MA678 Midterm Project Report

Fan Feng

2020/12/8

Abstract

Dianping is one of the most popular applications in china for rating and reviewing restaurants. The overall rating, provided by the Dianping platform, is an important reason for deciding whether costumers choose this restaurant. However, we have little understanding of the mechanism behind the ratings. In order to better use this application, our project aim to analyze the key factors influencing the overall rating based on the datasets crawled from Dianping website. We selected restaurants belong to ten different food styles. Thus, we developed multilevel linear mixed regression and multilevel ordinal regression for analysis with food style as random effect. According to our results, we found that some feature have strong positive relationship with the rating while some predictors are not that corrlated with ratings. We also performed the model comparison and model validation in the end.

Introduction

A. Background and Goal

As the most popular platform for reviewing restaurants, dianping is important in our daily life. I am the deep user of dianping as well. This application can help people explore and choose the restaurants that is more suitable for them from a holistic view. Overall Rating, computed by the dianping platform, is often one of the most important and most convenient reference standard for people to choose a restaurant. However, everyone has different evaluation criteria and we need more specific understanding of the mechanism behind the rating numbers. Thus, it is meaningful to analyze what the overall rating of a restaurant represent in details. We can

Our goal of this project is to analyze the factors that affect the ratings of the restaurants located in Shanghai, and the degree of influence of each factor by using multilevel linear regression and multilevel ordinal categorical regression.

B. Dataset and Features

The Dataset, provided by Professor Yongfeng Zhang as json files, was initially scarched from the dianping website. The data contains the basic information of restaurants all over China, including the rating, reviews, flavor, service, environment, average cost, food style, district, time and so on. We transformed and cleaned the data by using the methods from tidyverse in R because of some unknown problems of the raw json files.

In our project, in order to focus on the problem of our interest, we only select the businesses located in Shanghai city and take ten styles of food into consideration.

In addition, according to the rules in dianping platform, we removed the samples with rating equals zero because zero rating means that dianping platform has stopped providing service to these restaurants. For

better modeling, we performed the feature extraction and add some variables by applying the string detect methods to the contents of review, such as wifi, delivery, free_parking and so on.

In the end, we have 5746 samples of restaurant in total and 12 features included for each restaurant. Each restaurant selected belongs to one certain food style and we have ten food in total. The numbers of restaurants from different food styles are different according to Figure 1.

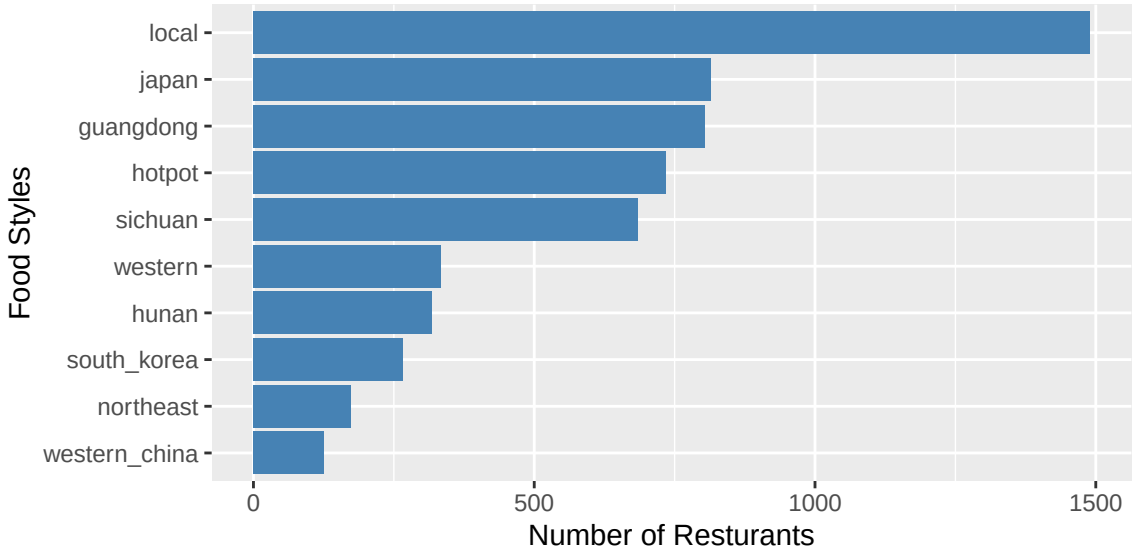


Figure 1: The barplots of the numbers of restaurants from different food styles in descending order

Method

We considered two types of models based on the two ways of treating the rating variable: continuous and ordinal categorical. For the continuous condition, we apply the multilevel linear regression; for the latter one, we make use of the multilevel ordered categorical regression.

