MA684 Final Project - Yelp Data Challenge

Ziyan Li, Mark

December 15, 2016

Introduction

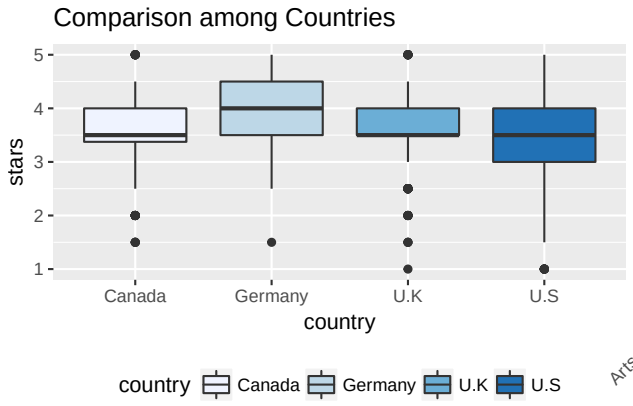
The Yelp challenge dataset is so profound that many topics could be addressed. This project will narrow down the scope to a particular research question and focus on the implementation of mixed effect model. The topic going to be discussed is the cultural differences among four countries (U.S, U.K, Germany and Canada) in terms of the preferences when people go to bars and enjoy their nightlife. More specificly, what will affect people's ratings and is such a impact varies among different countries?

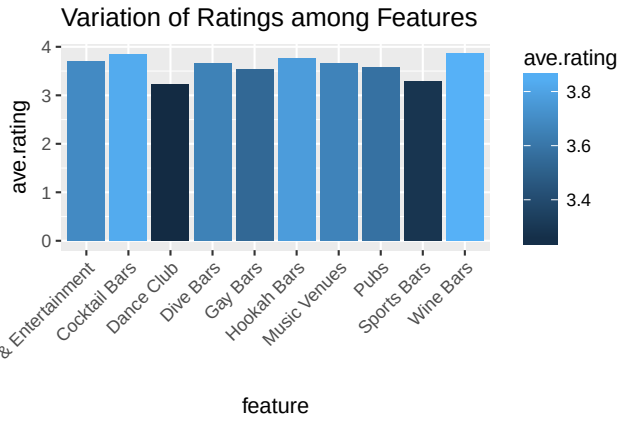
Data Cleaning and Recoding

Let us subset the business dataset by searching key words of "Bars" and "Nightlife" in categories, for the reason that it is the most effecient way to filter out the business that we are interested in. Merely using either one of them for filtering will cause some trouble, since we are not interested in over 800 sushi bars or a pizza shop opens at night. Additionally, the highly frequent co-occurence of "Bars" and "Nightlife" ensure that we don't lose much information and finally we captured 4742 businesses. Meanwhile, since we are interested in the cultural difference which makes more sense on a country level instead of a state or city level, we need to recode the location information into countries using: $U.K = \{EDH, ELN, MLN\}$, $Germany = \{BW\}$, $Canada = \{QC, ON\}$ and $U.S = \{PA, IL, NC, NV, WI, AZ, SC\}$. Recoding variables are not trivial in this case and I will address that in the Design Matrix section.

