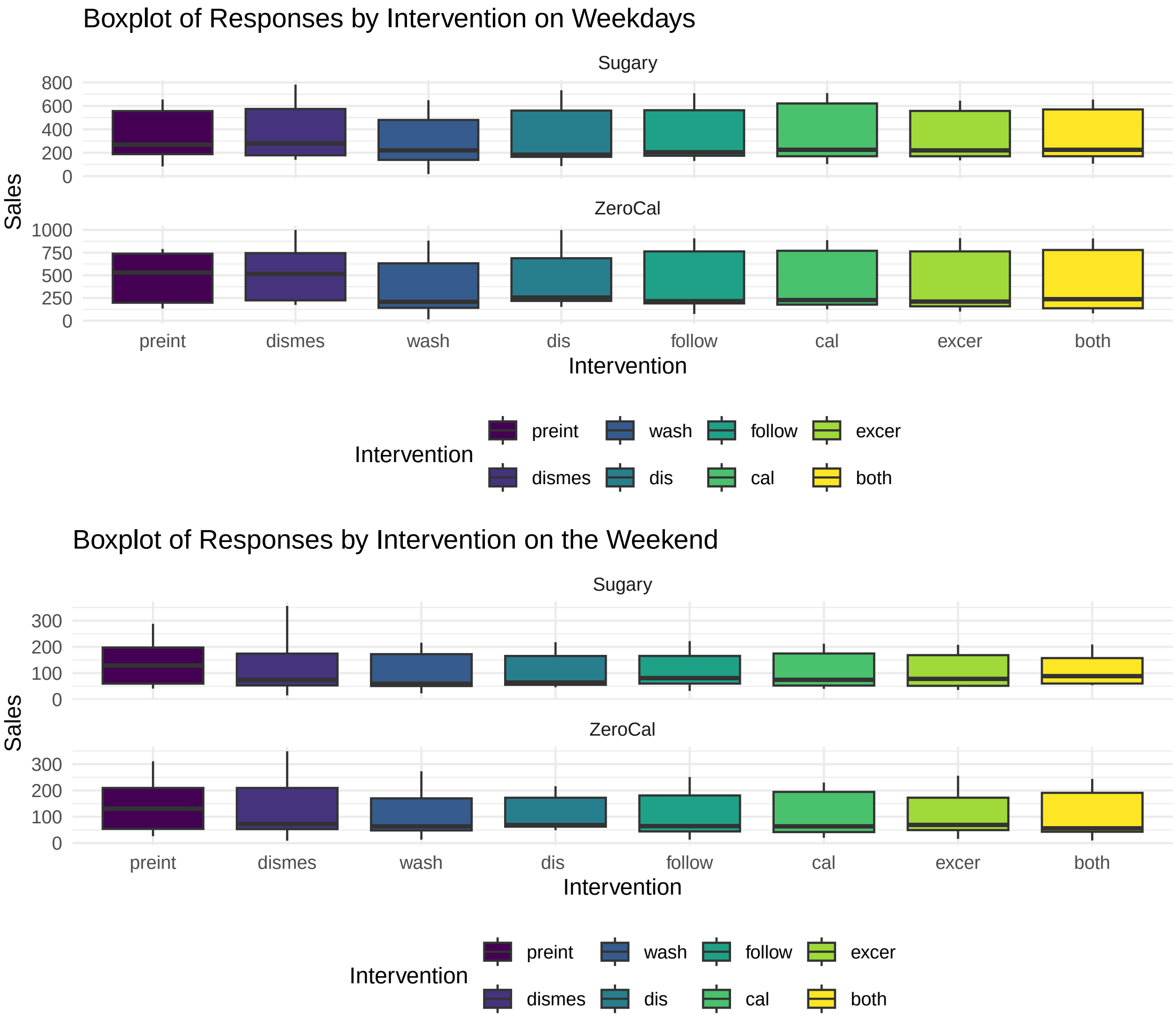

Figure 2: *The boxplots display the sales distribution of Sugary and Zero-Calorie beverages across various interventions. Each intervention is colour-coded and shows the range of sales data with the central line representing the median sales. Note that the plots only illustrates weekday sales.*

