STAT550 Homework No 2: Statistical Advice on Investigating the Impacts of Interventions on Sugary and Zero-Calorie Beverage Consumption

Son Luu (71843379), Xihan Qian (54285556) and Javier Mtz.-Rdz. (94785938)

March 1, 2024

1 Introduction

Zero-calorie beverages have been promoted as an alternative to sugary drinks to reduce the harmful effects of artificial sweeteners on consumers. The advised study investigates the impact of interventions, such as messaging and discounts, on the buying behaviour of zero-calorie and sugary drinks.The statistical advice for that study outlines how to quantitatively evaluate four research questions on this topic. The statistical questions to be answered include the following: 1) whether the interventions lead to increased consumption of zero-calorie beverages and decreased consumption of sugary beverages; 2) whether the effects vary across different sites 3) whether combining interventions will have a larger effect; and 4) whether calorie or exercise messaging results in larger effects.

2 Data Description and Summaries

To evaluate the effects of interventions, the study gathered data on beverages sold at four cafeterias and three convenience stores across three sites. Daily sales of these beverages were recorded for a period of 221 days and summarized by site. The starting date of observations for each site varied: site A on day 1, site B on day 14, and site C on day 20.[1] In total, there are 631 observations in the dataset.

The dataset includes variables related to time, sales, site, and intervention. The time variables are the count of days since the start of the study and the day of the week. The sales variables include zero-calorie, sugary, 100% juice, orange juice, sports, and total beverages sold, but only zero-calorie and sugary beverages are considered for this analysis. Each site and intervention has a variable. Table 1 summarizes the variables in the dataset, including they are classified and measured.

Table 1: **Description of variables**

Variable	Type
Day of the quasi-experiment	Ordinal
Day of the week	Categorical
Site	Categorical
Intervention	Categorical
Sugary beverages sold	Discrete
Zero-calorie beverages sold	Discrete
Other beverages sold	Discrete

In addition to the periods that were not recorded at the beginning of the study in sites B and C, there are nine missing values for sales of zero-calorie and sugary beverages. The missing observations are from the last week of site B and two days of site C. Aside from the missing information at the beginning and end of the study, the other missing values are unrelated to any specific factor. Furthermore, the sales data for other beverages and the total number of sales have several missing values, but they do not affect this analysis.

[1] In this report, Site A corresponds to chop in the dataset, Site B to NF and Site C to NS.

3 Exploratory Analysis

Given that the data consists of a time series of sales across three sites, it was necessary to carry out a time-based analysis. In that sense, Figure 1 helps visualize the beverages sold and the shadows behind the lines display the distinct intervention periods by site. In particular, it shows some important characteristics of the dataset. Firstly, the measurements for each site begins at different time points. Secondly, the site C shows an increase in sales during most of the first intervention, but afterwards, sales stay at a lower and more stable level. Lastly, a weekly seasonal effect can be noticed.