For both models, we choose flavor, environment, service, share, avg_cost, wifi and free_parking as predictors. To be specific, among all the predictors, only wifi and free_parking are binary(0 means no and 1 means yes). In order to better interpret the main effects, the binary predictors are left as they are and the continuous predictors are linearly transformed by centering and standardizing.

Method 1: Multilevel Linear Regression

To begin with, considering the multilevel structure of the data, I firstly apply the multilevel linear regression, by setting the food_style variable as random effect.

However, as we all know, the outcome variable of linear regression should be continuous rather than discrete. In our data, even though we have 7 levels of rating, we can treat the rating of restaurants as continuous because the overal rating is actually the rounded value calculated from all the ratings and reviews from different costumers. In order to compare multilevel linear model with other methods, we can discrete the predictions from the multilevel linear model to the nearest half star. In this case, as Andrew Gelman said, it worth a try to treat ratings as continuous outcome variable.

Method 2: Multilevel Ordered Categorial Regression

If we treat the rating of restaurants as the ordinal categorical variable, considering the multilevel structure in our data, we can naturally apply the ordered ordered categorical regression with random effects. We realized this method by using the cumulative link mixed models(CLMMs) in 'ordinal' package[] in R. The maximum likelihood based on Newton-Raphson method is used to estimate the model parameters.

When fitting the model, it should be pointed out that the model has the proportional odds assumption which can simplify the model. Under this assumption, the effect of predictors are the same for each increase in the level of the outcome variable.