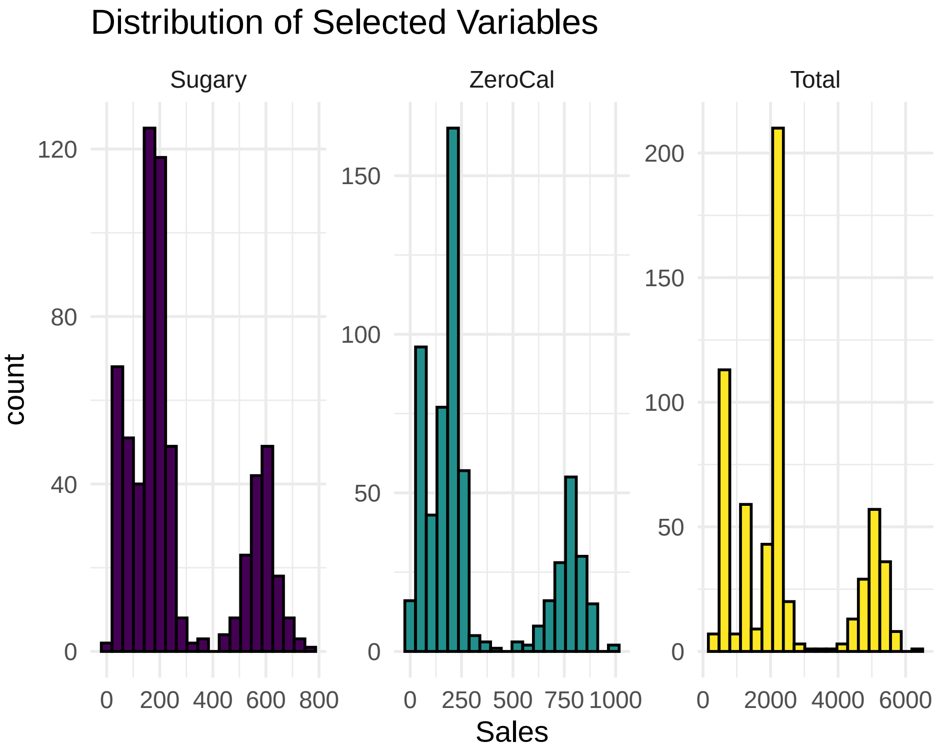

Figure 3: *Sales Distribution Analysis: Histograms displaying the frequency of sales for Sugary (purple), Zero-Calorie (teal), and Total combined (yellow) beverages. Each histogram reveals the distribution pattern of sales volumes, highlighting the bimodal nature of sales across all types.*

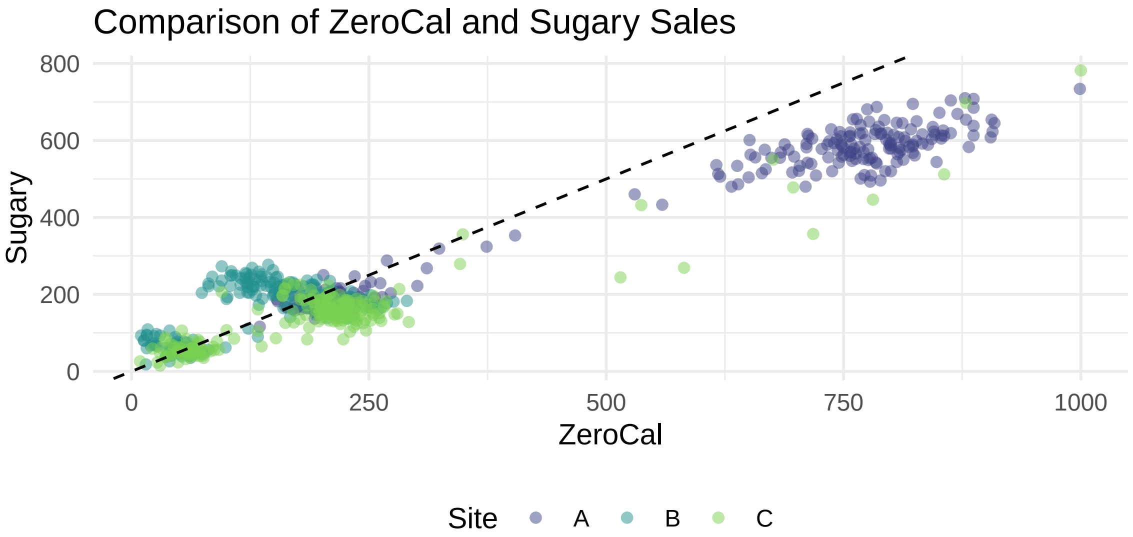

Figure 4: *This scatter plot contrasts zero-calorie and sugary beverage sales, colour-coded by the site. Each point represents the paired sales data for a given day, with the site-specific colour coding (chop in purple, HF in blue, NS in yellow) illustrating the sales trend at each location. The dashed diagonal line marks the parity where the sales of both beverage types are equal. Deviations from this line highlight the predominance of one beverage type over the other in daily sales.*

