Review Website DianPing

Zijie Huang

2020/12/9

1 Abstract

DianPing is a widely used review website. Customers give total score, flavor score, environment score and service score to restaurants. Total score is the most important criterion when people are selecting restaurants. We developed a mutilevel ordinal logistic model to assess how flavor score, environment score and service score attribute to the total score. Results showed that all ratings have positive effect on total score and flavor rating attributes most. Results also showed that rating effects vary between cities. Including more data may improve the model and give us preciser estimations.

2 Introduction

DianPing is a famous Chinese restaurant review website. Consumers can make comments on restaurants in *DianPing*. Consumers are allowed to give four scores to restaurants, namely total score, flavor score, environment score and service score. Based on all these reviews, *DianPing* calculates these four scores for each restaurant. In the thumbnail page of each restaurant, only the total score is shown to the customer. One restaurant may get a very high total score because of its high flavor score, even this restaurant has a very low environment score. In this report, we seek to investigate how flavor rating, environment rating and service rating affect the total score. We use a mutilevel ordinal logistic model to investigate how much does each rating attribute to the total score while accounting for differences between cities.

This dataset is downloaded from **The Wise Lab Dataset, Rutgers University**. This dataset is crawled from **dianping.com**. It contains detailed business meta data information of 338026 restaurants in 2011.

3 Methods

3.1 Data Cleaning and Selection

Since all data were stored in the format of *json_line*, we cleaned the data using the tidyverse package in R. After data cleaning, we mananged to get a well-organized dataset. We deleted all meaningless data that have zero value for the total score. Then, our dataset included 89693 restaurants across 246 cities. However, it is very difficult and time-consuming to fit such a large dataset to our model. Therefore, we decided to shrink our dataset. We wanted to focus on major cities that have more restaurants. Therefore, we first deleted all the cities that have less than 100 restaurants. After that, our dataset included 86717 restaurants across 46 cities. Then, we sampled 4000 data from our dataset. Our final dataset included 4000 restaurants across 46 cities. In order to make sure that our sample does not lose the features of orginal dataset, we compared the distribution of four ratings before and after sampling. They gave us similar distribution. Results are shown in *Appendix: Figure 4 and 5*. Below is the description for each variable.

Variable	Description	Remark
restId	An unique seven-digits ID for each restaurant	
cityId	An unique ID for the city this restaurant locates	

Variable	Description	Remark
score	The overall rating of the restaurant	0 - 5 (min division value: 0.5)
flavor	The flavor rating of the restaurant	0 - 40 (integer)
environment	The environment rating of the restaurant	0 - 40 (integer)
service	The service rating of the restaurant	0 - 40 (integer)

3.2 Model Selection

Dependent variable, the total score ranges from 0 to 5 with minimum division value to be 0.5. We decided to use multilevel model. There are two reasons.

1. For the rating of restaurants, it is reasonable to consider it as close to complete pooling. Since for restaurants in different cities, they will not differ that much. Multilevel model is effective in this case and it allows variation between cities.

2. For many cities, they have very few restaurants. The number of restaurants in each city is shown in *Appendix: Figure 6*. It would be difficult to model using classical regression. Multilevel model can give us reasonable estimates even for cities with small sample sizes.

We first tried multilevel linear model. Our predictors are city ID, flavor rating, environment rating and service rating. However, the parallel lines in residual plot indicated that treating total score as a continuous variable is not suitable. Results are shown in *Appendix: Figure 7*. Therefore, we decided to treat total score as an ordered categorical variable and used multilevel ordinal logistic regression model. We ran Bayesian model using *brms* in R to build this model. We compared models with and without certain variable to check the proportional odds assumption. We assessed the model using Bayesian posterior predictive checking.

4 Results

4.1 Estimations and Assumptions

Below is the model fitted in R.

$$brm(score \sim (1 + flavor + environment + service|cityId), data = rest\_sub, family = cumulative("logit"))$$

We allowed all variables to vary between cities. In *Figure 1*, we can see the estimates with one standard error confidence intervel for flavor rating, environment rating and service rating. All of the intervals lie on the right hand side of zero. They all have positive value. Among them, flavor has the highest value. One unit increase in flavor rating increase the log odds of having total score to be in category higher than 0.5 versus category 0.5 by on average 1.11.