Result

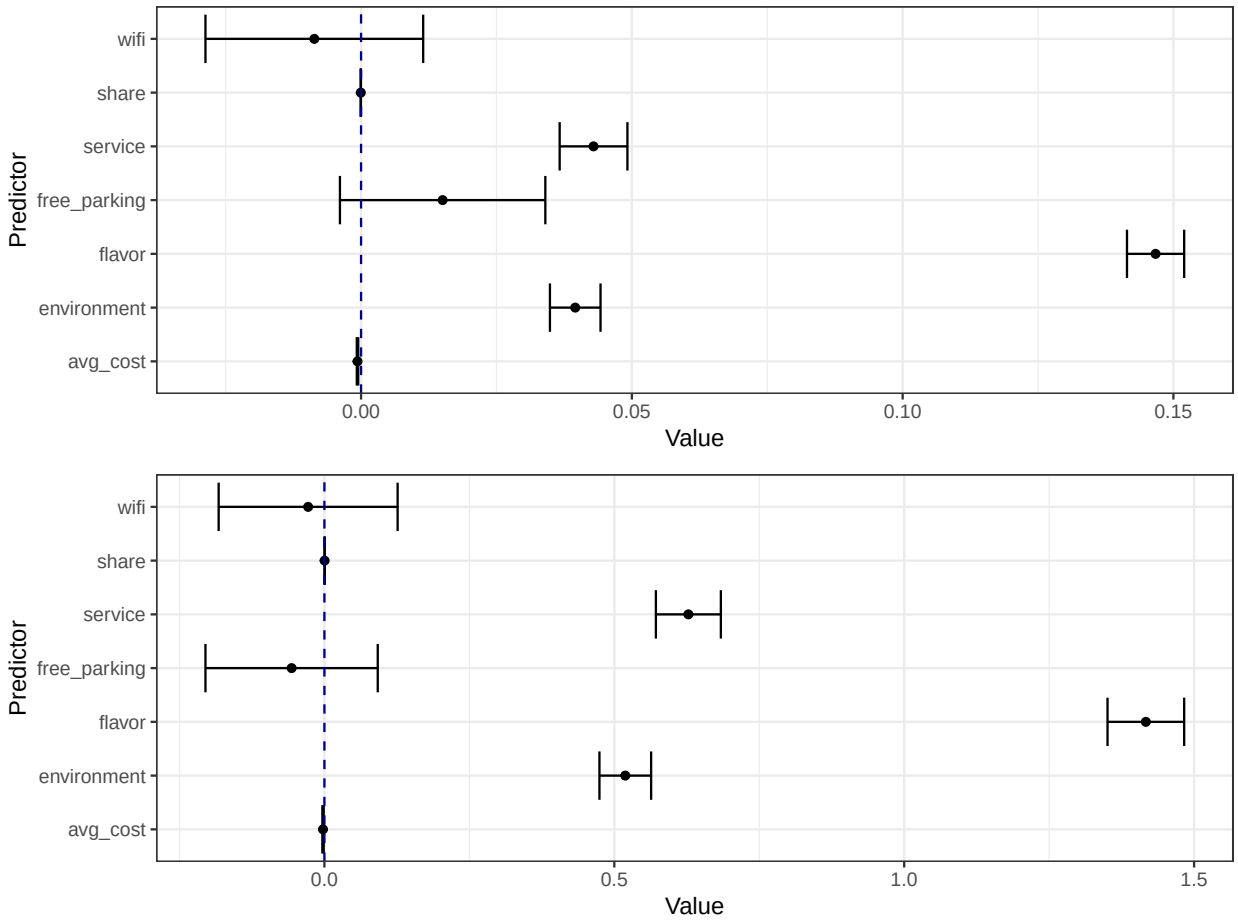


Figure 2: The error bars of the the Multilevel Linear Model

A. Model Comparison and Estimation

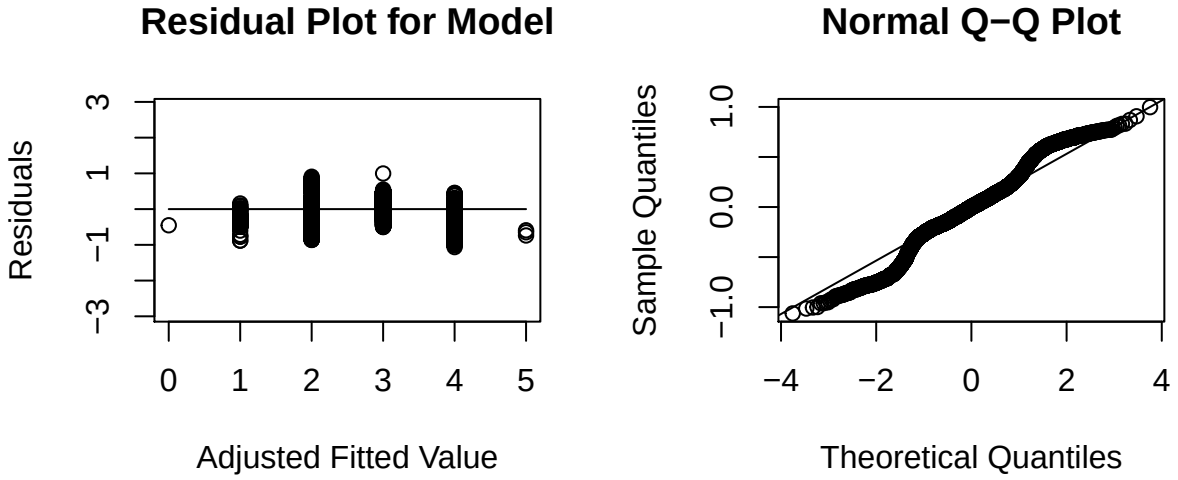
As is shown in Figure 2, the estimation of each predictor from two models is relatively similar, even though we treated the respond variable in a different way. It should be pointed out that the x axis of two error bars have different meaning because of the different link functions used.

In terms of the estimation, we can find the estimated coefficients of each variable and the 95% confidence interval. As predictors, Service, environment and flavor have strong evidence of positive effect on the ratings. We can find that the score of flavor is the strongest effect on rating of a restaurant. While the predictors wifi and free_parking do not have strong evidence of the effect on the rating. In addition, the avg_cost is seems to have some evidence of showing very slightly negative effect on the rating.

B. Validation

Firstly, if we look at the random effects, we find that the variation between different food styles is very close to zero, which means that the correlation with each food style is very small. However, we did not remove the random effect even thou it is not that significant.

In addition to that, we made the residual plot and Q-Q plot of the linear mixed model. I should point out that the fitted is adjusted by rounding to the closest half value. Then we can find that there are typical pattern in the residual plot when we modeling the multinomial outcome as the continuous variable, but in general, the residuals for each fitted value is evenly distributed around 0. Also the Q-Q plot shows that the assumptions about the normality of errors is statified in general.



