Comparing intervention effects on zero-calorie beverages consumption

Yian Lin, Haibo Wu, Wakeel Adekunle Kasali

2024-03-29

1 Introduction

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

The statistical questions of interest are :

1. Whether these interventions significantly influence the consumption of zero-calorie and sugary beverages;
2. How to determine which interventions are more effective, and specifically, whether the combination of interventions lead to larger effects;
3. If the effectiveness of different interventions vary between sites.

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 applies formal analysis methods, and Section 5 summarizes our recommendations. The detailed analysis is shown in Appendix.

2 Data Summary

This is a quasi-experimental study to investigate 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 discrete count. On the other hand, based on the objectives of the study, the site and intervention are to 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 Exploratory Data 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 a table be created to show the data missingness by Site for variables that have missing data (see Table 1). To see site by site which intervention type has the missing data, an expanded table can be made to show the data missingness by `Site` and `Intervention` for variables with missing data. For crucial variables to be used in the analysis to answer the research questions, it is important to find out why the data are missing. To explore the change of daily sales of zero-calorie beverages and bottled sugared beverages over time, a time series plot for daily count of `zero-calorie` beverages and `sugary` beverages can be made for each of the three sites (see Figure 2). To compare the consumption of a selected type of beverages (zero-calorie beverages or sugary beverages) under different `intervention` categories, boxplots with sales in count 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.

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

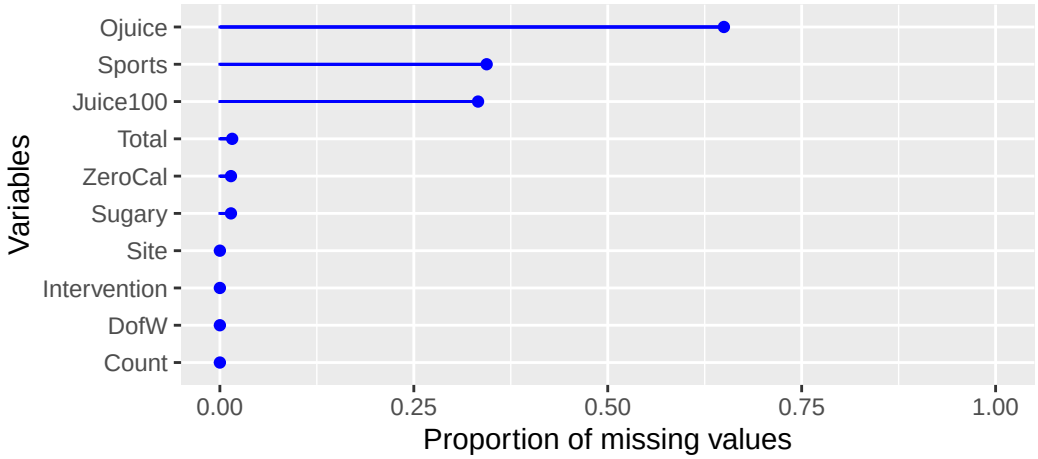

Figure 1: The proportion of missing values for all variables in the dataset.