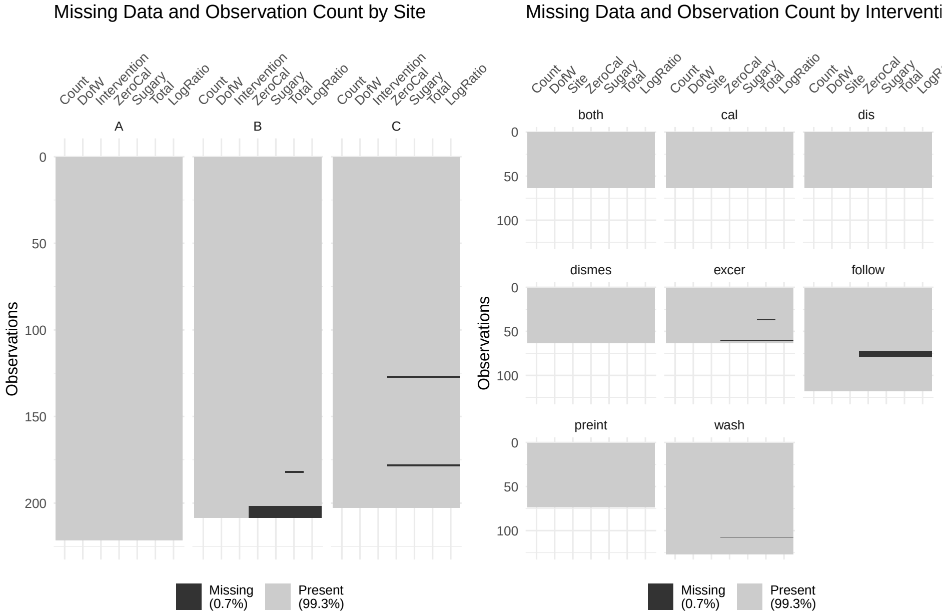

Figure 5: *This plot provides insight into the frequency of missingness within the dataset. Black indicates missing data. Additionally it shows the quanitity of data available by site and by intervention.*

B Appendix : Technical Analysis

B.1 Mathematical Formulation

The core of our analysis utilizes a combined approach of Interrupted Time Series Analysis (ITSA) and Linear Mixed Models (LMMs) to assess the effects of health interventions on the choice between zero-calorie and sugary beverages. This approach enables the examination of both the immediate and the longitudinal impacts of the interventions across different sites. The mathematical model integrating ITSA principles within an LMM framework is formulated as follows:

$$LogRatio_{st} = \beta_0 + \beta_{int} \times INT_{st} + \beta_{int:time} \times INT_{st} \times Time_s + \beta_{dow} \times DOW_{st} + b_s \times SITE_t + \epsilon_{st}$$

where model components are defined based on the Table 1. Note that s and t indices represent the time and site.

This formulation allows for a nuanced analysis that captures the dynamic effects of interventions over time, adjusted for site-specific variations and weekly sales patterns. By employing this ITSA within an LMM framework, the model adeptly handles the hierarchical data structure and accounts for both fixed and random effects.

B.2 Implementation

In our study, we opted for the `nlme` package in R for its comprehensive support in fitting Linear Mixed Models (LMMs), which allows for the inclusion of both fixed and random effects, essential for the analysis of complex data structures like ours. The `lme` function from this package was specifically chosen due to its flexibility in specifying a wide range of variance and correlation structures, critical for accurately modeling the intricate dynamics of our data. By preprocessing

Table 1: *(#tab:table 1)Variable Description*

Term	Description
$LogRatio_{st}$	Log ratio of zero-calorie to sugary beverage sales
$Time_s$	Days since the start of the intervention at site s
$Site_t$	Indicator function of site at time t
INT_{st}	Indicator function of intervention type at time t and site s
DOW_{st}	Indicator function of day of week at time t and site s
β_0	Intercept, representing the baseline log ratio
β_{int}	Coefficients of fixed effect of the intervention
$\beta_{int:time}$	Coefficients of interaction effect between interventions and time
β_{dow}	Coefficient of fixed effect of the day of the week
b_s	Random effect for site s
ϵ_{st}	Error term for the model at time t and site s

the data to omit missing values, we ensure the robustness of our model against potential biases that missing data could introduce. The decision to condense the day of the week into two categories, 'Weekday' and 'Weekend', and to simplify the intervention levels by grouping non-intervention periods under a single category 'noint', was driven by the desire to streamline the analysis and enhance interpretability without sacrificing the granularity necessary for meaningful insights.

```r
library(nlme)
clean_data <- na.omit(beverage_sales)
# Recode DofW to two levels: Weekday and Weekend
clean_data$DofW <- factor(ifelse(clean_data$DofW %in% c(1, 2, 3, 4),
                                 'Weekday',
                                 'Weekend'))
# Recode Interventions to mark all no interventiopn periods as "noint"
levels(clean_data$Intervention)[levels(clean_data$Intervention) %in%
                                  c("preint", "follow", "wash", "wash2")] <- "noint"

# Adding a variance structure to the model
model <- lme(LogRatio ~ Intervention + Intervention:Count + DofW ,
             data        = clean_data,
             random      = ~ 1 | Site,
             weights     = varPower(form = ~ I(1/ZeroCal) + Sugary),
             correlation = corRatio(form = ~ Count | Site),
             method      = "ML")

# summary(model)
```

