Midterm Project—Yelp

Ang Li

2017/12/2

1. Project Description

1.1 Overview

The love of people for delicious food has never decreased. Yelp, as an informative platform, gives us opportunity to view, learn, search and comment for restaurants or other business easily. As it becomes more and more popular as an application, the reviews and comments that customers leave on Yelp have larger and larger impact on other customers' choices. When we are looking at the list of restaurants, the first thing we will check is the "stars" that they have, scaled from 0 to 5. There are a lot of factors that can influence a restaurant's score. In this way, I come up with the question: WHAT DETERMINES A RESTAURANT'S SCORE? Beside the extent of love of people for the food it provide, a restaurant's facilities, food categories or customers' personal behaviors can also become factors that influence the score on Yelp. I managed to extract data in 10 kinds of cuisines, such as Chinese, Mexican, Janpanese, American(Traditional) and etc, as I will show later in the project.

1.2 Research Question Statement

In this project, my goal is to study and find out which factors that may have an impact on the stars for restaurant, and the extents those factors contribute, under different type of cuisines.

1.3 Data and Big Data Challenges

1.3.1 Data choosing and cleaning

I use the data on the website of Yelp, which are used for "Yelp Data Challenges". The data has been divided into several datasets based on its categories, such as business, user, review, etc. Each of them is a large dataset with up to 4.7 million rows. The information that I want is dispersed in serveral datasets, so I first used the platform of SQL on R to extract columns I want from each dataset, and then I merged those columns into one single dataset, using two identifiers: business_id and user_id. I also used dplyr package to sort and clean data so that I can a tidy result for each identifier.

My data contains basically 3 things: Basic information,such as their names and state location; Conditions, such as whether they have TV; And stars they receive from customers.

Here is an example,which shows the first 5 rows of the dataset:

```
##              business_id                 business_name state noiselevel
## 1 --6MefnULPED_I942VcFNA John's Chinese BBQ Restaurant    ON          1
## 2 --S62v0QgkqQaVUhFnNHrw                       Denny's    OH          2
## 3 --SrzpvFLwP_YFwB_Cetow          Keung Kee Restaurant    ON          2
## 4 --qJNlGWyvPJfBrqwp9c0w                   Bella Vista    BW          1
## 5 -01XupAWZEXbdNbxNg5mEg 18 Degrees Neighborhood Grill    AZ          3
##   Delivery WiFi average_stars   Avuser TV Alcohol OutdoorSeating
```

```
## 1        0    0      3.233333 3.522000  0       1              0
## 2        0    0      2.000000 3.205000  1       0              0
## 3        0    2      3.631579 3.570526  0       0              0
## 4        0    0      3.266667 3.686333  0       1              0
## 5        0    2      2.888889 3.487500  1       2              1
##   Reservations Attire              category
## 1            1      1               Chinese
## 2            0      1 American (Traditional)
## 3            1      1               Chinese
## 4            1      1               Italian
## 5            0      1 American (Traditional)
```

1.3.2 Data explanation

Here are the explanations for some crucial variables (The entire explanation can be found at Appendix):

average_stars: The average stars that customers give for the restaurant on Yelp given the data we have online.

Avuser: For a single business, the average score that its users(customers) usually give for commenting restaurants on their accounts. For each restaurant, I summarize the average stars that each of commented customer gives for commenting on restaurants, adding them together and then calculated a mean of this for each restaurants.

category: The type of food that restaurant provide.

