Comparing intervention effects on zero-calorie beverages consumption

Yian Lin, Haibo Wu, Wakeel Adekunle Kasali

2024-03-01

1 Introduction

Sugary beverage consumption poses health concerns because of its association with several health challenges like obesity and diabetes. Different intervention methods are proposed to steer consumers toward zero-calorie beverages that are thought healthier. This study aims to investigate the efficacy of five different intervention methods, three of which focus on visual calorie messaging via posters, and the other two focus on price discount. To assess and compare different intervention methods, sales of zero-calorie beverages and sugary beverages under different interventions were monitored in three different sites.

The statistical questions of interest are:

1. Do the interventions significantly influence the sales of zero-calorie and sugary beverages?
2. Which interventions are more effective? Specifically, does the combination of interventions lead to larger effects? Does presenting calorie-equivalent messaging have larger impacts than presenting the simple calorie admonishment?
3. Does the effectiveness of different interventions vary between sites?

The structure of the remainder of this report is as follows. Following a brief description about the dataset in Section 2, suggestions on exploratory data analysis (EDA) are presented in Section 3. Section 4 proposes formal analysis methods, and Section 5 summarizes our recommendations.

2 Data Summary

This is a quasi-experiment study for investigating the effects of five interventions on consumer inclination towards zero-calorie beverages. The dataset contains the daily sales of different types of beverages (including zero-calorie, sugared and others) under various interventions

across three sites, from October 27 to May 23. In this study, the daily sales of each type of beverage are used for measuring the consumers' preference on this type of beverage. The crucial variables are the daily sales of zero-calorie beverages, sugared beverages and total of all the beverages, which are measured as continuous data. On the other hand, based on the objectives of the study, the site and intervention should be treated as categorical variables for further analysis, and the days of week are also recorded as a categorical variable to determine potential impacts of time on the outcomes. Additionally, the dataset includes a time-counting variable that captures the timeline of the entire experiment, indicating that each site has around 220 time points. For the entire dataset, the records of these key variables are relatively complete, with sugared and zero-calorie beverages having nine observations missing. However, records for other types of beverages show a certain proportion of missing data, which could be because they are not the items of interest and excluded in recording.

3 EDA Analysis

Before doing any analysis, it is important to first assess data missingness. A graph (like Figure 1) can be drawn to visualize the proportion of missing values for all columns in the dataset. If there are missing data, it is suggested to create a table to show the data missingness by `Site` for columns that have missing data (see Table 1). To see for each site from which intervention type the missing data comes from, an expanded table can be made to show the data missingness by `Site` and `Intervention` for columns with missing data (see Table A1 in Appendix). For columns that will be used in the analysis to answer the research question, it is important to find out why the data are missing.

Figure 1: The proportion of missing values for all columns in the dataset.

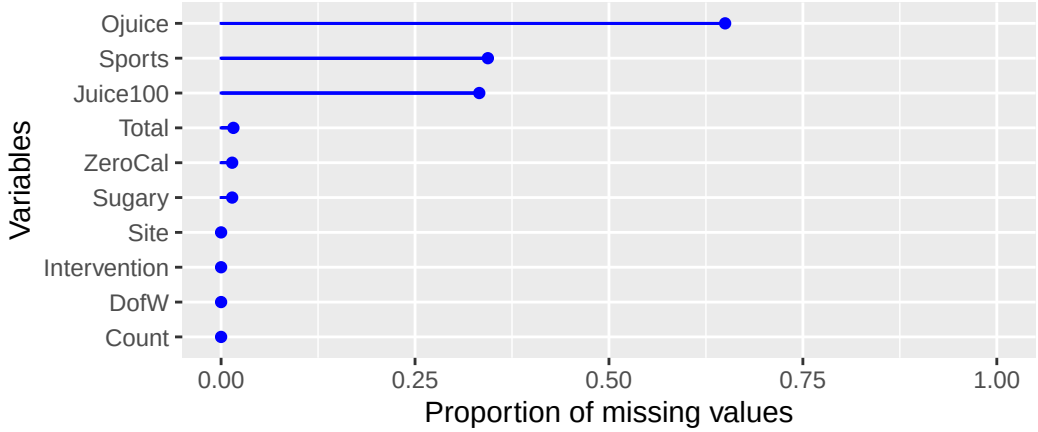


Table 1: Number of missing observations for columns with missing data present by site

Site	ZeroCal	Sugary	Juice100	Ojuice	Sports	Total
chop	0	0	0	0	0	0
HF	7	7	208	208	15	8
NS	2	2	2	202	202	2

To explore the change of daily sales of zero-calorie beverages and bottled sugared beverages over time, time series plot for `ZeroCal` and `Sugary` can be made for each of the three sites. It is important to note that the increase in the raw counts of say zero-calorie beverages being sold does not necessarily mean the decrease in the sales of sugared beverages. It could be the case that more customs came to the shops that day so that sales of all things tended to be higher. Therefore, in addition to visualize the change of daily sales using the raw counts, it is even more important to explore the change of daily sales in percentage terms (i.e. `ZeroCal`/(`ZeroCal`+`Sugary`)*100% and `Sugary`/(`ZeroCal`+`Sugary`)*100%) (see Figure 2 as an example). Change in the zero-calorie beverages percentage can better represent the change in zero-calorie beverages consumption relative to the sugared beverages consumption, and vice versa.

