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 how this varies across different 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).

Table 1: Summary of data

Data Name	Data Type	Data Description
Count	Categorical	The day from the beginning of study
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
zero calorie	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

This dataset has missing entries, especially in the 'Juice100,' 'Ojuice,' and 'Sports' columns. However, since the study does not primarily focus on these measurements or their connection to the consumption of zero-calorie/sugary beverages, the missing values in these columns shouldn't be emphasized. 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.

Exploratory Data Analysis

Firstly, to address the influence of location on zero-calorie beverage consumption, 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.

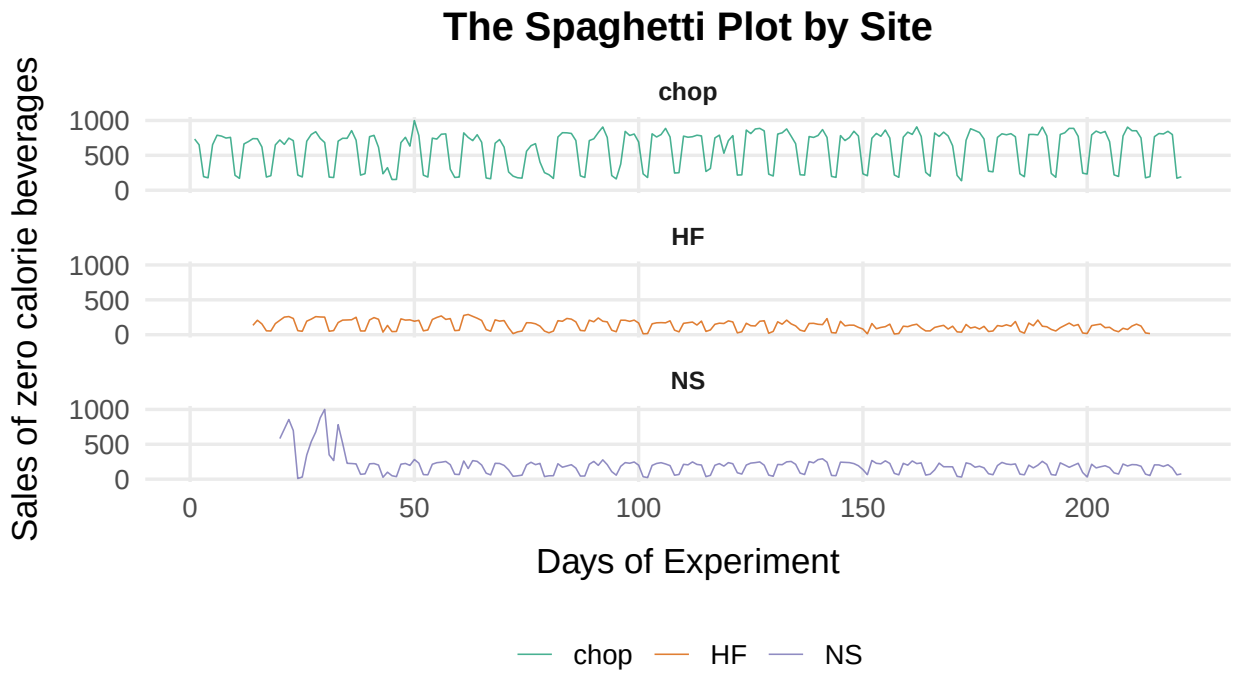


Figure 1: Examples of Recommended EDA Plots

Formal analysis

It is recommended to implement a Generalized Linear Mixed Model (GLMM) to address the primary problems and a Generalized Linear model to address the secondary problem. 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, fixed effects are variables that we expect to have an effect on the response variable. Random effects are typically grouping factors for which we are attempting to control. The link function can ensure that the range of model outputs matches the range of values

that can be observed, which is an advantage of GLMM over LMM. A GLM is essentially the same as a GLMM except the inclusion of random effects.