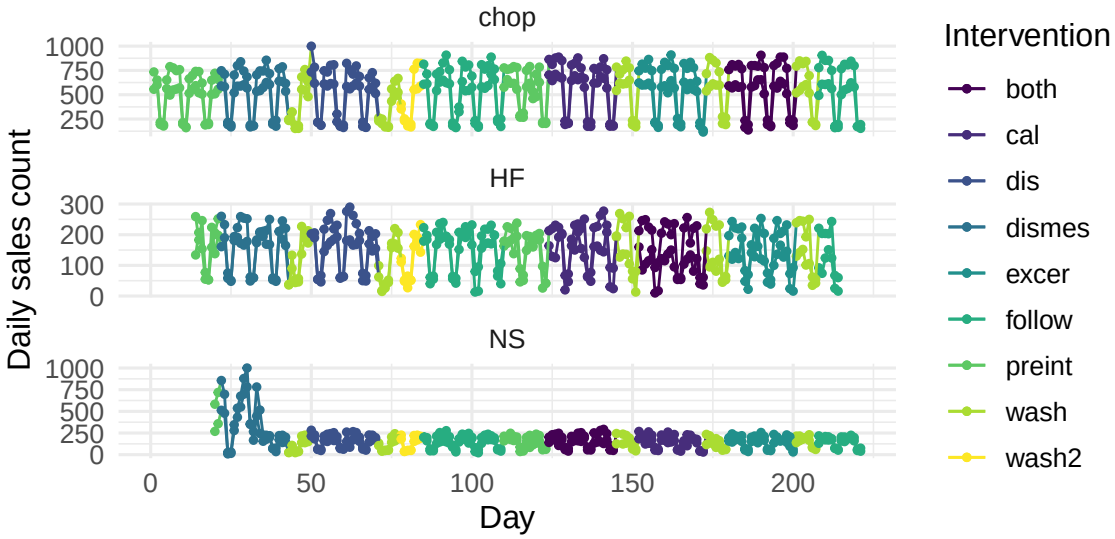

Figure 2: Daily count of zero-calorie beverages sales over time for each of the three sites.

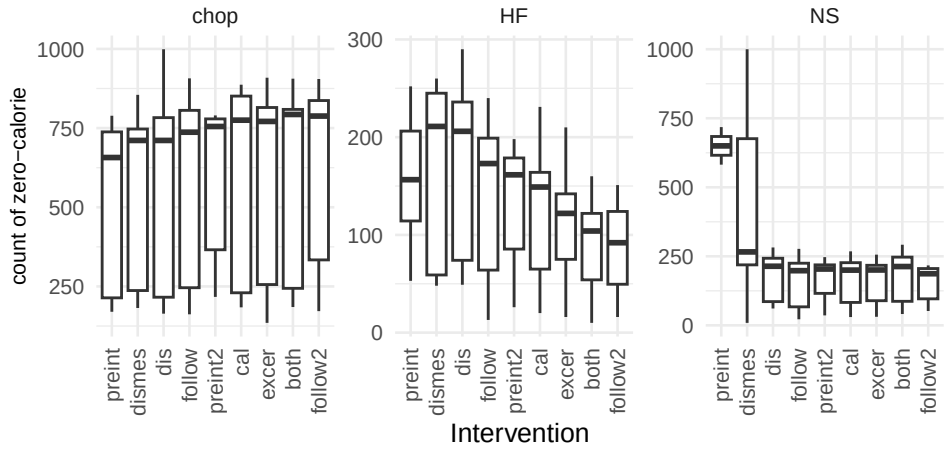

Figure 3: Boxplots of daily percentage of zero-calorie beverages sales for each of the three sites.

Moreover, histograms can be utilized to illustrate the overall distribution of count variables. For instance, Figure 4 reveals that beverage consumption exhibits a bimodal shape. This visualization assists in identifying potential overdispersion within the count data, guiding the selection of a more suitable model for further analysis.