To compare the consumption of a selected type of beverages (zero-calorie beverages or sugary beverages) under different `intervention` categories, boxplots with sales in percentage terms as y-axis and with `intervention` as x-axis can be made (see Figure 3 as an example). Similarly, boxplots can also be used to compare the consumption of a selected type of beverages among days of week to see if there are any patterns.

Figure 2: Daily percentage of zero−calorie beverages sales over time for each of the three sites

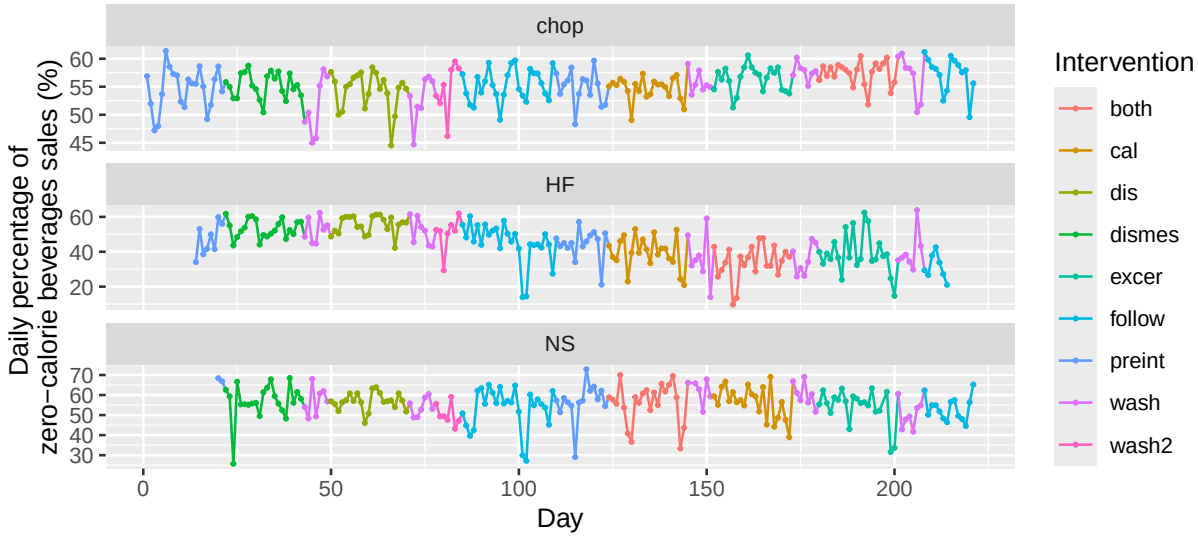
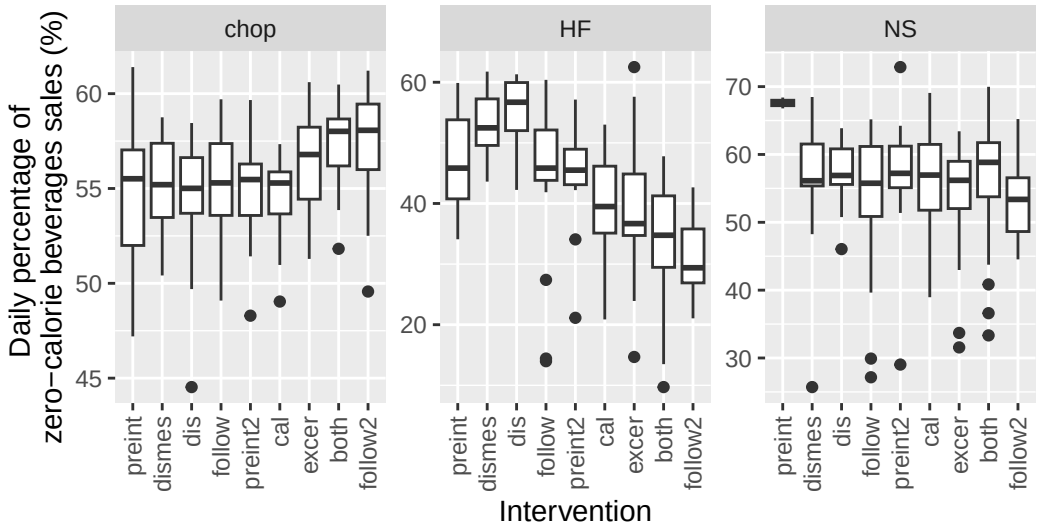


Figure 3: Boxplots of daily percentage of zero−calorie beverages sales for each of the three sites



4 Formal Analysis

To see which interventions lead to increases in zero-calorie beverages consumption and/or decreases in sugary beverages consumption, it is recommended to use generalized least squares regression (GLS). GLS is similar to linear regression. The key difference is that GLS allows the residuals to be correlated while linear regression assumes that the residuals are independent. For time series data, as the one that we have, observations close to one another may be more likely to be similar than observations that are relatively remote. Therefore, it may not be reasonable to assume that the residuals of a regression model are independent, and so GLS is preferred over linear regression.