4.2 Predictions

We assess the model by Bayesian posterior predictive checking. *Figure 2* shows the overlaid density estimates of the data and the replicates. If our model fits the data well, we would expect that the data and the replicates will overlap perfectly. From the result, we can see that they almost overlap perfectly. Our model captures the patterns very well. Though, we still lose some features in those peaks. The overall prediction performance is great.

5 Discussion

In general, we would expect that all three ratings have positive effect on the total score and flavor rating plays the most important role in attributing to the total score. The results in section 4.1 line up with our expectation. The results also show us that environment rating and service rating have nearly equal effect on the total score.

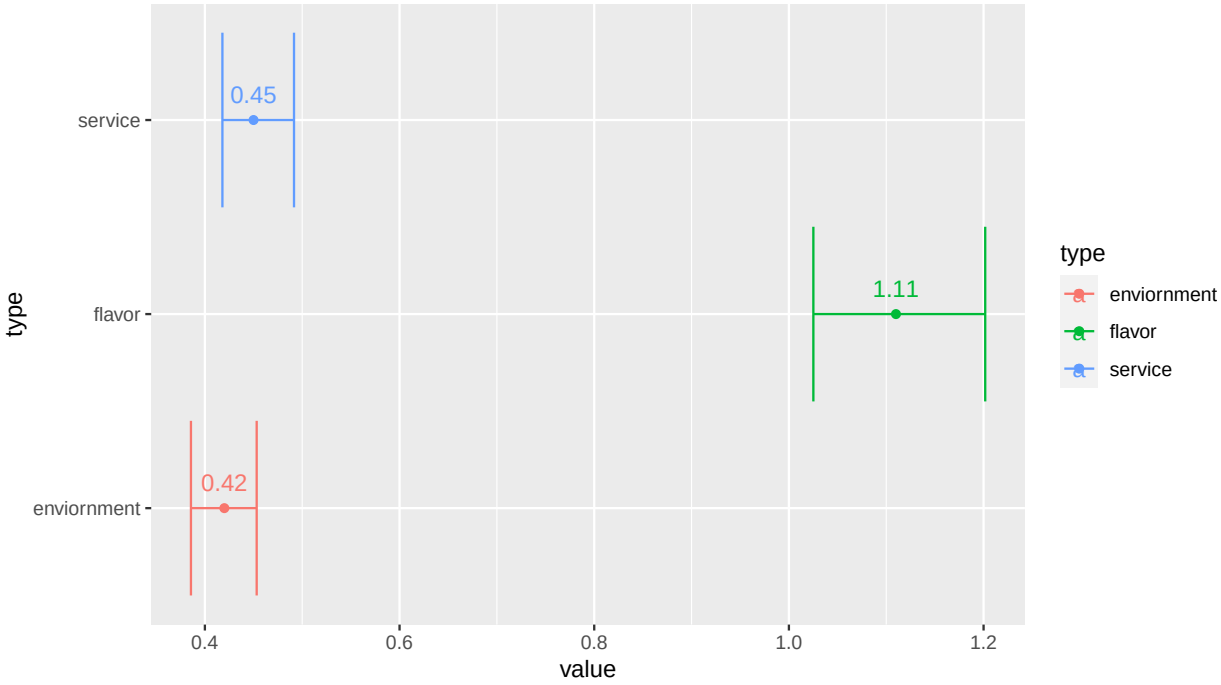

Figure 1: The predicted coefficients

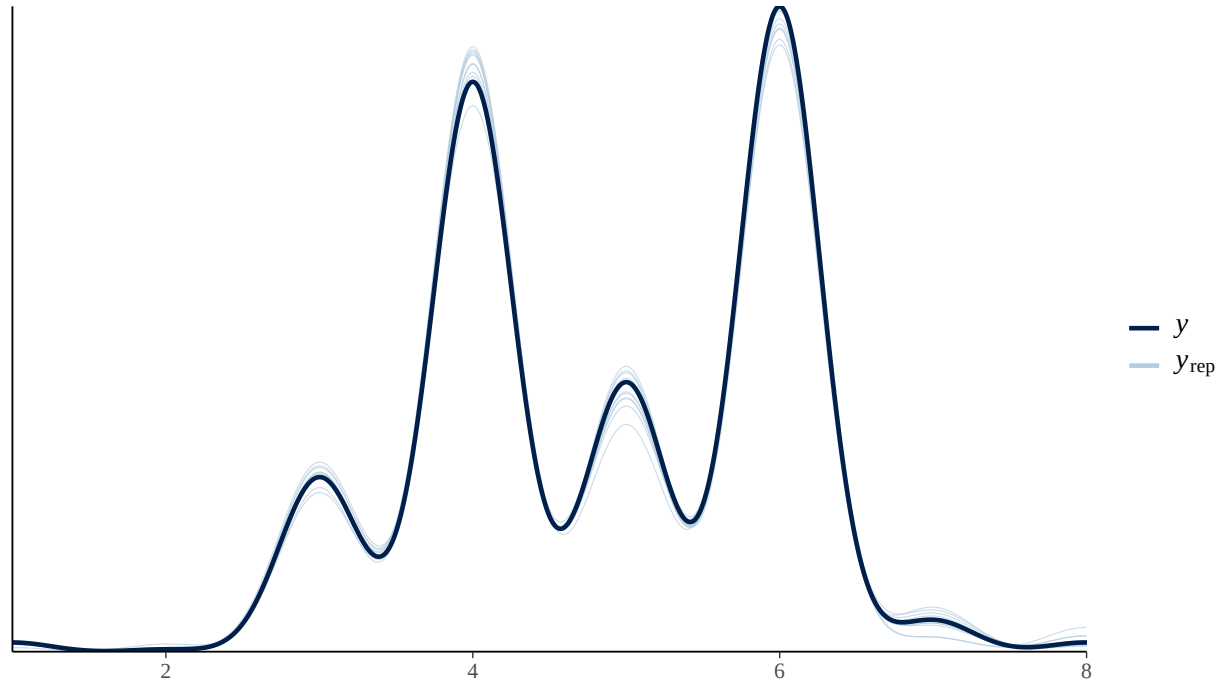

Figure 2: Density estimates of the total score data y and replicates

Then, we look at the city level estimations. Estimations in city level tell us that this rating effects vary between cities especially for flavor rating which has the widest confidence interval. We look at the cities that flavor rating has a relatively large effect on the total score. In *Figure 3*, cities are sorted by the coefficient of flavor rating. The top five cities are *Beijing*, *Chongqing*, *Changchun*, *Hefei* and *Xiamen*. In China, except for *Xiamen*, the rest of them are all provincial capital. These cities are well developed and have high standard of living. Higher value of coefficients may indicate that people in these cities think flavor is attributing more to the total score.

CityId	Flavor Coefficient
110	2.022756
9	1.820622
2	1.805768
70	1.800442
15	1.799524

Noticing that in order to fit a Bayesian model, we sampled the dataset. This may cause our estimations to be not that precise. Though it looks like that these three ratings are the most relevant variables, we can still include more variables to improve our model. For example, the type of restaurant and the favourable comment ratio.

6 Bibliography

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Paul-Christian Bürkner (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Yongfeng Zhang, Haochen Zhang, Min Zhang, Yiqun Liu and Shaoping Ma. Do Users Rate or Review? Boost Phrase-level Sentiment Labeling with Review-level Sentiment Classification. 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.

Gelman, A., & Hill, J. (2006). Multilevel linear models: The basics. In Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research, pp. 251-278).

Gelman, A., Hill, J., & Vehtari, A. (2020). Assumptions, diagnostics, and model evaluation. In Regression and Other Stories (Analytical Methods for Social Research, pp. 153-182).