In this study, our primary purpose is to examine the impacts of various interventions. Hence, it is advisable to consider the intervention variable as a fixed effect. Furthermore, Day of Week should also be treated as a fixed effect, since they are consistent across all three sites. The random effect in this analysis should be Site, which accounts for variability in the observations over time. Although the total consumption volume is a plausible predictor for both variables of interest, it is more fitting to incorporate it as an offset term in the model. This approach ensures that its influence on the model is "structural," with a fixed coefficient of one, reflecting its proportional effect on the response variables.

Given that the consumption data consist of non-negative integers, a viable choice for the link function is the logarithm transformation applied to the daily sales of zero calorie, sugary, and total beverages. This transformation ensures that the model's outputs remain positive, aligning with the nature of the consumption data. Hence, the proposed models are in Equation 1 and 2. Note that the response variables can be assumed to follow either Poisson or Negative Binomial distribution using log link function.

Based on the analysis results provided in the appendix section, we conclude that a negative binomial distributed GLMM is a better model due to its lower AIC. We suspect this improvement may be attributed to the negative binomial distribution allowing the conditional variance of the outcome variable to be greater than its conditional mean, providing greater flexibility in model fitting.

To investigate whether the intervention effects are differed by sites, which is our secondary purpose, we propose a generalized linear model on a negative binomial distribution. This model should incorporate an interaction term between site and intervention. This will allow us to discern whether the intervention impacts are indeed site-specific. In addition, we also include the day of the week, and an offset term of total daily sales in this model, which allows to normalize the effect per unit sale and account for the weekly fluctuation. The following analysis can be conducted by examining the coefficient of the interaction terms.

Conclusions

Overall, to assess the impact of various interventions on the consumption of zero calorie and sugary beverages at the three sites, we propose a GLMM with a negative binomial distribution for both variables of interest. According to the model outputs, only discount along with discount messaging has significant impact on the beverage sales. It is effective in promoting the consumption of zero calorie beverages, and counter intuitively the consumption of sugary beverages as well. However, the intervention by sites analysis reveals that although these intervention strategies have some effect in promoting zero calorie beverage consumption, the magnitude of the effects highly depend on sites. For more detailed insights, please refer to the statistical appendix.

References

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 zero calorie 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.

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

where "zero calorie" is the count for the daily sales of zero calorie 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).

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

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	zero calorie 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.

GLMM model output

Shown in table 3 and 4 is the result obtained from generalized linear mixed model using negative binomial distribution.

Based on the two tables above, the GLMM with a negative binomial distribution outperform other models in terms of the Akaike Information Criterion (AIC), a widely used criterion for model selection. For zero-calorie beverage consumption, the model indicates that only the "discount+messaging" intervention has a statistically significant effect. Similarly, for sugary beverage consumption, the effect of the "discount+messaging" intervention is also significant. Additionally,

Table 3: Percent change in sales of zero-calorie beverages under intervention

Intervention	Estimate	Standard Error	p-Value
Discount	6.26%	3.57%	0.079
Discount + discount messaging	14.87%	3.44%	0.000
Calorie messaging	1.60%	3.57%	0.65
Exercise equivalent messaging	-3.44%	3.64%	0.34
Calorie + exercise equivalent messaging	-4.10%	3.59%	0.25

Table 4: Percent change in sales of sugary beverages under intervention

Intervention	Estimate	Standard Error	p-Value
Discount	0.45%	3.35%	0.89
Discount + discount messaging	8.05%	3.24%	0.013
Calorie messaging	4.26%	3.24%	0.19
Exercise equivalent messaging	-0.90%	3.33%	0.79
Calorie + exercise equivalent messaging	-1.96%	3.30%	0.55

the effects associated with "DofW=6" (Saturday) and "DofW=7" (Sunday) are significant, suggesting that, on average, sugary beverage consumption may vary between weekdays and weekends. Lastly, by including the total beverage consumption as an offset in our analysis, we observe that the day of the week typically does not exert a significant influence on the response variables.

GLM for Evaluating Intervention Effects by Sites

As shown in equation 3, the GLM model uses a logarithmic link function on the response variables.

$$\log(\text{zero calorie}) = \beta_0 + \beta_1 \cdot \text{Intervention} : \text{Site} + \beta_2 \cdot \text{DoW} + \log(\text{Total}) \tag{3}$$

Here shows the selected model output for zero calorie beverages.