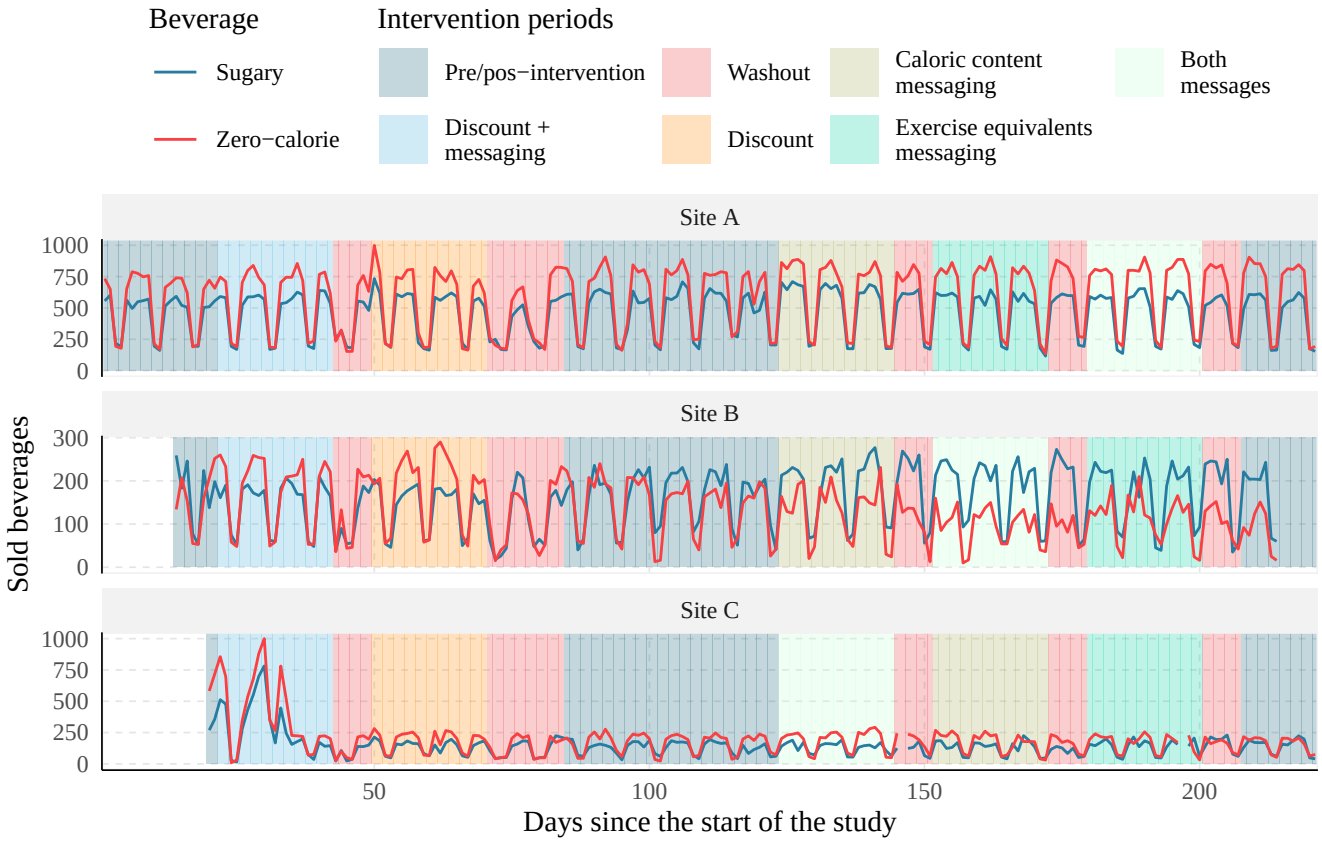

Figure 1: **Sale of sugary and zero-calorie drinks by site**

4 Formal analysis

A preparation step to address some elements that are identified in the Exploratory Analysis along with two statistical methods, Generalized Linear Models (GLM) and Generalized Linear Mixed-Effect Model (GLME), are suggested for the formal analysis of the data. In the following subsections, the necessary data adjustments, a detailed explanation of each recommended model and the suggested procedure to conduct and analyze the results are presented.

4.1 Pre-analysis

Before conducting the statistical analysis, it is important to address some elements that are identified in the Exploratory Analysis that may impact the results. As sites B and C have a short pre-intervention period, it is necessary to compensate for the missing information. This can be done by using pos-intervention and pre-intervention categories as the baseline for the study. Lastly, we are not interested in the washout periods to reduce the effects of previous interventions, so such data can be ignored.

4.2 Generalized Linear Models

Generalized Linear Models expand traditional linear regression to accommodate various data types, including those that do not fit the normal distribution mold. GLM uses link functions to connect predictors to the response variable in a way that is suitable for the distribution of the data. This approach is especially useful for analyzing count data, which count occurrences of events and often do not align with linear regression assumptions.

For count data, GLM commonly employs the Poisson and Negative Binomial distributions. The Poisson distribution fits count data assuming events occur independently at a constant rate, suitable for evenly distributed counts. When counts exhibit more variability than Poisson can handle, the Negative Binomial distribution steps in. It offers an additional parameter to better model data with variance exceeding the mean, making it a flexible choice for analyzing count data within the GLM framework. In addition to the distributional assumption, GLM assumes that the error terms are independent.