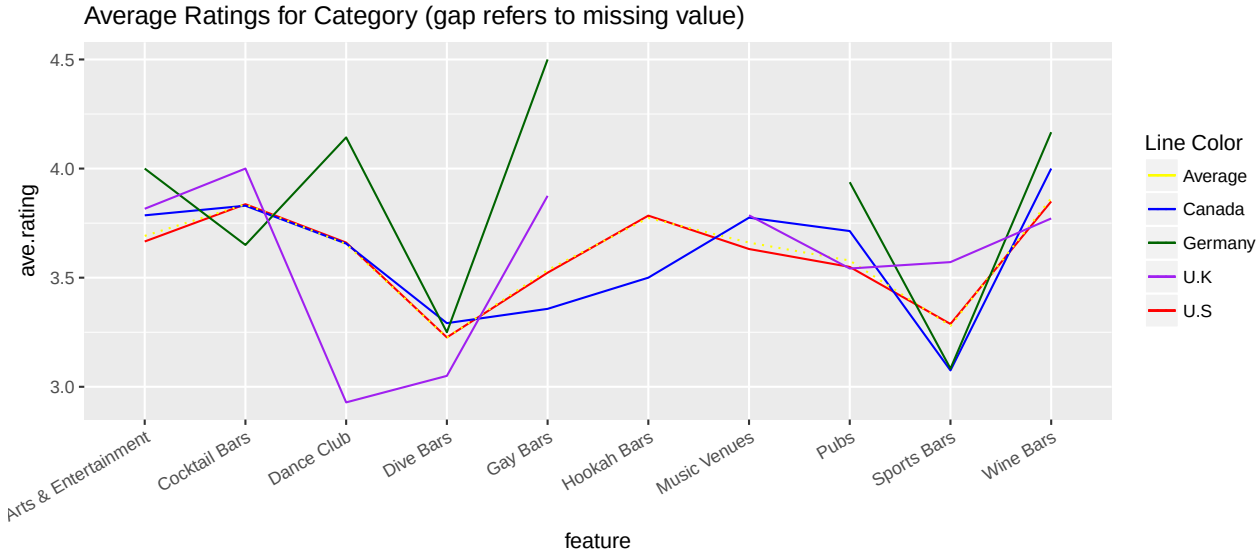
Exploratory Data Analysis

The exploratory data analysis aims to find the source of variation for the response variable which in this case is the business average rating. First of all, the variation might come from the categories that one business belongs to. From a business owner's perspective, will adding a certain category label to its business profile affect the expected rating and what category will it be? Here I select 10 features that we might be interested in and probably will affect the average rating from the top 20 most frequent ones. [see Appendix.1] The follwing two figures shows variation of ratings among countries and our selected features.

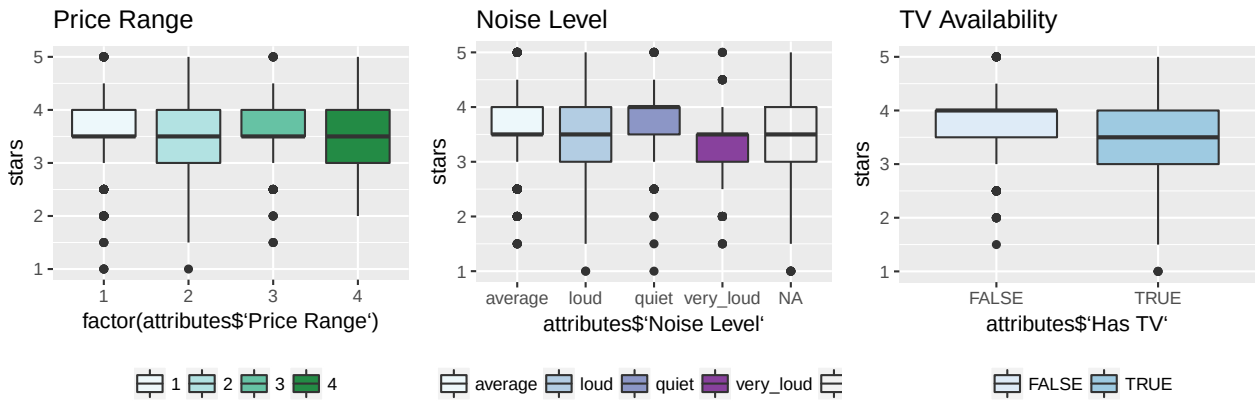




If we break it down, we will see more variation on the country level which might lead us to treat them as random effect in our model.



Besides, after exploration fro other variables,the following atributes might have a fixed effect to the average business rating and the variation is visualized in the following.



