What is the best way to promote zero calories beverages?

Ivy Liu, Zefan Liu, Tom Tang

Introduction

Zero calorie beverages are healthier compared to sugared beverages because they contain less sugar and significantly fewer calories. Over the past decades, hospitals have been striving to promote zero-calorie beverages for a healthier lifestyle. This study employs five interventions to help increase such promotion. These interventions include a 10% price discount, a price discount combined with messaging that explains the reason for the discount, messaging displaying the caloric content in sugared beverages, messaging indicating the amount of exercise needed to burn off the calories in a sugared beverage, and messaging containing both caloric content and the amount of physical activity needed.

The study aims to investigate whether these interventions help encourage people to choose zero calorie beverages over sugared beverages and if such effects varies across hospitals.

Data description and summaries

This is an experimental design where data are collected over 30 weeks, from October 27 to May 23, with a follow-up period of 14 days, gathered from two urban hospitals and one suburban hospital. During this time period, the sales of different types of drinks are recorded (see Table 1).

There are missing entries in this dataset, with a significant portion of the missing data attributed to the 'Juice100,' 'Ojuice,' and 'Sports' columns. Since neither these measurements themselves nor their links with the consumption of ZeroCal/Sugary beverages are the main focus of the study, the missing values in these columns should not be given specific attention. Additionally, there are seven entirely missing consecutive entries at HF hospital during the follow-up period and two isolated cases of missing entries. Such occurrences could be due to public holidays or temporary closures for renovation. In general, such missing data can be classified as missing at random.

Table 1: Summary of data

Data Name	Data Type	Data Description
DofW	Categorical	The day of the week, 7 levels
Site	Categorical	The location of the hospital, 3 levels
Intervention	Categorical	The intervention applied, 9 levels
ZeroCal	Integer	The number of zero calorie drinks sold
Sugary	Integer	The number of sugary drinks sold
Juice100	Integer	The number of Juice100 sold
Ojuice	Integer	The number of Ojuice sold
Sports	Integer	The number of sports drinks sold
Total	Continuous	The sum of all drinks sold

Exploratory Data Analysis

Firstly, to address the influence of location on zero-calorie beverage consumption, which is of primary interest to the study, a spaghetti plot is recommended. A spaghetti plot is typically used to demonstrate the change of multiple flows over time and how they vary across the three sites. Based on Figure 1, it is evident that the time-series data for zero-calorie beverage consumption at the site "chop" are significantly higher than those at the other two sites, suggesting substantial variation in consumption distributions under different interventions. Additionally, the spaghetti plot reveals that except for the initial period at the site "HF," there is no apparent trend at all three sites, as the long-term averages seem to be stable. However, strong weekly variations are evident, implying that it is reasonable to include the day of the week (DofW) and exclude the time (count) in the model.

Formal analysis

It is recommended to implement a Generalized Linear Mixed Model (GLMM) to address the problems. A GLMM, an extension of the Linear Mixed Model (LMM) (Winter 2013), can incorporate both fixed and random effects, and also accommodate the response variable being a count variable through a link function (Dobson and Barnett 2018). Specifically, while the fixed effects assess the association of covariates with the response variable across the overall population—serving as a baseline—the random effects account for variations in some associations across different entities or individuals.

In this study, the primary focus is to analyze the effects of the interventions, and therefore, it is suggested to treat the intervention as a fixed effect. Additionally, days of the week should also be considered as fixed effects because they apply to all three sites. To allow for a more comprehensive understanding of the intervention effects, it is advised to include sites and total as two random effect in the model. These covariates control for variability that might influence the patterns of zero-calorie/sugary beverage consumption. For instance, site differences (which