4.3 Generalized Linear Mixed-Effect Model

Generalized Linear Mixed-Effects model is a statistical model that accounts for fixed effects, common trends that are present at all levels of the data, and random effects, group-specific error terms that quantify correlation within groups and variation between groups. The inclusion of random effects makes GLME especially adept at handling hierarchical and longitudinal data, where observations are grouped into different levels. Besides the additional random effects, GLME models have the same distributional and error assumptions as GLM.

In addition, GLME models can accommodate non-normal responses using different link functions to control the relationship between the response mean and the other variables. In these models, we control the variance of the response by assuming the response follows a specific distribution. For this study, we use the log link function and Poisson or Negative Binomial regression family due to the response being count data.

4.4 Recommendations

For the first question regarding the effects of each intervention, we recommend the following GLM model

$$\log(\text{Sales}) = \beta_0 + \beta_1 \times \text{Intervention} + \log(\text{Total Sales})$$

with the control period being the reference and the total sales as the offset. Both the pre-intervention and follow-up periods are considered as the control period in this case. For the

second question regarding the difference in intervention effects between sites, we recommend the same model as the first question with the addition of the interaction term between intervention and site. For the third question comparing individual interventions against their combination, we recommend the following GLM model

$$\log(\text{Sales}) = \beta_0 + \beta_1 \times \text{Calorie} + \beta_2 \times \text{Exercise} + \log(\text{Total Sales})$$

with the combination period being the reference. If the interest is also the effect of combining discount and messaging, then the following model is recommended

$$\log(\text{Sales}) = \beta_0 + \beta_1 \times \text{Discount} + \log(\text{Total Sales})$$

with the discount plus messaging as the reference. For the final question comparing the effects of calorie messaging and exercise messaging, we recommend the following GLM model

$$\log(\text{Sales}) = \beta_0 + \beta_1 \times \text{Exercise} + \log(\text{Total Sales})$$

with the calorie messaging period being the reference.

In the model for the second question, we recommend first running a preliminary GLM for each of the interventions and the pre-intervention period. If there are significant differences in estimated coefficients between interventions, a GLME model with intervention-dependent random effects is recommended. A similar strategy is recommended for the other questions where the preliminary GLM is performed for each site and site-dependent random effects are added when differences in estimated coefficients between sites are significant. The reason for these preliminary analyses is to determine if GLME is necessary or not.

After the models are fitted, the residuals need to be visualized via scatter plot, autocorrelation plot and QQ-plot to confirm independent, equal variance and normal error assumptions for both GLM and GLME. For hypothesis testing and interpretation, the Bonferroni correction should be applied to the significance level. This means dividing the significance level by the number of hypotheses being tested to counter false rejections. For model diagnostics, we recommend using QQ-plots to confirm the distributional assumption and scatter plot of the residuals for the independent error assumption.

5 Conclusion

Poisson and Negative Binomial GLME models are suggested to analyze the effectiveness of interventions on beverage sales across hospitals, addressing four research questions. Preliminary GLM models are used to identify differences between intervention and site effects. GLME models are recommended for formal analysis when significant variations are observed, and GLM models are suggested otherwise. This approach ensures a thorough investigation into how each intervention influences beverage choices, the impact of site characteristics and the effectiveness of calorie information and exercise prompts. The methodology recommended would provide a clear path to understanding which interventions work best in promoting healthier beverage consumption.

A Appendix

A.1 Formal analysis

A.1.1 Bonferroni correction

When using the Bonferroni correction, the significance level is adjusted by dividing 0.05 by the number of hypotheses being tested, in this case, 46. The confidence level for the confidence intervals is $1 - 0.05/46$.

A.1.2 Link function

Poisson GLME models exhibit overdispersion for this data (mean being less than variance). Therefore, the results above are obtained using negative binomial GLME, which eliminates overdispersion. In addition, negative binomial models have better QQ plots, which is illustrated in Figure A.1 and Figure A.2.