The choice of the logarithm of the ratio of zero-calorie to sugary beverage sales as the response variable was motivated by its ability to symmetrically quantify shifts in consumer preferences, providing a balanced measure that reflects relative changes in sales rather than absolute values, which is particularly insightful for evaluating the effectiveness of health interventions. The incorporation of a power variance function with the sales data as covariates addresses the potential issue of heteroscedasticity, ensuring that the model adequately reflects the variability in the data. The rational

__Table 2:__ Significant Model Coefficients and Their P-Values

term	estimate	p.value
Interventiondis	0.880	0.044
Interventiondismes	0.875	0.014
Interventionexcer	1.792	0.001
Interventionnoint	0.719	0.032
DofWWeekend	-0.087	0.000
Interventionboth:Count	0.004	0.039
Interventionexcer:Count	-0.007	0.008

quadratic spatial correlation structure, specified through the `corRatio` function, was selected for its suitability in capturing the decay of correlation over time within each site, a feature that aligns well with the temporal nature of our data. This correlation structure, coupled with the defined variance model, encapsulates our understanding of the data's underlying patterns, ensuring that the model's assumptions about residual variance and within-site correlations are well-aligned with the observed behavior of our dataset, thus facilitating a more nuanced and accurate interpretation of the impact of health-focused interventions on beverage selection.

B.3 Results and Diagnostics

The coefficient of the model are presented in Table 2. Moreover the residual plot in Figure 6 shows pattern-free behaviour which indicate an appropriate variance structure.

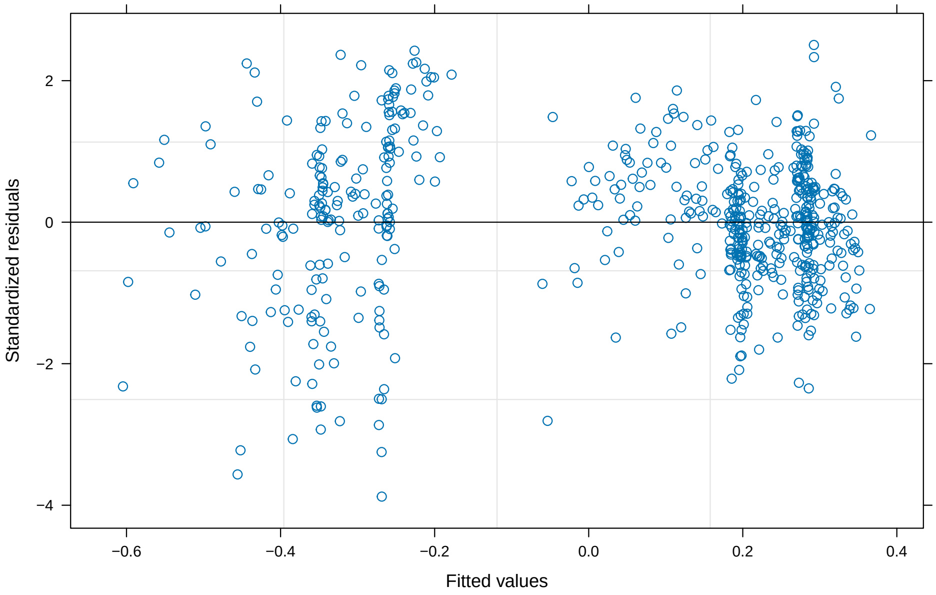

__Figure 6:__ The residual plot provides insight into the fitness of the model, reveal potential outliers, and help validate assumptions. Here, the residual plot is pattern-free and scattered roughly symmetrically around zero. The heteroscedasticity seen in residual plot before applying the appropriate variance structure is not apparent anymore.

C Contributions

Parham Pishrobat (71097927): Introduction, data, Formal Analysis, Appendix B, Conclusion, other EDA plots, formatting

Johnson Chen (85784080): LMEM, Correlation, appendix, formatting, edits after peer review, proofread.

Sarah Masri (97415681): GEE, Missing Values, conclusion, formatting, edits after peer review, proofread.