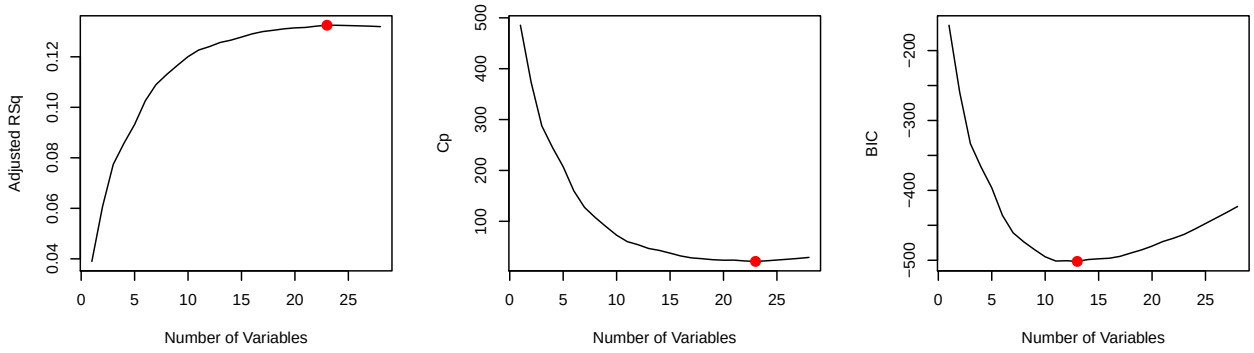
Design Matrix and Variable selection

1.Design Matrix

One major challenge here is that there are lots of NAs since most of the information input are optional for those business owner. Simply omitting them is not reasonable and not necessary and it will dramatically shrink the number of obsercations. Thinking it carefully, in this case, these NAs are not missing value. Instead, they are more like baselines, as it refer to the situation where business owners are not willing to provide information. From a modeling perspective, adding a piece of information will probably increase the expected rating but it is possible to the other way around since we do not know the sign of the coefficient before we fit the model. Thus, "no response" will not affect the expected rating no matter what the coefficient will be. Therefore, the more reasonable way is to recode variables into several columns and treat NAs as baseline (i.e, 0), and construct a design matrix for modeling.

2.Variable Selection

As for variable selection, we can start with all the reasonable ones and use the best subset selection method as a reference. Here, I start with 10 category variables (mentioned previously), noise level, smoking availability, happy hour, live music, outdoor seating, has TV or not and price range. After recoding construct a design matrix with 28 columns and apply the best subset selection. Here is the result.



Let us take the model with minimum BIC with 13 variables, which seems good as it reduce the dimension of our data by half. The 13 variables as well as their coefficients from a linear model are shown below.

	round(coef(fit.full, 13), digit = 4)
(Intercept)	3.6649
Arts...Entertainment	0.1781
Cocktail.Bars	0.2590
Dance.Club	-0.2238
Hookah.Bars	0.1992
Sports.Bars	-0.2754
Wine.Bars	0.2340
noise.quiet	0.1706
noise.average	0.1168
noise.very_loud	-0.2425
smoking.yes	0.0880
dj.yes	-0.0831
TV.no	0.2139
price	-0.1044

Mixed Effect Model

1.Linear Mixed Effect Model

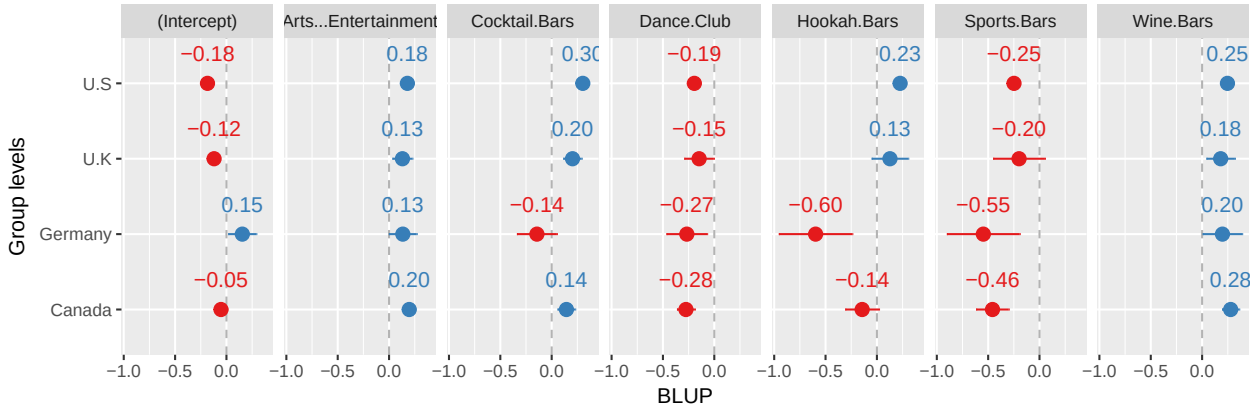
The model can be expressed as:

$$y_i = \beta_{0j[i]} + \beta_1 X_i^{noise.quiet} + \beta_2 X_i^{noise.average} + \beta_3 X_i^{noise.very_loud} + \beta_4 X_i^{smoking.yes} + \beta_5 X_i^{dj.yes} + \beta_6 X_i^{TV.no} + \beta_7 X_i^{price}$$
$$+ u_{1j[i]} X_i^{Arts.Entertainment} + u_{2j[i]} X_i^{Cocktail.Bars} + u_{3j[i]} X_i^{Dance.Club} + u_{4j[i]} X_i^{Hookah.Bars} + u_{5j[i]} X_i^{Sports.Bars}$$
$$+ u_{6j[i]} X_i^{Wine.Bars} + e_i$$

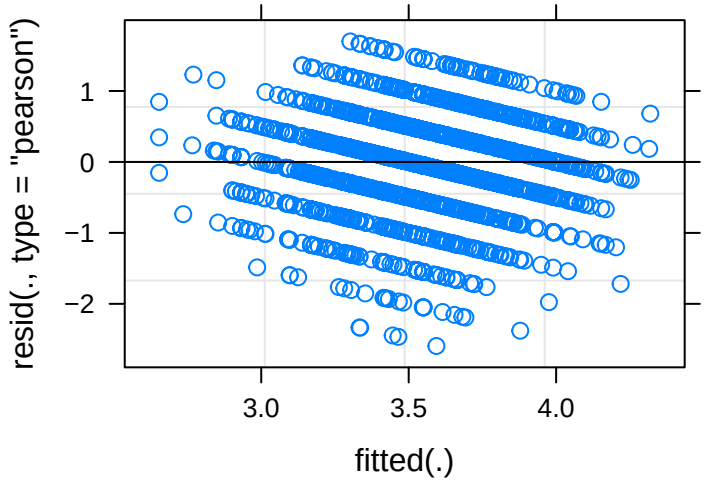
2.Interpretation

fixef(fit)	
(Intercept)	3.8268598
noise.average	0.1332457
noise.quiet	0.1836686
noise.very_loud	-0.2369519
smoking.yes	0.1009735
dj.yes	-0.0841427
TV.no	0.1927062
price	-0.1124450

The fixed effect is shown above, and the estimated coefficients are quite similar to those in the previous linear model. Generally, noise.average, noise.quiet, smoking.yes, and TV.no have positive effect while the rest have negative effect. Note that in this model, taking TV as an example, a TV.yes is equivalent to no information, which means that adding a "TV.yes" to one's business profile will not affect the expected rating. And interestingly, having a "TV.no" feature will even lead to a higher expected rating. Similarly interpretation can be made to other coefficients in the same way. Meanwhile, the random effect is shown in the following:



As we expected before, there seems some cultural differences in terms of people's attitudes towards diferent type of bars. In general, Germen are giving higher rating to bars compared to people from the other three countries, but cocktail bars and hookah bars seem not as welcomed in Germany. Besides, add a sports bar feature will lower your expected rating no matter where. Other interesting interpretations can be made in the same way.