It is recommended to fit GLS models separately for each of the three sites, with zero-calorie beverages percentage or sugary beverages percentage as the response variable. In addition to the reason mentioned in the second paragraph of section 3, another advantage of using percentage terms over raw sales is that we just need to fit models for either zero-calorie beverages percentage or sugary beverages percentage instead of for both. An increase in zero-calorie beverages percentage implies a decrease in sugary beverages percentage, vice versa.

Regarding the predictor variables, we recommend using `Intervention` and `DofW`. For `Intervention`, it is important to first rename some of the categories. For the observations made in the second pre-intervention period and the second follow-up period, the values of `Intervention` should be respectively replaced by something like "preint2" and "follow2". This way we are not mixing up the first and second pre-intervention and follow-up period in our analysis. Also, "wash" and "wash2" can be combined into one category as they are not of the main interest in the study. For `DofW`, Monday through Friday can be merged

into a category called weekday, and Saturday and Sunday can be merged into a category called weekend. When adding text label of `DofW` to each of the data points in Figure 2, it can be seen that most of the valleys of the time series happen at weekends. Therefore, it is important to control for day of week when making comparisons between intervention and pre-intervention periods. In summary, the non-residual part of each site-specific model can be specified as ZeroCal_percent ~ Intervention_updated + DofW_updated or Sugary_percent ~ Intervention_updated + DofW_updated.

When it comes to the residuals part of each site-specific model, it is suggested to compare four different correlation structure: independent, the first-order autoregressive process (AR(1)), the first-order moving-average process (MA(1)), and the combined autoregressive-moving-average processes (ARMA(p=1, q=1)). A GLS with an independent correlation structure is equivalent to a linear regression which assumes independence among residuals. A GLS with a AR(1) allows the residual at time point t to depend on the residual at the time point t-1. A GLS with a MA(1) is another way to capture temporal correlation among residuals, and ARMA(p=1, q=1) is the combination of AR(1) and MA(1) (Fox 2016). For each site, the GLS model with the smallest AIC values can be chosen as the final model. As a result, there will be three final models, one for each site, and the estimated coefficients of different interventions can be used for intervention effect assessment and comparison.

One limitation of this approach is the challenge in statistically testing the differences in intervention effects across different sites. However, we could still learn differences between sites in a sense that maybe statistically significant effects are detected for some interventions in some sites but not in other sites.

5 Conclusions

It is recommended to use daily sales in percentage terms as response variable for the analysis, to better represent the relative consumption of zero-calorie and sugared beverages. Besides the tables showing the missingness of the dataset, the time series plots can be displayed to visualize the daily changes of zero-calorie drinks consumption, and the side by side boxplots can illustrate the relationship between the daily sales and various interventions at each site.

We suggest to implement GLS for each site to capture the association between changes in beverage consumption and the adopted interventions, allowing us to account for the dependence structure of residuals of the model. Specifically, `Intervention` and `DofW` should serve as the predictor variables. The optimal model for each site can be determined by comparing the AIC across several candidate models, with each model specifying one of four different correlation structures. Differences in the effectiveness of various interventions and sites are identifiable by comparing the coefficients of the final models.

References

Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. Sage.

Statistical Appendix

Table A1: Number of missing observations for columns with missing data present by site and intervention type.

Site	Intervention	ZeroCal	Sugary	Juice100	Ojuice	Sports	Total
chop	both	0	0	0	0	0	0
chop	cal	0	0	0	0	0	0
chop	dis	0	0	0	0	0	0
chop	dismes	0	0	0	0	0	0
chop	excer	0	0	0	0	0	0
chop	follow	0	0	0	0	0	0
chop	preint	0	0	0	0	0	0
chop	wash	0	0	0	0	0	0
chop	wash2	0	0	0	0	0	0
HF	both	0	0	21	21	0	0
HF	cal	0	0	21	21	0	0
HF	dis	0	0	21	21	0	0
HF	dismes	0	0	21	21	0	0
HF	excer	0	0	21	21	0	1
HF	follow	7	7	39	39	7	7
HF	preint	0	0	22	22	8	0
HF	wash	0	0	35	35	0	0
HF	wash2	0	0	7	7	0	0
NS	both	0	0	0	21	21	0
NS	cal	0	0	0	21	21	0
NS	dis	0	0	0	21	21	0
NS	dismes	0	0	0	21	21	0
NS	excer	1	1	1	21	21	1
NS	follow	0	0	0	39	39	0
NS	preint	0	0	0	16	16	0
NS	wash	1	1	1	35	35	1
NS	wash2	0	0	0	7	7	0