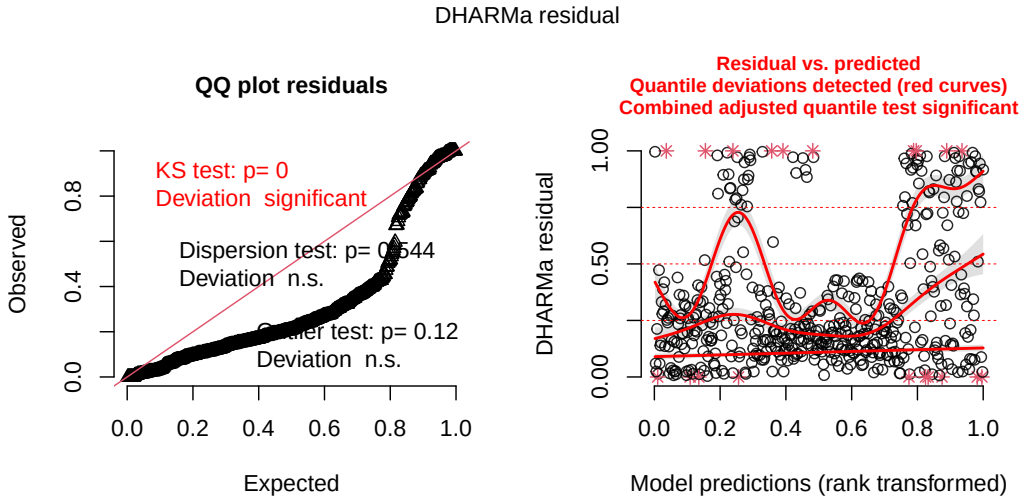

Figure A.1: **QQ and residual scatter plots for poisson GLME model of first question**

A.2 Preliminary GLM

In the preliminary analysis examining the efficacy of interventions designed to encourage the consumption of zero-calorie beverages, we are particularly interested in understanding how these interventions perform across different hospital sites. To make informed decisions about our model structure, specifically whether to include random slopes or random intercepts, we generate plots that visualize the effects of interventions across the sites.

With only three sites in our study, incorporating random effects is not advisable due to the limited data available. A small number of groups can lead to overfitting and unreliable estimates of variability. Therefore, despite variations in the preliminary analysis, we will not include random effects for sites, keeping our model straightforward against the challenges of a small sample size.

Now we proceed to determine if intervention-specific random effect is needed. By using `ggcoef_compare`, we are able to produce plots that display the confidence intervals for the

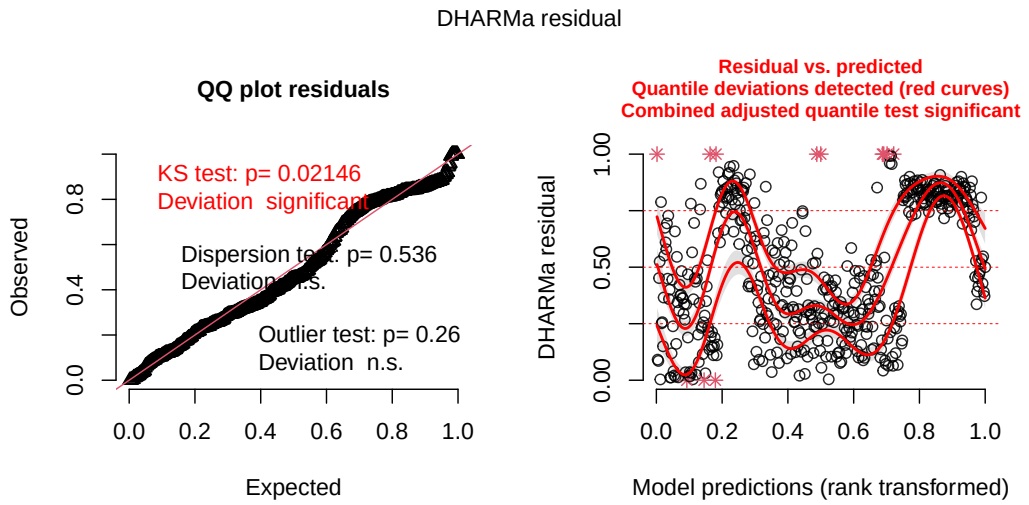

Figure A.2: **QQ and residual scatter plots for negative binomial GLME model of first question**

incidence rate ratios (IRR) of each intervention, broken down by site. These plots are a powerful diagnostic tool—they help us to visually inspect if there is a notable variation in the effects of interventions within the different hospital sites. With the plot containing both intercept and slope, we will be able to determine whether we need random intercept or slope based on the variability.