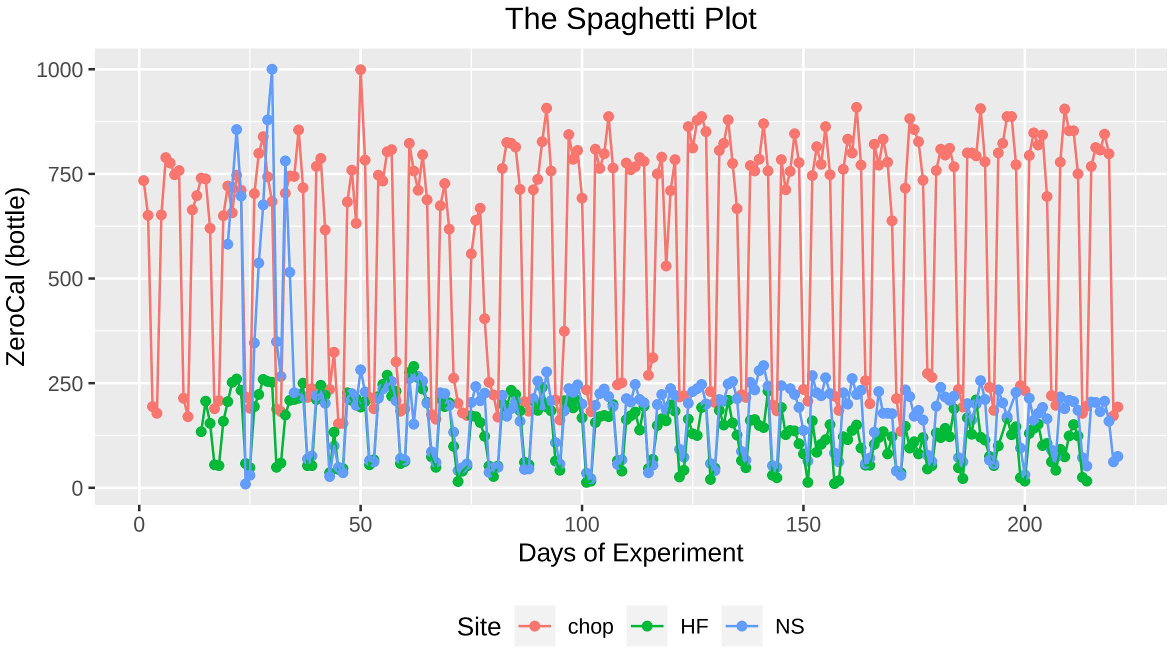


Figure 1: Examples of Recommended EDA Plots

are the secondary statistical question) may impact purchasing behavior due to demographic variations between urban and suburban areas. Similarly, accounting for total daily beverage consumption helps isolate the effect of human traffic on zero-calorie/sugary beverage choices. Lastly, given that the consumption of ZeroCal beverage is count (non-negative integer), it is natural to assume that it follows a Poisson distribution. Subsequently, using an exponential link function to guarantee a positive output from model is suggested.

Hence, the proposed models are

$$
\begin{aligned}
ZeroCal &\sim Intervention(Site) + DofW + Total \\
Sugary &\sim Intervention(Site) + DofW + Total.
\end{aligned}
$$

Fixed effects include Intervention, DofW, and Total, while the random effect measures how the effect of Intervention and baseline consumption varies across sites. Note that these models are simplified versions, especially with the omission of the link function. Please refer to the appendix for the formulated models.

However, there are several limitations to the models mentioned above. Firstly, GLLMs do not accommodate autocorrelation in time series data, which refers to the dependence between consecutive observations. Moreover, as mentioned, the model proposed implicitly assumes that the response variables follow Poisson distributions (Coxe, West, and Aiken 2009), where the expectation and variability of response variables should equate, but this assumption can be easily violated in practice.

Conclusions

We recommend using GLMMs to fit the relationship between covariates and response variables. Based on different research interests, response variable should be either ZeroCal beverage consumption or sugary beverage consumption. The fixed-effect explanatory variables suggested are intervention, Total and DofW, and the random-effect explanatory variable is Site.