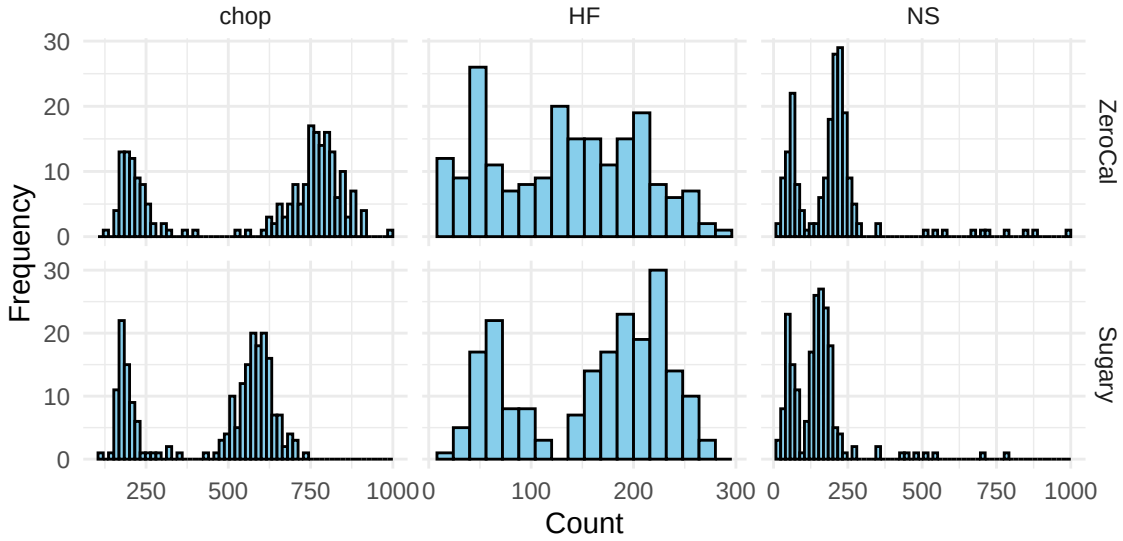

Figure 4: Daily count of beverages over time for each of the three sites.

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 linear mixed models (GLMMs), which introduce random effects to the corresponding generalized linear model and are applicable to count longitudinal data. The individual-specific random effects in GLMMs can incorporate the between-individual variations and capture the correlation in the repeated outcomes over a period of time, which aligns with the analytical requirements and the inherent characteristics of the data The major assumption of GLMMs is the normality distribution of random variables.

The count of zero-calorie beverages and sugary beverages are response variables. This count data directly aligns with the objectives to quantify changes in consumer behavior by assessing the actual volume of beverage sales under different interventions. Regarding the predictor variables, we recommend using `Intervention` and `DofW`, where `Total` of all beverages can be added as offset.

It is recommended to fit GLMM by introducing the random effects to sites, because the considerable differences can be seen (Figure 2) in the baseline of beverage consumption among the three sites, and the random effects can account the variability between different sites. Another critical aspect that warrants further consideration in GLMMs is the selection of the link function. For modeling count data, the log link function is commonly used with the Poisson or Negative binomial distribution (Yirga 2020). The log link ensures that the predicted counts are always positive. The choice can be theoretically justified based on the distribution of the response variables being modeled. Figure 4 in EDA shows there is potential overdispersion in the data, which means the Negative binomial could fit the nature of data better. Moreover, in this case, model fit statistics (e.g., AIC) are suggested to determine the most suitable formation of model. Based on the results of the optimal model, the estimated coefficients of different interventions can be used for intervention effect assessment and comparison. On the other hand, the estimated random effects of each site can illustrate the affects caused by location variability. Differences in the effects of interventions across sites can be accounted for by incorporating an interaction term between site and intervention in an alternative model. Significant coefficients of this interaction term can illustrate the extent to whether the effects of interventions vary by site, offering insights into how site-specific factors influence the effectiveness of interventions.

One limitation of this approach is the challenge in statistically identifying the differences in intervention effects across different sites. Nevertheless, by examining the random effects across sites, we can discern statistically significant differences in the impact of certain interventions at some sites but not at others. This analysis allows us to understand the variability in intervention effectiveness due to site-specific characteristics. On the other hand, whether the data fits the model's assumptions well also needs to be cautiously verified.

5 Conclusions

It is recommended to use daily count of zero-calorie and sugary beverages as response variables for the analysis, and `Intervention` and `DofW` should serve as the predictor variables. 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 GLMM with site as random effects to capture the variability across sites, it allows us to account for the differences in the effects of interventions across sites, by introducing the interaction term to the model. According to the results of statistical analysis (see Appendix), the Negative binomial GLMM has a better fit for this dataset than Poisson model based on the AIC score. The results indicate that only the combination of discounts and messaging effectively enhances the consumption of zero-calorie beverages. The data also reveal that posters displaying both calorie and exercise information tend to deter consumption. The model incorporating the interaction term between site and intervention demonstrates that the effects of interventions vary across sites. Overall, the model fits the data well, but some assumptions and conclusions still need further verification.