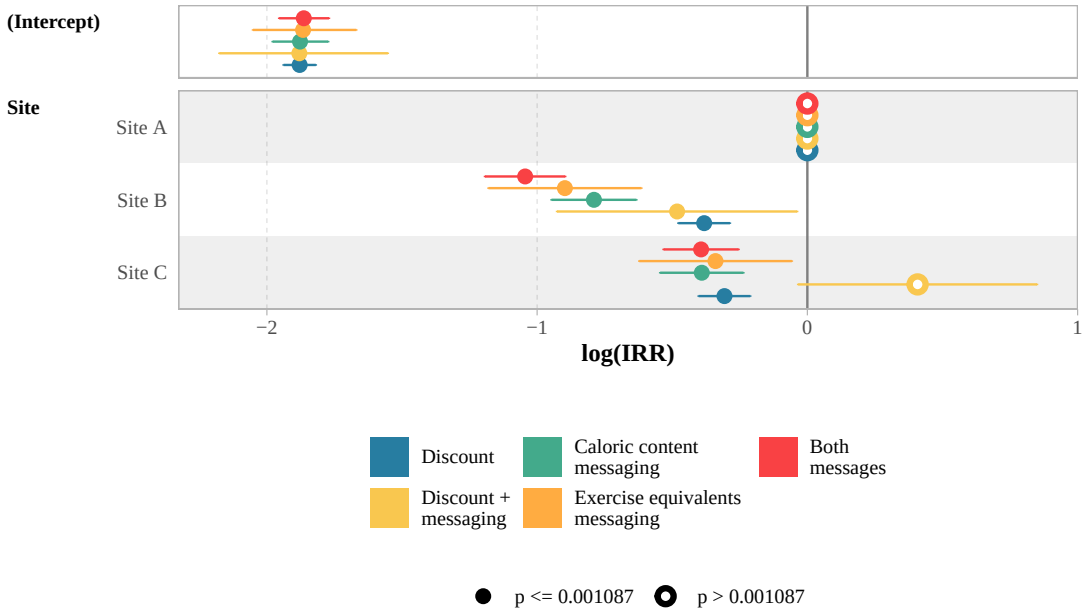

Figure A.3: **CIs for consumption of Zero Calorie Beverages**

From the plot provided relating to consumption of zero-calorie beverages, we see that the confidence intervals within each site demonstrate variation in the effect sizes of the interventions. This pattern suggests substantial heterogeneity in the treatment effects within sites, indicating that the inclusion of random slopes for the interventions would be appropriate to model this within-site variation. Conversely, the relatively consistent placement of the confidence interval

centers across the different sites argues against the need for random intercepts, as it implies little variation in the baseline effects between sites. Thus, a mixed-effects model with random slopes for interventions would be a good choice. A comparable plot can be produced to examine the consumption of sugary beverages, which similarly indicates the necessity of incorporating random slopes into the model to account for the variability in effects.

A.2.1 Results

For the first question regarding the effects of each intervention, there is a significant increase in zero-calorie beverage sales when discount plus messaging is implemented and no similar increase in sugary beverage sales as can be observed on Figure A.4. Other interventions have no significant effect. This suggests that discount plus messaging is effective at encouraging people to buy zero-calorie beverages.