3.Diagnostics

The pearson residual plot of our model looks okay, since we are fitting a continuous model to a ordinal variable. But there is no obvious weird pattern except the spread increases in the middle, which is reasonable as most ratings fall into this part. We can also look into other diagnostics such as the correlation matrix for fixed effect [see Appendix.2] and qq-plot for random effect [see Appendix.3]. The correlation matrix does not suggest any high correlation for fixed effect because we did not include all the recoded dummy variables. The qq-plot looks okay, too. All the points fall into both sides of the q-norm line and the assumption for normality of random effect is not violated.

4.Further Discussion

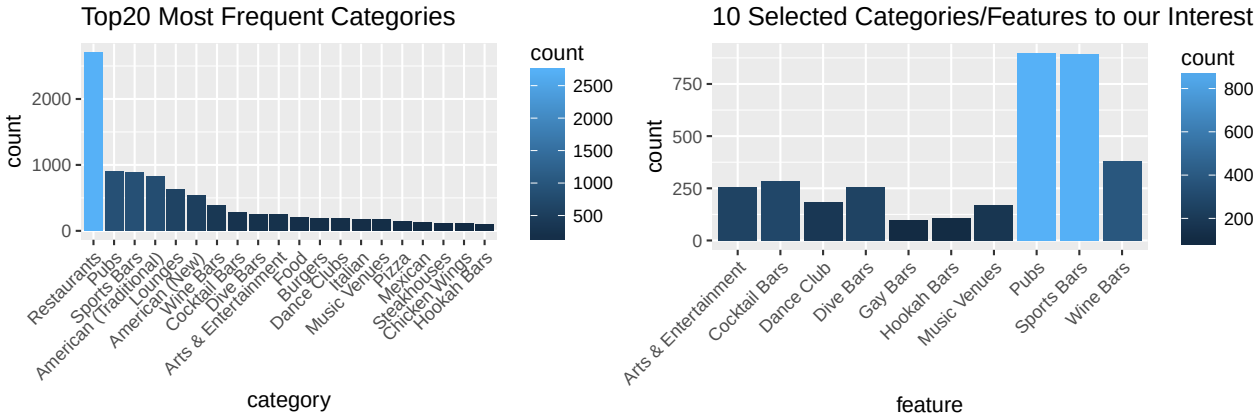
Though unable to find a pakage that can be used to fit a multinomial fixed effect model with multiple group level predictors, the clmm2{ordinal} allows to fit a cumulative link mixed model with one random effect. In the appendix [see Appendix.4], I tried different multinomial mixed models and run a likelihood ratio test to confirm that both the category variables and the country term are significant. In addition the AIC for the multinomial mixed effect model is quite large compare to the linear mixed effect model due to the absence of group level predictors, which in a sense support the significance of the random effects. Besides, I did not use the user data and the review data in this project. It is pretty likely that user will bring much variation to our model which might lead to a totally different result. Hence, in the future, it will be good to look at these two data set and conduct some further analysis.

