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 are consistent across hospitals. The primary statistical question in this context is: "Is there any association between interventions and the increase in zero calorie beverage consumption or decrease in sugary beverage consumption?" The secondary statistical question is:"Does the effect of intervention differs by site?"

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.

Data Name	Data Type	Data Description
Count	Integer	The number of days
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

Data Name	Data Type	Data Description
ZeroCal	Integer	The number of zero calorie drink sold
Sugary	Integer	The number of sugary drink sold
Juice100	Integer	The number of Juice100 sold
Ojuice	Integer	The number of Ojuice sold
Sports	Integer	The number of sports drink sold
Total	Continuous	The sum of all drink sold

There are missing entries in this dataset, with a significant portion of the missing data attributed to the 'Juice100,' 'Ojuice,' and 'Sports' columns. These columns will not be included in the model because of their irrelevancy, so the missing entries in these columns will not affect the statistical analysis. Additionally, there are seven entirely missing consecutive entries at HF hospital during the follow-up period. This is possible due to the hospital's cafeteria and convenient shop close down for renovation. Apart from these, there are two isolated cases of missing entries. These can be reasonably attributed to temporary closures for public holidays.

Exploratory Data Analysis

Specific to the statistical questions listed in the introduction, the following exploratory data analysis procedures are suggested. 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 clear 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 that the distributions of consumption may vary substantially under different interventions. Additionally, there are two observations. Firstly, the spaghetti plot reveals that except for the beginning 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 presented, implying that it is reasonable to include the day of the week (DofW) and to exclude the time (count) in the model.

Formal analysis

This section outlines general strategies for quantifying and comparing the impacts of five different interventions on the consumption of zero-calorie and sugary beverages across three sites.

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

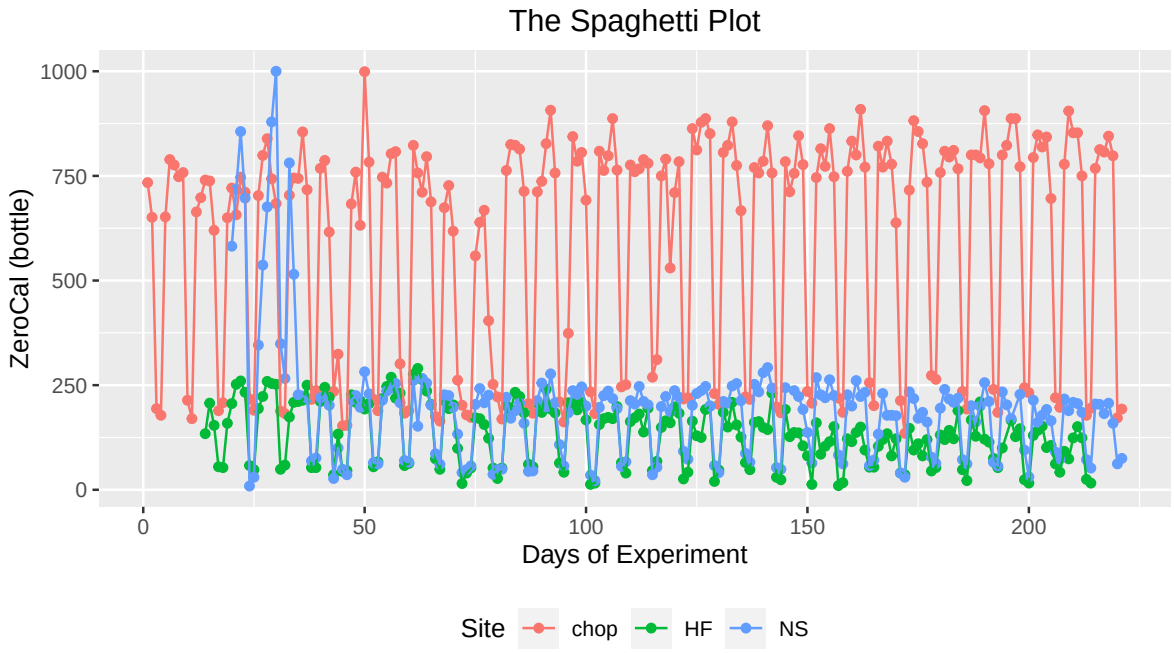

Figure 1: Examples of Recommended EDA Plots

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 and total daily beverage consumption 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 as a 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 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. Last but not least, given that the consumption of ZeroCal beverage is a non-negative integer, using an exponential link function to guarantee a positive output is suggested.

Hence, the proposed models are

$$ZeroCal \sim Intervention(Site) + DofW + Total$$
$$Sugary \sim Intervention(Site) + DofW + Total.$$

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, 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

Here is a formulation of the Generalized Linear Mixed Model within the context of this research. As suggested earlier, random effects can be placed on both the intervention's slope and intercept. Taking $response = ZeroCal$ as an example, the model will take the format of

$$E(zero\_calorie_{ij}) = exp((\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\, Intervention + \beta_2\, DofW + \beta_3\, Total)$$

where $b_{0i} \sim N(0, \sigma_1)$ and $b_{2i} \sim N(0, \sigma_2)$. Here, $i \in \{1, 2, 3\}$ is used to index the three sites, and $j \in \{1, 2, \ldots, 221\}$ is used to index the date. Both σ_1 and σ_2 can be estimated using data,

and $zero\_calorie_{ij}$ is assumed to have poisson distribution where its expectation should be equal to its variance. When such an assumption is violated, usually quasi-maximum likelihood estimation is used. This method estimates the regression parameters without the need to specify a distribution for the outcome. When such an assumption is violated, usually quasi-maximum likelihood estimation is used. This method estimates the regression parameters without the need to specify a distribution for the outcome.