References

Yirga, Melesse, A. A. 2020. "Negative Binomial Mixed Models for Analyzing Longitudinal CD4 Count Data." *Scientific Reports* 10 (16742).

Statistical Appendix

In this appendix, we mainly focus on implementing the suggested models to approach the proposed statistical questions, and also provide detailed interpretations for the results.

5.1 Model Construction

To develop a suitable GLMM for this dataset, the primary consideration should be the selection of random effects. The dataset clearly shows significant variability in beverage consumption across the three sites, indicating that incorporating a random intercept for each site in the model is a reasonable approach, which allows the model to capture the variation in beverage consumption attributable to differences among the sites. Another consideration is setting random slopes for intervention within site to allow the effect of interventions to vary across sites, e.g., some interventions might be more effective in certain sites than in others, reflecting the unique contexts or characteristics of those sites. Moreover, the selection of the specific type of model also requires careful consideration. Both the Poisson and the Negative binomial distribution are commonly utilized in GLMMs when dealing with count data as the response variable. Due to the Negative binomial model's additional dispersion parameter, the estimates can be more reliable when the data exhibit overdispersion. However, determining which method aligns more closely with the nature of data requires further analysis. The AIC can be employed to select the most suitable model for the final analysis by providing a quantitative measure to compare the relative quality of different models. Generally, we suggest to construct models with the following formation,

$$\log(\text{Zero-calorie}_s) = \beta_0 + \beta_1 \cdot \text{Intervention}_s + \beta_2 \cdot \text{DofW}_s + \beta_3 \cdot \log(\text{Total}_s) + u_s \qquad (1)$$

where β_0 is the overall intercept, β_1 and β_2 are the fixed effect coefficients of `DofW` and `Intervention`, respectively, and u_s is the random effect for `Site`, indicating variation in the log count of zero-calorie that is specific to each site but not explained by `DofW`, `Intervention`, or `Total`. If the goal is to investigate the impact of interventions on the daily count of sugared beverage consumption, the response variable should be directly replaced with the daily count of `sugary` beverages.

5.2 Model Selection

In this section we evaluate the performance of Poisson model and Negative binomial model based on the AIC score. The results in table A1 illustrates the AIC of Negative binomial model is significant lower than that of Poisson model, which implies Negative binomial provides a better fits, likely due to potential overdispersion within the count of beverage. Thus, it is suggested to use Negative binomial as the final model for further analysis according to this criterion.

Table A1: AIC score under different link functions

Model	AIC
Poisson	14875.4
Negative Binomial	6916.3

5.3 Results analysis

By fitting the Negative binomial GLMM for daily count of zero-calorie beverages, the estimated fixed effect coefficients are shown in table A2. The interventions are incorporated as fixed effects in the model to quantify the average impact of various interventions on beverage consumption. The results indicate that only the combination of discounts and relevant explanatory messages significantly promotes the consumption of zero-calorie beverages, comparing to the baseline (pre-intervention). Conversely, providing calorie information alongside exercise details appears to have a somewhat negative effect on consumption. Meanwhile, the most of coefficients of `DofW` are not significant, indicating that the daily consumption of beverages is not strongly associated with the day of the week.

Table A2: The fixed effect coefficients of zero-calorie

Intervention	Estimate	Standard Error	p-value
Discount	0.085	0.048	0.075
Discount & Explanatory messaging	0.328	0.048	0.000
Calorie information	-0.081	0.047	0.090
Exercise information	-0.086	0.049	0.078
Calorie & Exercise information.	-0.137	0.048	0.004

On the other hand, site is treated as a random effect on intercept of the model, with an estimated variance of 0.105 and a standard deviation of 0.324, which indicates a moderate level of variation in the baseline counts of zero-calorie beverages across different sites. The similar models can developed for the count of sugary beverages.