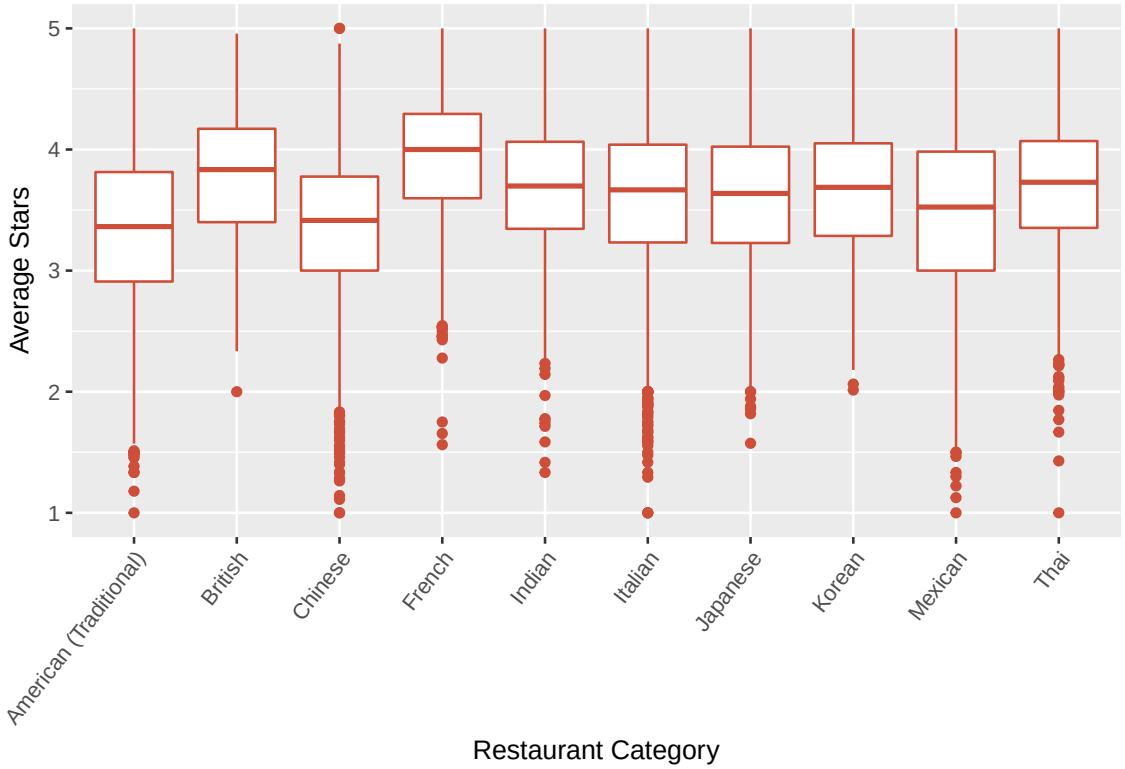
NoiseLevel: The noise level of the restaurant; 4—very loud; 3—loud; 2—average; 1—quiet.

WiFi: The wifi type for restaurant; 2—free; 1—paid; 0—no.

Alcohol: The alcohol serving type in the restaurant: 2—full_bar; 1—beer_and_wine; 0—none.