Conclusion

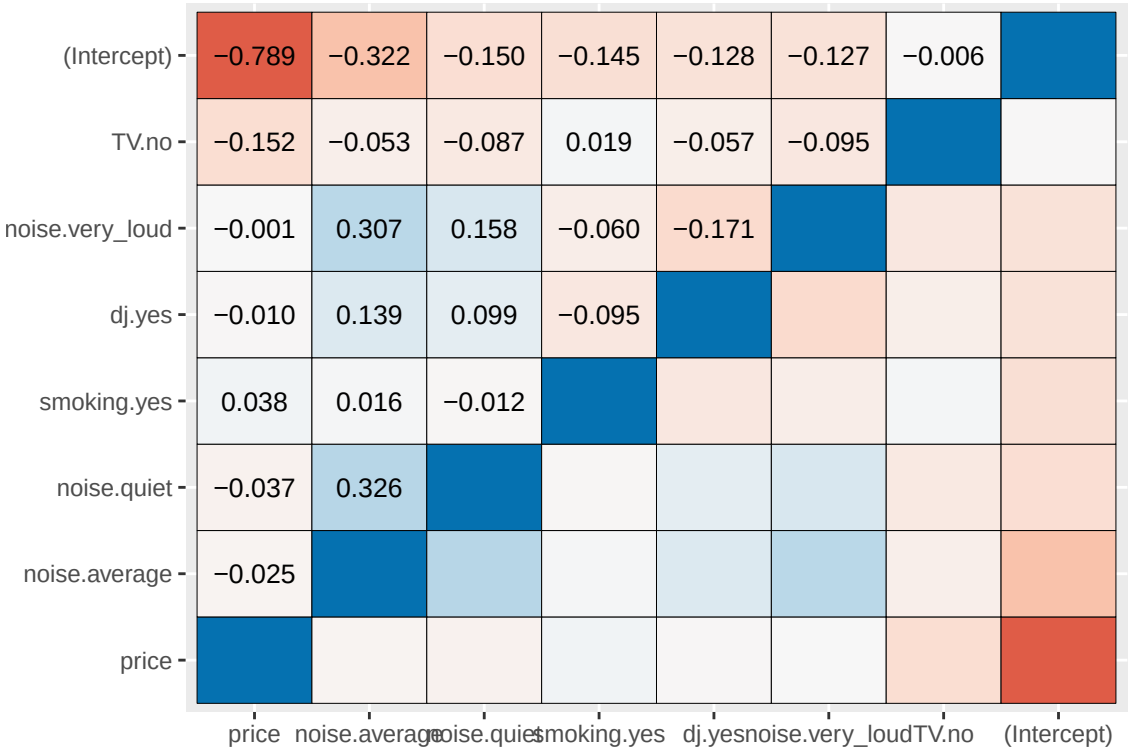
Although the mixed effect model might not be a perfect approach to fit rating data, the outcome seems well after checking the goodness of fit. It seems like there is a cultural difference among these countries in terms of people's attitudes toward bars. In general, German bars receive a higher rating on average. But a particular category might be welcomed in a country but is not as the same in others, such as cocktail bars. There are other interesting conclusion can be drawn from the model. Meanwhile, the limitation of the model has been discussed before and the further step is to look into other datasets where other appraoches and models will be adopted.