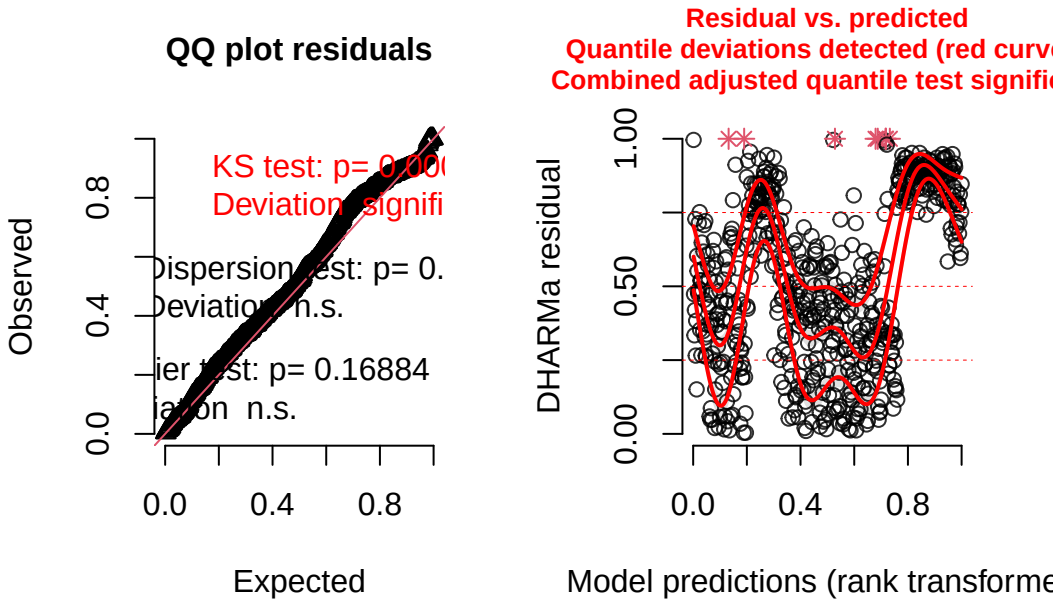
```
                           Estimate Std. Error z value Pr(>|z|)
Sitechop:Interventionboth   0.4807499  0.1075268   4.471 7.79e-06 ***
SiteHF:Interventionboth    -0.5612425  0.1099167  -5.106 3.29e-07 ***
SiteNS:Interventionboth     0.0816578  0.1086373   0.752 0.452258
Sitechop:Interventioncal    0.4665212  0.1075422   4.338 1.44e-05 ***
SiteHF:Interventioncal     -0.3313211  0.1092486  -3.033 0.002424 **
SiteNS:Interventioncal      0.0744004  0.1087687   0.684 0.493960
Sitechop:Interventiondis    0.4611542  0.1075998   4.286 1.82e-05 ***
SiteHF:Interventiondis      0.0781893  0.1086428   0.720 0.471715
SiteNS:Interventiondis      0.1638958  0.1086102   1.509 0.131292
Sitechop:Interventiondismes 0.4634593  0.1075745   4.308 1.65e-05 ***
SiteHF:Interventiondismes  -0.0124398  0.1087009  -0.114 0.908889
SiteNS:Interventiondismes   0.8742991  0.1077941   8.111 5.03e-16 ***
Sitechop:Interventionexcer  0.4776633  0.1075572   4.441 8.95e-06 ***
SiteHF:Interventionexcer   -0.4207657  0.1102873  -3.815 0.000136 ***
SiteNS:Interventionexcer    0.1254518  0.1095253   1.145 0.252038
```

Here shows the selected model output for sugary calorie beverages.