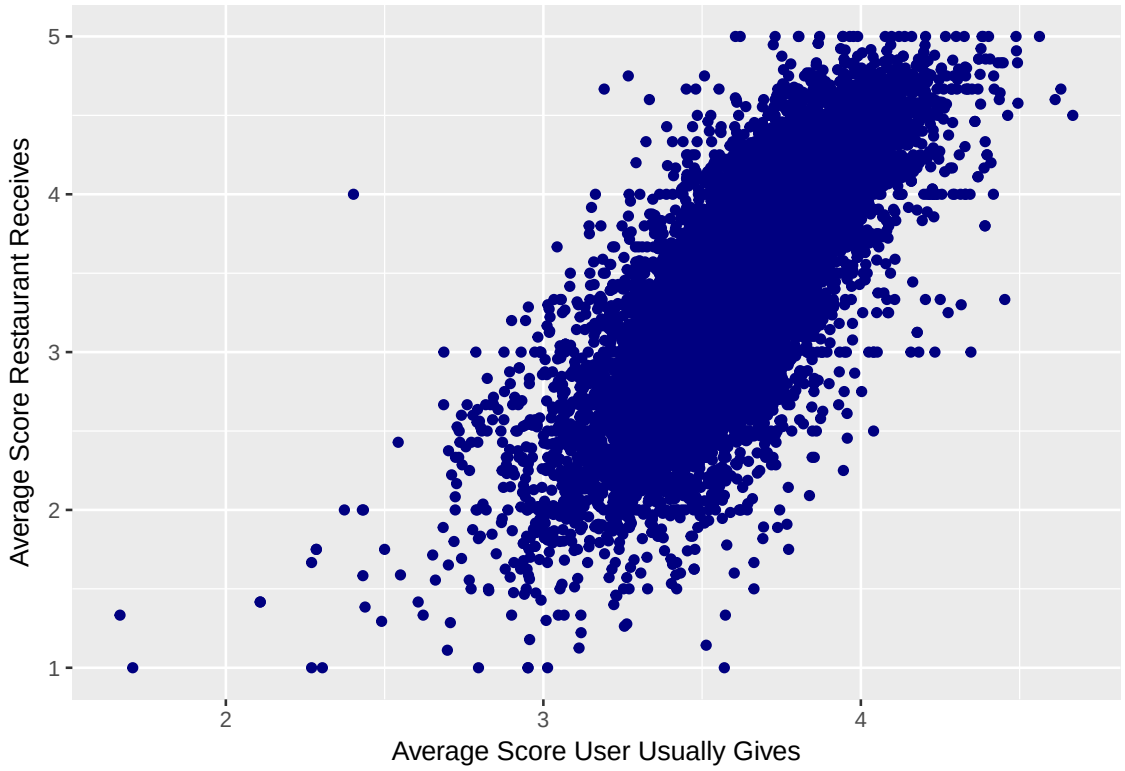
2. EDA

2.1 Boxplot of Average stars for each category



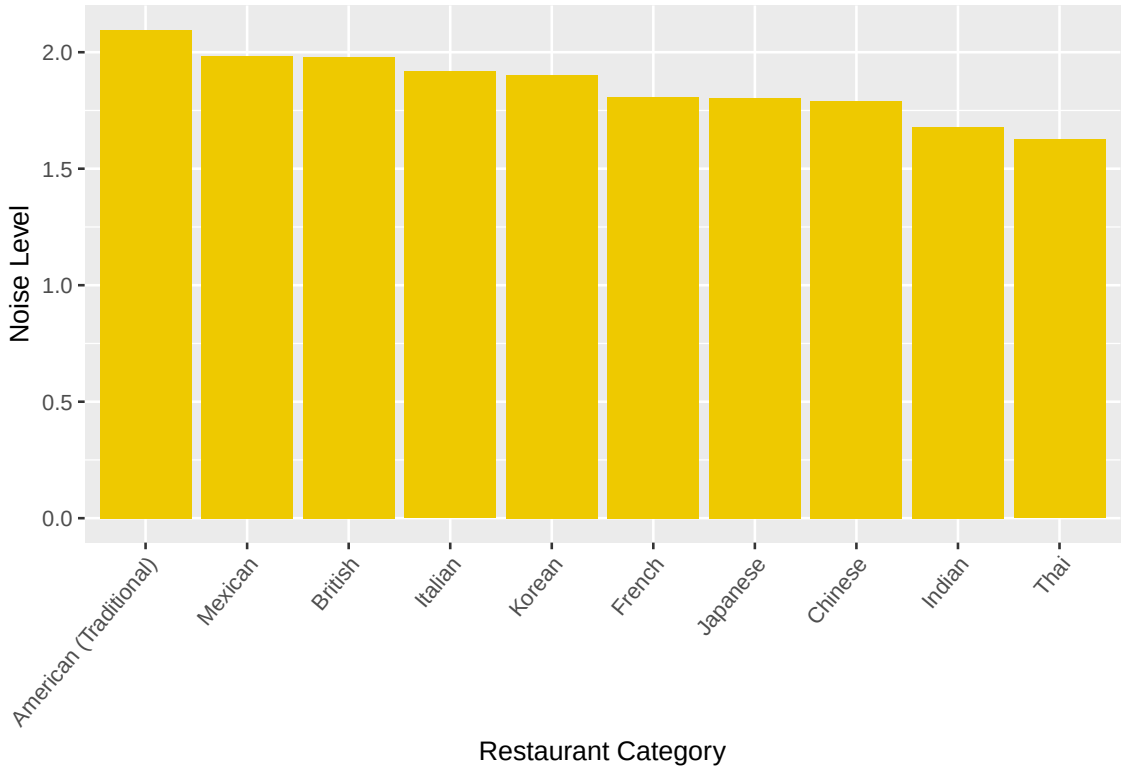
This boxplot shows the comprison of average stars that restaurant received for each cuisine category. From this plot we can find out that the median of average stars that each type of cuisine received are basically close to each other. Among these scores, French cuisine has a highest median score, While American Tradisional cuisine and Chinese cuisine receive relatively lower median scores. And according to the shape of rectangles between upper and lower quartiles, the rectangle for Mexican restaurants seem to be a little longer compare to other type of restaurants. This suggests that, the scores that Mexican food restaurants vary a little bit more compare to other type of restaurants. Also, there exist some outliers for each category, which are numerically distant from the rest of data. This is also reasonable since there are some "bad" restaurants that do not serve customers well.