Another important statistical question is to determine if the effects of different interventions varying by sites. A potential approach to address this issue is by introducing site as random effects on interventions. However, models structured in this manner often struggle to achieve convergence, making it challenging to obtain reliable results. A viable alternative approach is to incorporate an interaction term between site and intervention, which allows for assessing how the effect of interventions varies across different sites. Specifically, we can construct an alternative model with the formation as

$$\log(\text{Zero-calorie}) = \beta_0 + \beta_1 \cdot \text{Intervention} \times \text{Site} + \beta_2 \cdot \text{DofW} + \beta_3 \cdot \log(\text{Total}) \qquad (2)$$

The estimates of each interaction pair of intervention and site are displayed in A3, where the interaction items with **chop** (the name of one of sites) are treated as reference category. The results show that the effects of the interventions significantly differ by sites, indicating that the impact of each intervention varies depending on the location, which could be related to the unique circumstances of each site.

Table A3: The interactions between intervention and site

Interaction	Estimate	Standard Error	p-value
HF ×Discount	0.296	0.098	0.003 *
NS ×Discount	-0.135	0.102	0.187
HF ×(Discount & Explanatory messaging)	0.203	0.098	0.003 *
NS ×(Discount & Explanatory messaging)	0.574	0.102	0.000 *
HF ×Calorie information	-0.125	0.099	0.204
NS ×Calorie information	-0.243	0.102	0.017 *
HF ×Exercise information	-0.225	0.100	0.025 *
NS ×Exercise information	-0.205	0.104	0.048 *
HF ×(Calorie & Exercise information)	-0.375	0.100	0.000 *
NS ×(Calorie & Exercise information)	-0.242	0.102	0.018 *

5.4 Model checking

After fitting the models, it's crucial to check the assumptions and evaluate the fit. First of all, it's essential to verify that the chosen Negative binomial GLMM adequately reflects the characteristics of the data, even though the AIC score suggests it provides a better fit. According to the outputs of model, the dispersion parameter for the Negative binomial distribution is 14, which indicates that there is considerable overdispersion within count data, and thus, the Negative binomial model is more appropriate than a Poisson model. On the other hand, we check the normality assumption about residual, the Q-Q plot in Figure 4 suggests that the residuals mostly adhere to the normal distribution assumption, although there are some observable deviations. These deviations could be attributed to outliers within the data. The residuals versus the predicted plot indicates that the model might not be fully adequate for the data because deviation between actual and predicted quantile, which can be further investigated and improved. As for the model (2), the Q-Q plot in Figure 5 indicates that the residuals exhibit a significant deviation from the normality. Moreover, the patterns observed in the residuals versus the predicted plot suggest that the developed model could be inappropriate for the data.

5.5 Discussion

In this appendix, we provide the detailed procedure for developing GLMMs and analyzing the results. The selection with AIC score indicates that Negative binomial distribution has a better fit for our data than Poisson. The model outputs reveal that only the combination of discounts with messaging effectively promotes the consumption of zero-calorie beverages. Furthermore, posters containing both calorie and exercise information appear to have inhibitory effects on consumption. The results of model with interaction term of site and intervention illustrates the effects of interventions varying by sites as we expected. However, this model has bad performance on assumption checking, the drawn conclusions should be further determined and verified.