7 Appendix

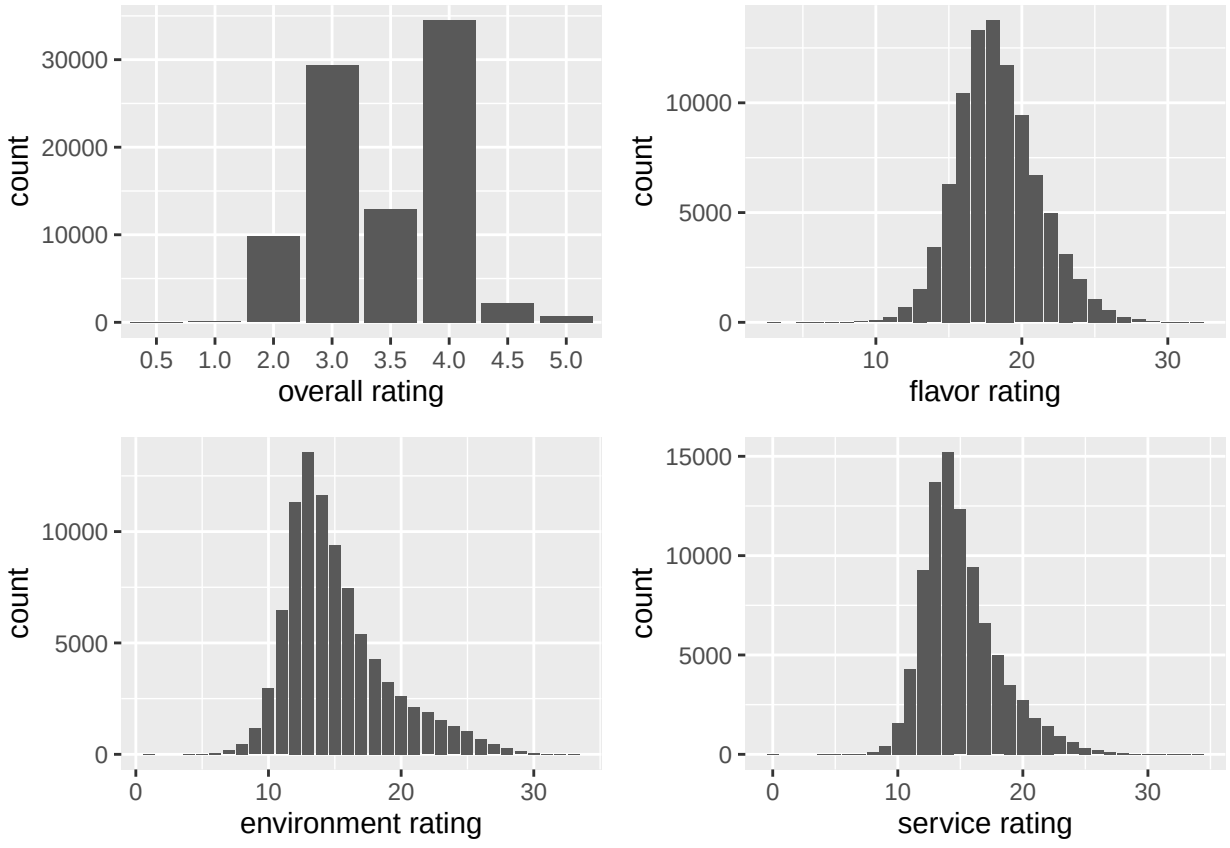

Figure 3: Distribution of overall rating, flavor rating, environment rating and service rating before sampling