References

Coxe, Stefany, Stephen G West, and Leona S Aiken. 2009. "The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives." *Journal of Personality Assessment* 91 (2): 121–36.

Dobson, Annette J, and Adrian G Barnett. 2018. *An Introduction to Generalized Linear Models*. CRC press.

Winter, Bodo. 2013. "Linear Models and Linear Mixed Effects Models in r with Linguistic Applications." *arXiv Preprint arXiv:1308.5499*.

Statistical Appendix

Mathematical Formulation of Models

There are two types of generalized linear mixed models fitted in this report, adapted for different structures of data. For illustration purpose, only the two models for zerocal beverage sales are shown. The first model (Equation 1) is GLMM with Poisson distribution, which is suitable for count data where the counts are grouped by random effects. In this case, the random effect is the site.

$$\log(\text{Zerocal}_{ijk}) = \beta_0 + \beta_1 \cdot \text{Intervention}_i + \beta_2 \cdot \text{DoW} + \log(\text{Total}_{ijk}) + u_j \qquad (1)$$

where "Zerocal" is the count for the daily sales of zerocal beverages, "DoW" encodes day of the week, and "Total" is the daily sales of all beverages. In addition, β_0 is the intercept, β_1 is the coefficient for intervention, β_2 is the coefficient for "DoW", and u_j is the random effect for the j-th Site. Notice that log(Total) is regarded as an offset because we want its coefficient to be fixed at 1.

The second model (Equation 2) is a GLMM with negative binomial distribution. Similar to the first model, it is also used for count data, but is more suitable when there is overdispersion (the variance is greater than the mean).

$$\log(\text{Zerocal}_{ijk}) = \beta_0 + \beta_1 \cdot \text{Intervention}_i + \beta_2 \cdot \text{DoW} + \log(\text{Total}_{ijk}) + u_j \qquad (2)$$

Model Selection

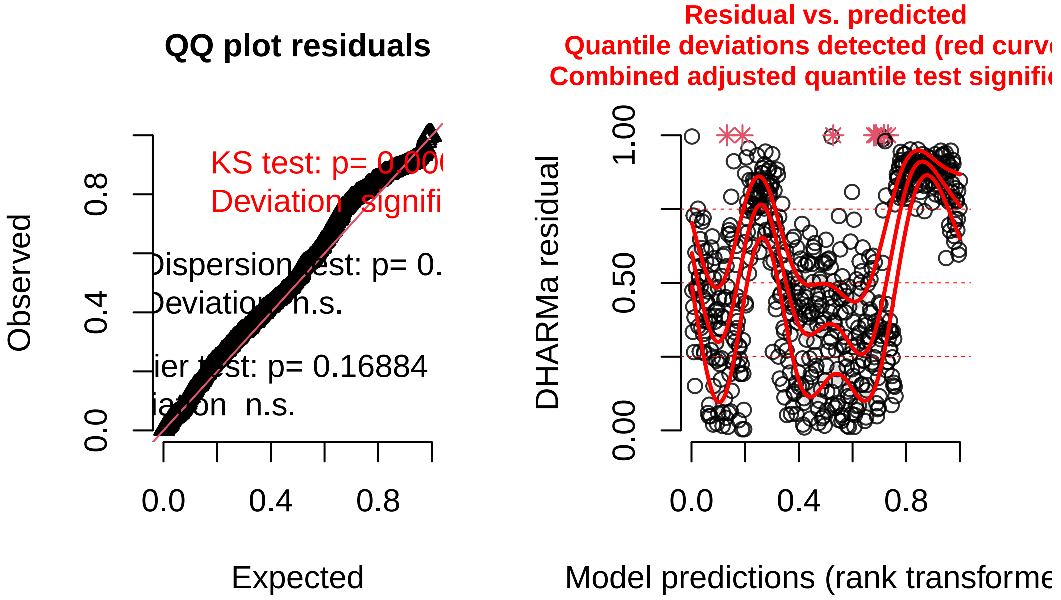
It is suggested in the formal analysis section that model should be selected based on AIC criterion. Table 2 shows the AIC comparison.