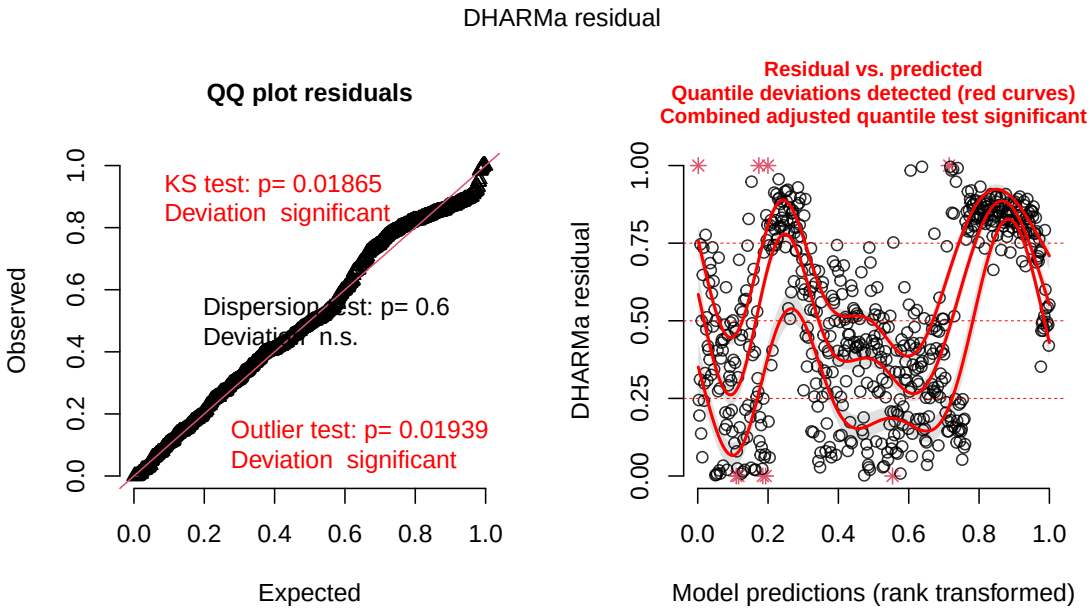

Figure 5: Diagnostics for model with site as a random effect on intercept.

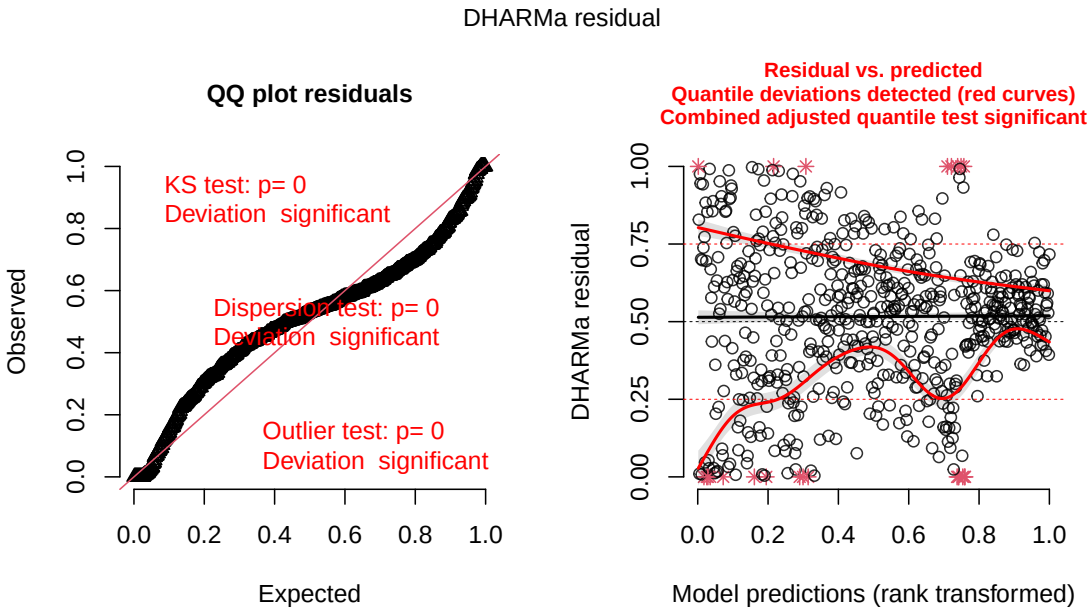

Figure 6: Diagnostics for model with interaction between site and intervention.

5.6 Contributions:

- Yian: wrote the EDA and formal analysis for Version 1;
- Wakeel: revised the sections before formal analysis for Version 2;
- Haibo: revised the formal analysis and wrote the statistical appendix for Version 2.