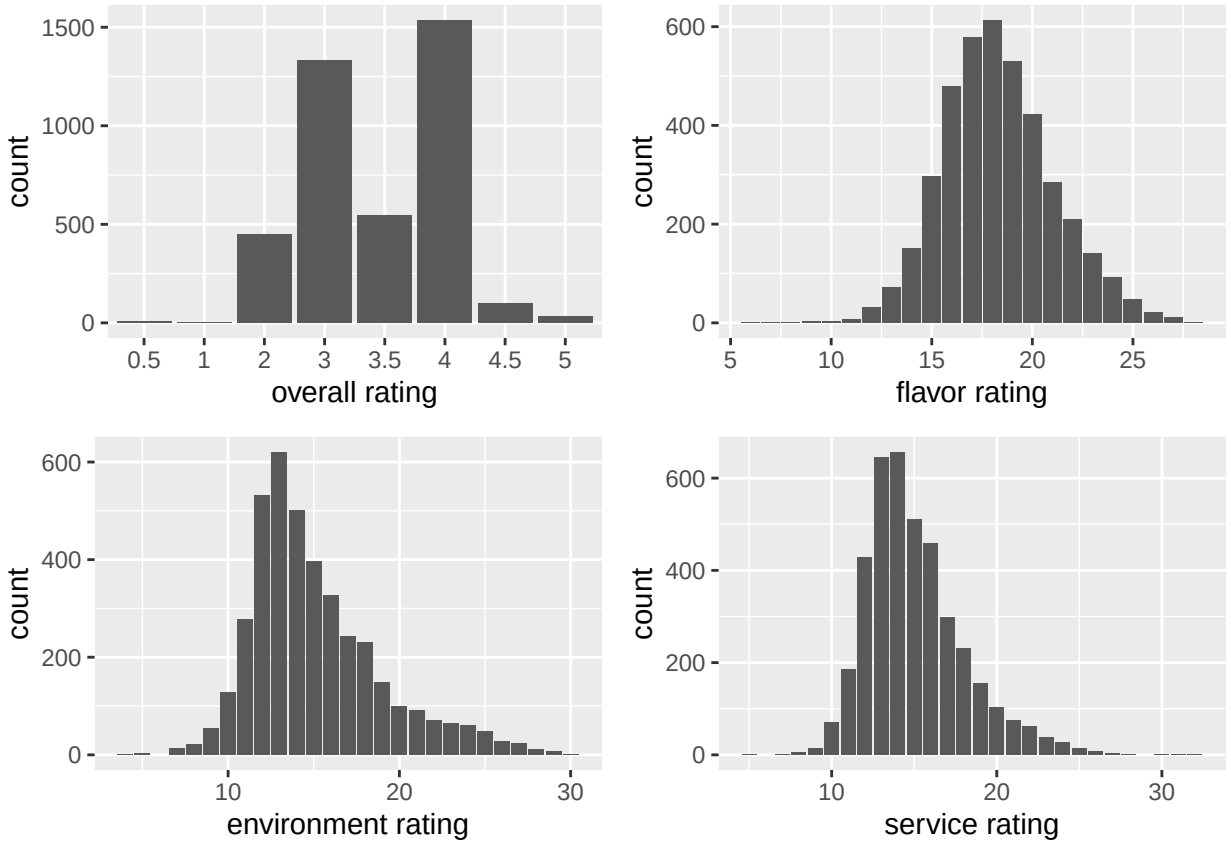

Figure 4: Distribution of overall rating, flavor rating, environment rating and service rating after sampling