	Estimate	Std. Error	z value	Pr(>|z|)	
Sitechop:Interventionboth	0.197987	0.104370	1.897	0.05783	.
SiteHF:Interventionboth	0.112814	0.105273	1.072	0.28389	
SiteNS:Interventionboth	-0.163545	0.105756	-1.546	0.12200	
Sitechop:Interventioncal	0.292919	0.104323	2.808	0.00499	**
SiteHF:Interventioncal	0.082066	0.105238	0.780	0.43550	
SiteNS:Interventioncal	-0.178256	0.105937	-1.683	0.09244	.
Sitechop:Interventiondis	0.297091	0.104387	2.846	0.00443	**
SiteHF:Interventiondis	-0.132208	0.105700	-1.251	0.21101	
SiteNS:Interventiondis	-0.101400	0.105765	-0.959	0.33770	
Sitechop:Interventiondismes	0.263538	0.104378	2.525	0.01157	*
SiteHF:Interventiondismes	-0.142257	0.105623	-1.347	0.17803	
SiteNS:Interventiondismes	0.583786	0.104731	5.574	2.49e-08	***
Sitechop:Interventionexcer	0.234858	0.104383	2.250	0.02445	*
SiteHF:Interventionexcer	0.017126	0.106071	0.161	0.87173	
SiteNS:Interventionexcer	0.091903	0.106241	0.865	0.38701	

From the above model output, we conclude that the intervention methods have different effects at different sites. The different impacts of interventions across sites may suggest underlying demographic differences. These could be in terms of age, income, health consciousness, or other cultural factors that influence purchasing behaviors. This highlights the necessity for customized strategies that consider the unique demographic and behavioral profiles of different populations.

Model Diagnostic

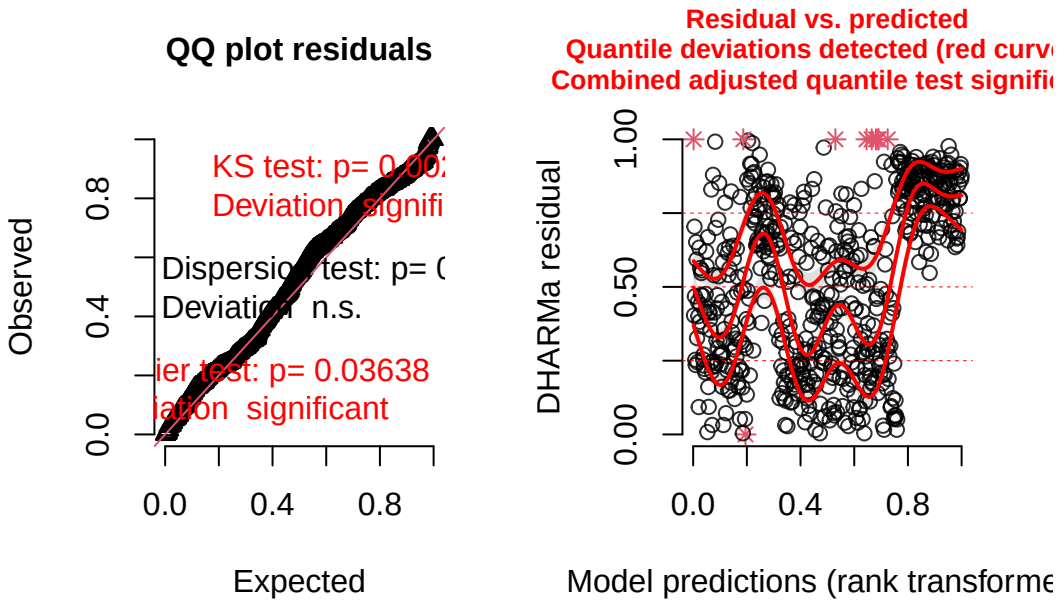



```
DHARMa nonparametric dispersion test via sd of residuals fitted vs.
```

```
    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-value
```




```
    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. Both dispersion numbers are close to 1, which indicates no major overdispersion.

Limitations

When utilizing a GLMM/GLM 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.

Contribution

Tom: Wrote introduction, data description and statistical appendix. Help edit everything else. Zefan: Wrote EDA and formal analysis. Help edit everything else. Ivy: Wrote formal analysis and statistical appendix. Help edit everything else.