2.2 Point plot of mean given stars and mean received stars



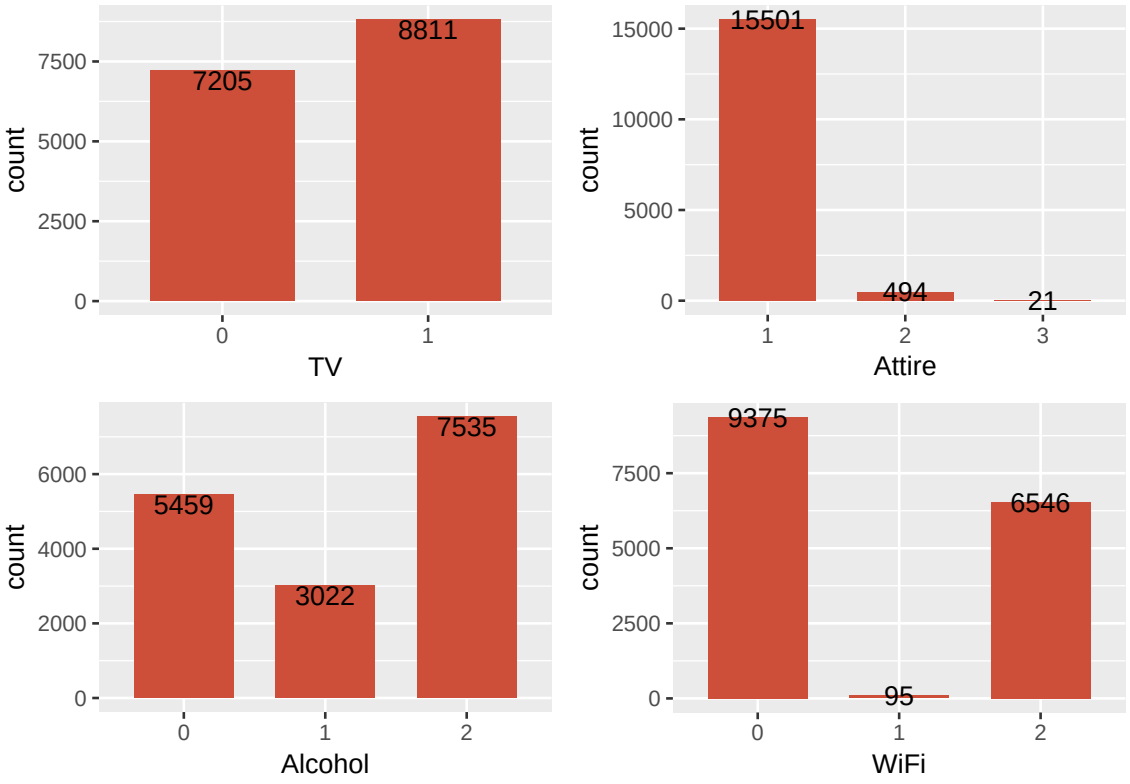
This plot shows the relationship between the average score that calculated from users' grading history and the average scores that restaurants receive. It is clear that there is a positive correlation. This also shows the subjectivity of grading restaurants from each users, which means that each Yelp app user has his or her own standard and preference, for example, 4 might be a low score for someone while 2 is a relatively high score for a "cynical" person.