Appendix

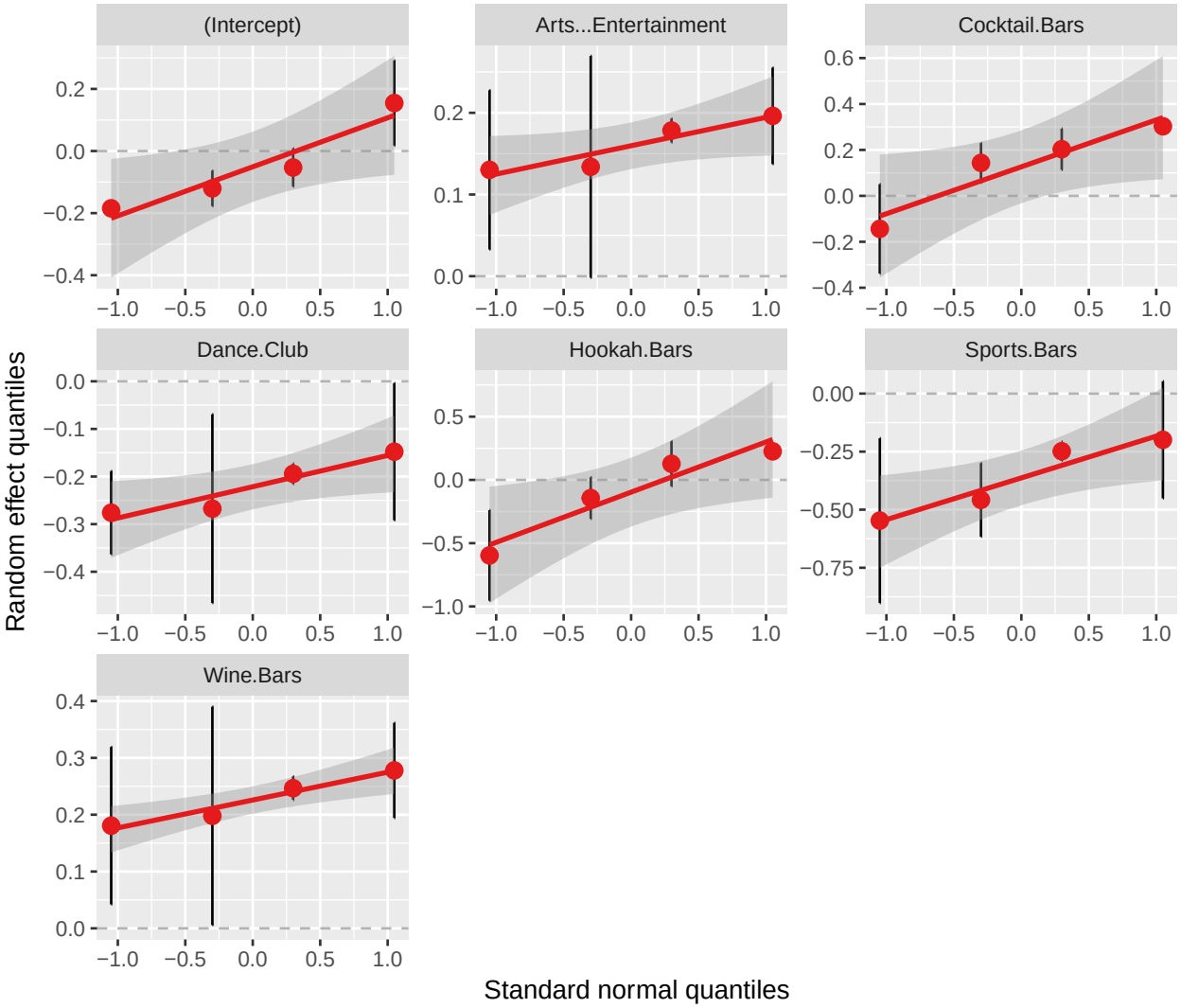
1.Exploratory Data Analysis (continued)



2.Fixed Effect Correlation Matrix



3.QQ-Plot for Random Effect



4.Multinomial Mixed Effect Model and Likelihood Ratio Test

```
# multinomial mixed effect model
fit.1 <- clmm2(factor(stars)~noise.average+noise.quiet+noise.very_loud+smoking.yes+dj.yes+TV.no+price, r
```

```
## Warning: clmm2 may not have converged:
##   optimizer 'ucminf' terminated with max|gradient|: 0.000582363082483623
```

```
fit.2 <- clm2(factor(stars)~noise.average+noise.quiet+noise.very_loud+smoking.yes+dj.yes+TV.no+price+Ar
fit.3 <- clmm2(factor(stars)~noise.average+noise.quiet+noise.very_loud+smoking.yes+dj.yes+TV.no+price+A
```

```
## Warning: clmm2 may not have converged:
##   optimizer 'ucminf' terminated with max|gradient|: 0.000203187969603235
```

```r
anova(fit.1, fit.2, fit.3)
```

```
## Likelihood ratio tests of cumulative link models
##
## Response: factor(stars)
##
## 1                                                                                                                noise.a
## 2 noise.average + noise.quiet + noise.very_loud + smoking.yes + dj.yes + TV.no + price + Arts...Entei
## 3 noise.average + noise.quiet + noise.very_loud + smoking.yes + dj.yes + TV.no + price + Arts...Entei
##   Resid. df -2logLik   Test    Df   LR stat.    Pr(Chi)
## 1      4500 14645.76
## 2      4495 14349.23 1 vs 2     5 296.527835 0.00000000
## 3      4494 14340.70 2 vs 3     1   8.529316 0.00349471
```