We also want to show the binned residual plot for the second model, but it seems to be difficult because the clmm method we used is still under development. However, we can try to plot the main predictors verse the ratings based on the multilevel ordinal model to gain some intutive insights from the model.

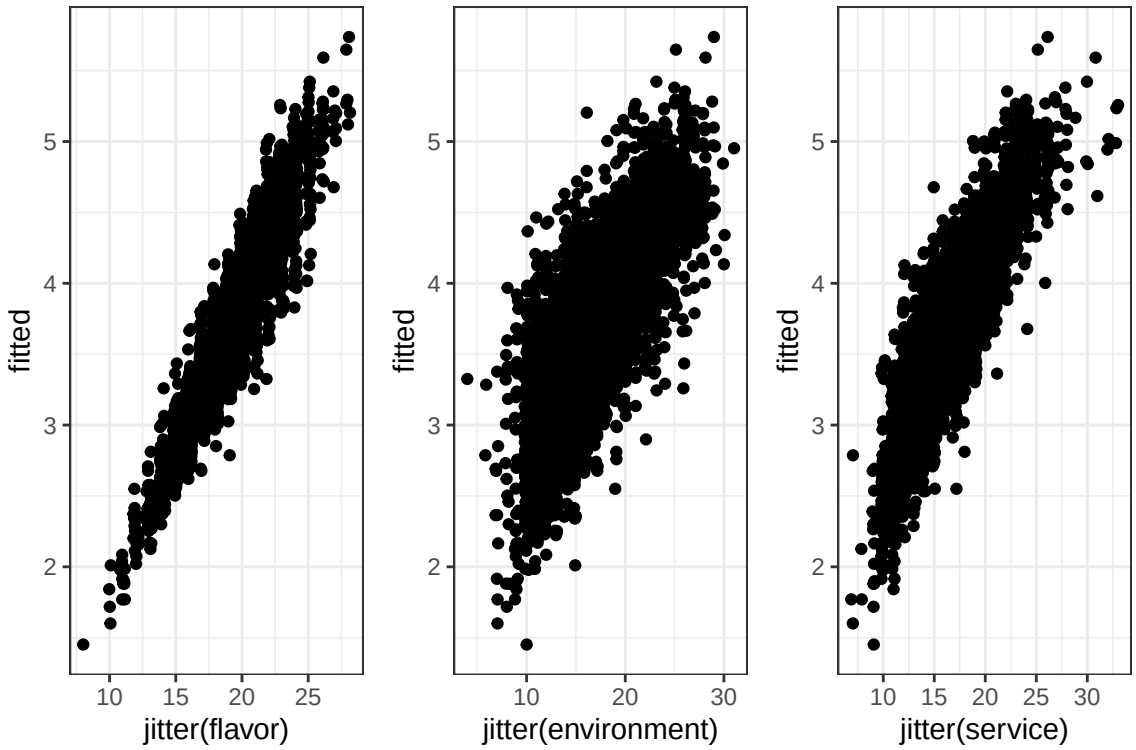


Figure 3: Fitted value of ratings verse the predictors with main effects.

Discussion

In the end, we can answer the question of this project: what are the factors influence the rating of a restaurant which is calculated based on the mechanism of Dianping?

Based on the whole analysis, we can say that the score of flavor has the most important positive effect on the rating in our data. The score of environment and service is also important with postive effect but not as important as flavor. To our surprise, the average cost of the restaurant has a very weak negative influence of the rating. It probably means that in shanghai, which is the biggest city in china, people are not very sensititive to the cost of most restaurants belong to the ten food styles.

Also, we find that the free_parking and wifi service is has no significant effect on the rating. It shows that the things that the costumer really care about is the experience around the food.

Bibliography

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu and Shaoping Ma. Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis. In Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2014), July 6 - 11, 2014, Gold Coast, Australia.

Bates, Douglas; Maechler, Martin; Bloker, Ben; Walker, Steve (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48.

Goodrich, Ben; Gabry, Jonah; Ali, Iamd; Brilleman, Sam (2018). "rstanarm: Bayesian applied regression modeling via Stan." R package version 2.17.4, http://mc-stan.org/.

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse Heenan, Adel; Williams, Ivor; Acoba, Tomoko; DesRochers, Annette; Kanemura, Troy; Kosaki, Randall;

Nadon, Marc; Brainard, Russel (2017). Long-term monitoring dataset of coral reef fish assemblages in the western central Pacific. figshare. Collection. https://doi.org/10.6084/m9.figshare.c.3808039.v1