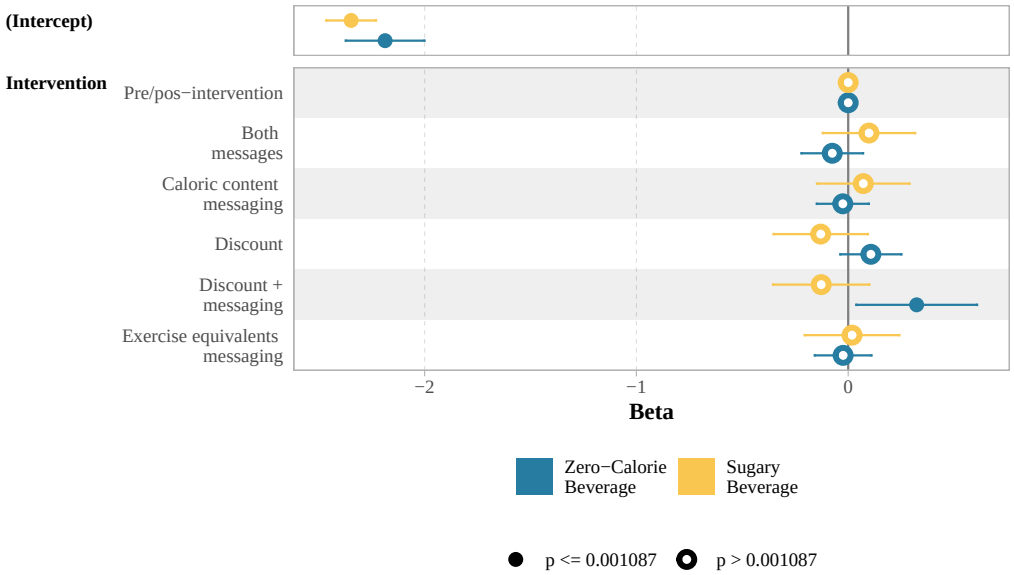

Figure A.4: **CIs for consumption of Zero-Calorie and Sugary Beverages by Intervention**

For the second question regarding the difference in intervention effects between sites, the results are summarized in Figure A.5. It is interesting to note that discount plus messaging has different effect across sites for both zero-calorie and sugary beverages.

For the third question comparing individual interventions against their combination, using both calorie and exercise messaging is not significantly different from implementing them individually. But discount plus messaging does improve sales for zero-calorie beverages compare to discount only.

For the final question comparing the effects of calorie messaging and exercise messaging, there are no significant results for both zero-calorie and sugary beverages.

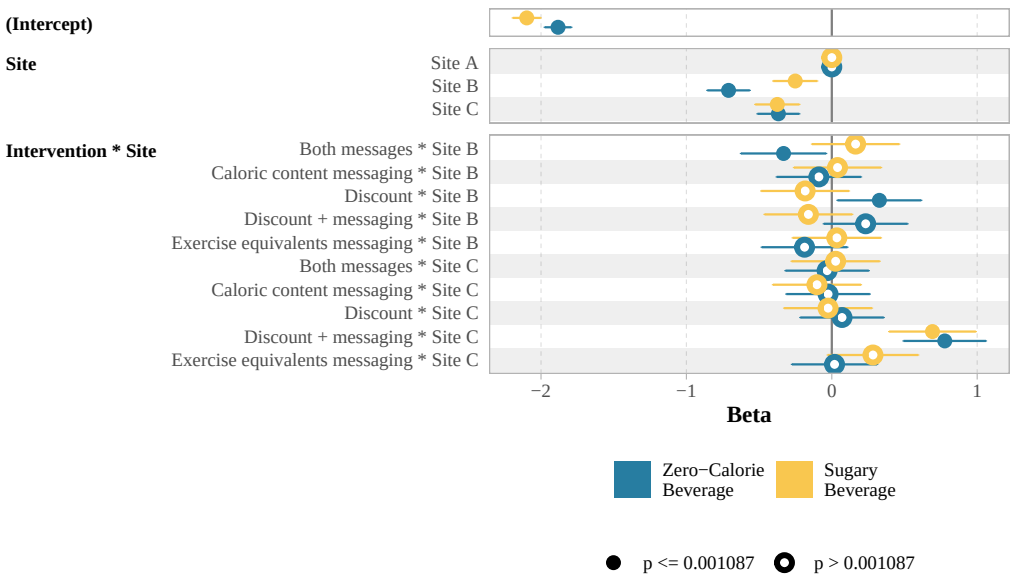

Figure A.5: **CIs for consumption of Zero-Calorie and Sugary Beverages by Intervention and Site**