Table 2: Model selection based on AIC criterion

GLMM Model	Zerocal beverage sales	Sugary beverage sales
Poisson	14874.39	12096.9
Negative Binomial	6765.097	6529.67

THe GLMMs using negative binomial distribution are suggested because they give lower AIC scores.

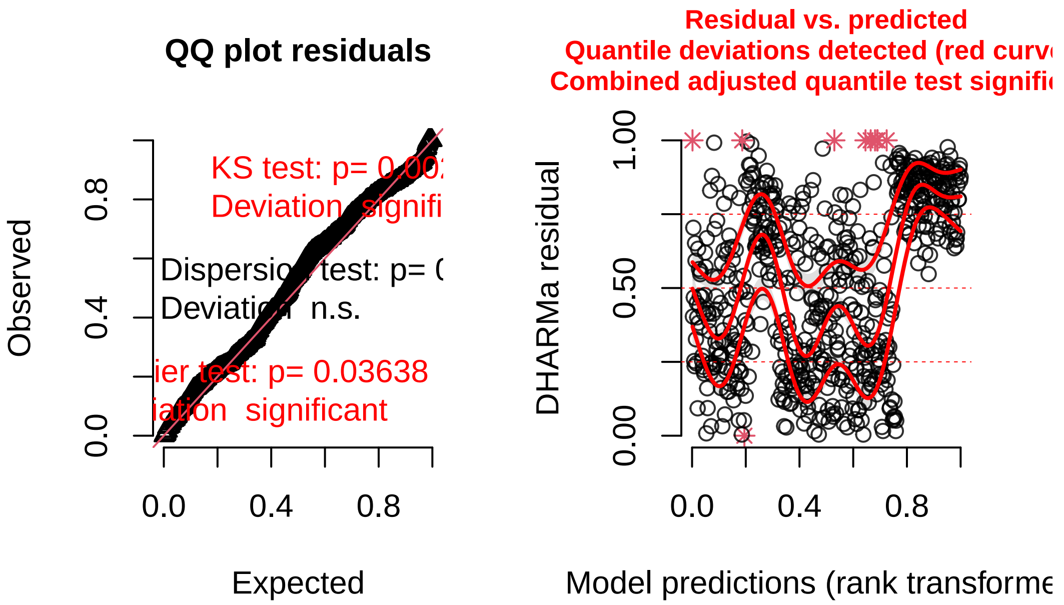
Model Diagnostic



```
    DHARMa nonparametric dispersion test via sd of residuals fitted vs.
    simulated

data:  simulationOutput
dispersion = 1.4593, p-value = 0.24
alternative hypothesis: two.sided
```

```
DHARMa:testOutliers with type = binomial may have inflated Type I error rates for integer-val
```



```
    DHARMa nonparametric dispersion test via sd of residuals fitted vs.
    simulated

data:  simulationOutput
dispersion = 1.4164, p-value = 0.2
alternative hypothesis: two.sided
```

For both models, we generated QQ-plots and residual plots, and we also conducted tests to determine if there's overdispersion. The QQ-plots indicate a good model fit as they form relatively straight lines. Regarding the residual plot, there are quantile deviations indicating that some data points fall outside the range of simulated values. However, since our focus is on inference rather than prediction, and we lack information on the extent of deviation from the model expectation, this issue is not considered major. This also suggests that despite not being the optimal model, this model is still adequate. Lastly, we want to test whether there's overdispersion, meaning whether the variance of the response is greater than what's assumed by the model. Both dispersion numbers are close to 1, which indicates no major overdispersion.

Limitations

When utilizing a GLMM model, a significant drawback arises from its inability to accommodate the time series structure of the data. In other words, this model fails to capture the dependent structure where the output variable is linearly dependent on its previous values.