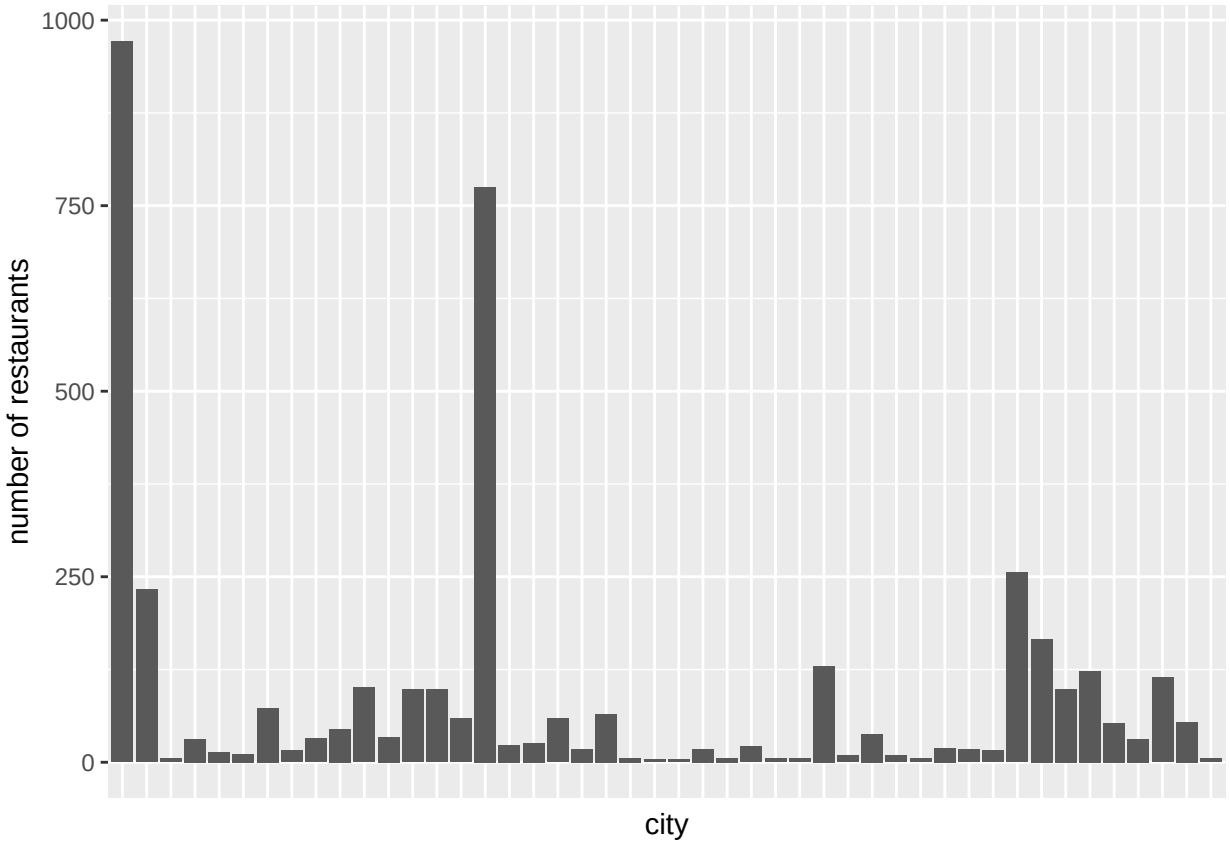

Figure 5: The number of restaurants in each city

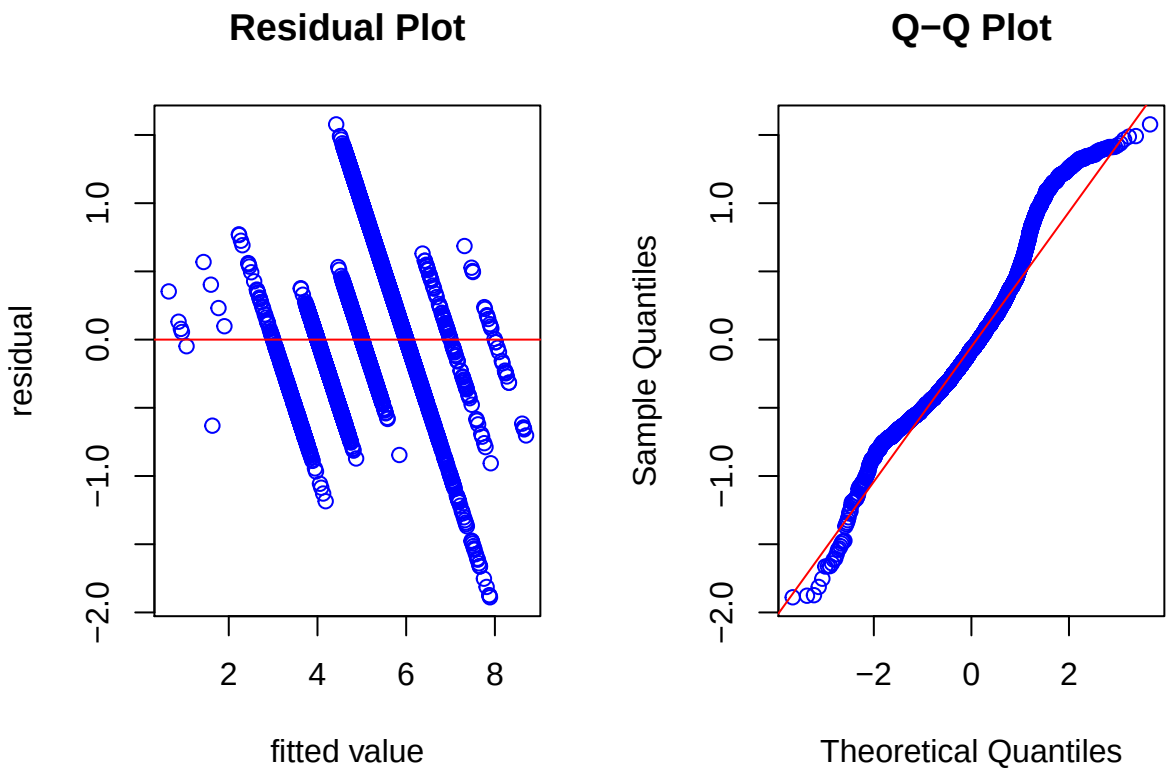

Figure 6: Residual plot and Q-Q plot of linear mixed effects model