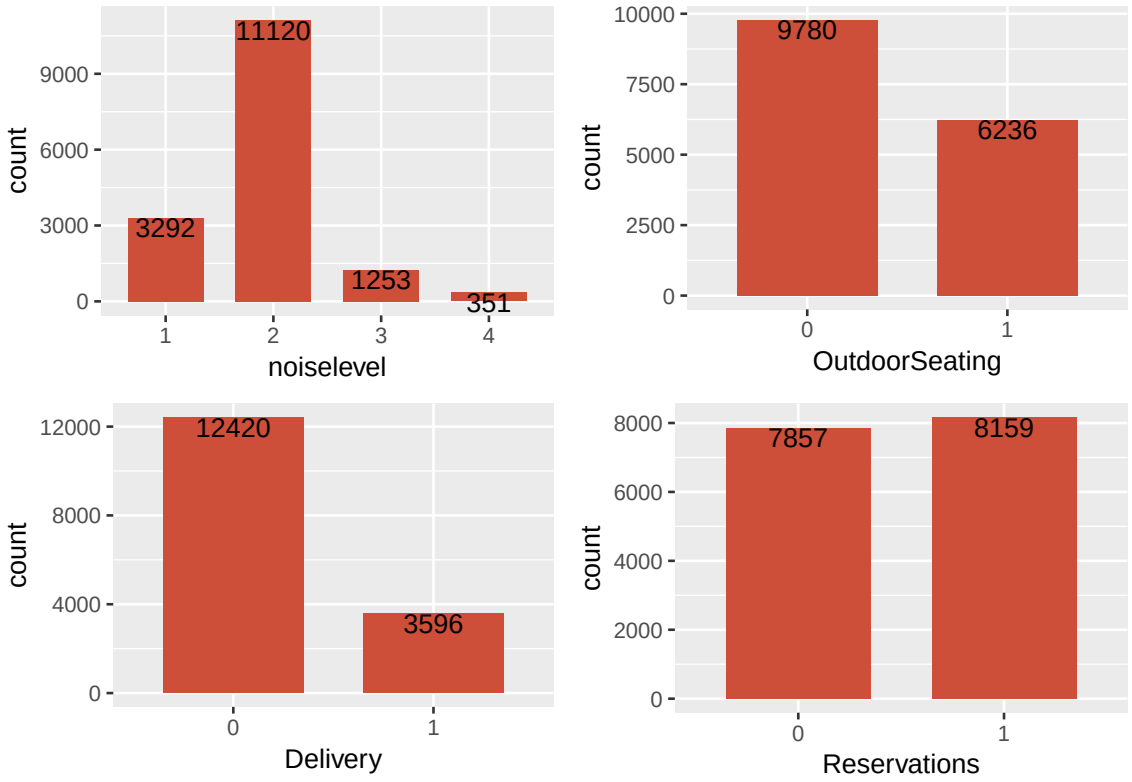
2.3 Barplot of mean noise level for each category



This plot shows the mean noise level for each type of restaurants. From the plot we can see that noise levels are ordered from high to low, and none of the types of restaurants has high level of noise. American Traditional has the relatively highest mean noise level(around 2.0), while Thai restaurants have the lowest. This might be a factor that can influence on the stars a restaurant receive, the extent may change according to category.

2.4 The barplot of count of each variable





These plots show the proportion of factors in each variable. From those plots we can notice that some factors for some variables have little observations, for example, for Attire, "dressy"(2) and "formal"(3) and for WiFi, "paid"(1) have little observations in the dataset. There are only 21 observations for "formal" in Attire and 95 only for "paid" in WiFi. These are the things that we should know before fitting models since with little data, the coefficients for these factors in our model might be less credible.

3. Model Selection

For the Model Selection part, first, I randomly selected 15000 observations out of 16016 to build the models, (other 1016s are planing for predictions.) I tested 3 models in total and try to find the best one. I will describe my process and list the output of my final model in the report. (Outputs for other former models are showed in Appendix).

3.1 First model —fit1, I treat category as a random effect and fit the model, which will fit the data into 10 sub-regression models (10 categories) with different intercepts. And this is how my model (fit1) looks like:

$$average - stars = Avuser + Delivery + WiFi + TV + Alcohol + OutdoorSeating + Reservations + Attire + noiselevel + (1|category)$$

From the information above, we can see that from ANOVA calculation table, for varaible TV, the P-value for its estimate coefficient is 0.7759105, greater than 0.05; and for variable Delivery, the P-value for its estimate coefficient is 0.1509271, also greater than 0.05. It means that these variables are not significant, after checking the summary table for model fit1, I decided that I should remove them from the model. For other varaibles in the model, although the coefficient are relatively small but siginificant.

For the coefficients table, we can see that those coefficients stay the same while intercepts vary for different types of restaurants, but those intercepts only vary a little and they are close to each other.

For the residual plot, although there are some points with relatively large distances from the y=0 line, most of the points are clustered around the zero line, which means the model we used is generally reasonable.

3.2 Second model —fit2, I still treat category as a random effect and fit the model without variables TV and Delivery, which will fit the data into 10 regression models (10 categories) with different intercepts. And this is how the model (fit2) looks like:

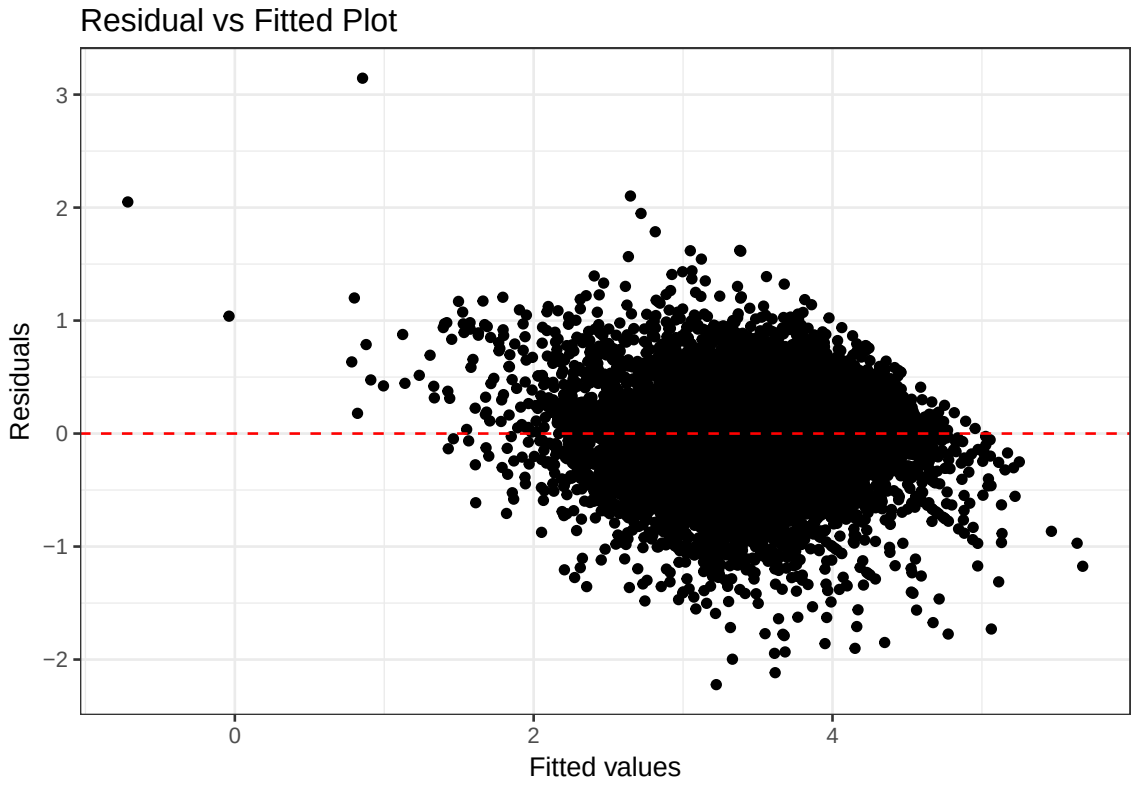
$$average - stars = Avuser + WiFi + Alcohol + OutdoorSeating + Reservations + Attire + noiselevel + (1|category)$$

The second model is basically the same with the first model besides some small changes on coefficients, and residual plot is also similar with the first one. There is also no big changes on the coefficient table, and after checking the ANOVA calculation table, I decide to keep all of my current variables. Also, I notice that the variable Avuser, which is that the mean score that a user give in his grading history, plays a crucial part in this model, since it has the largest coefficient in the model. I am wondering whether the influence of this variable varys from differnt types of restaurants. In this way, my next step will be allowing it to vary with group, which leads to my third model.

3.3 Third model —fit3, this is my final model. I treat the old category and add Avuser in as random effect that allow the slope to vary, and fit the model, which will fit the data into 10 regression models (10 categories) with different intercepts AND DIFFERENT COEFFICIENTS for the variable of average score of user. And these are the summary, coeficients and residual plot of fit3:

$$average - stars = Avuser + WiFi + Alcohol + OutdoorSeating + Reservations + Attire + noiselevel + (1 + Avuser|category)$$

By comparing the summaries of fit3 and fit2, I notice that for the random effect part, the variance of the category group increases from 0.004522 to 0.33686. This means that there are more different patterns that has showed up between each group, which also means I did not make those groups for nothing. This is also showed in the coefficient table, between the regression function of each group, intercepts vary more and coefficient of Avuser also varies, compare to the results from fit2. This model clearly shows the extent of impact of average score that user give on the average score one type of restaurant receives, for example, with 1 unit increase in Average user score, Mexican and American Traditional restaurants are likely to receive more credit and increase their stars more (about 0.3), compare to other types of restaurants.



I also check the residual plot for model fit3. Those points are still clustered around that area next to the horizontal line. For this situation, my guess is that this is a relatively large dataset, and the score numbers in dataset, are very close to each other, even with many same scores, in this way, they all stick together on the plot and eventually becomes a shape like this. And the model itself is fine since those fitted values are close to acutal ones and there is no obvious pattern for the points.

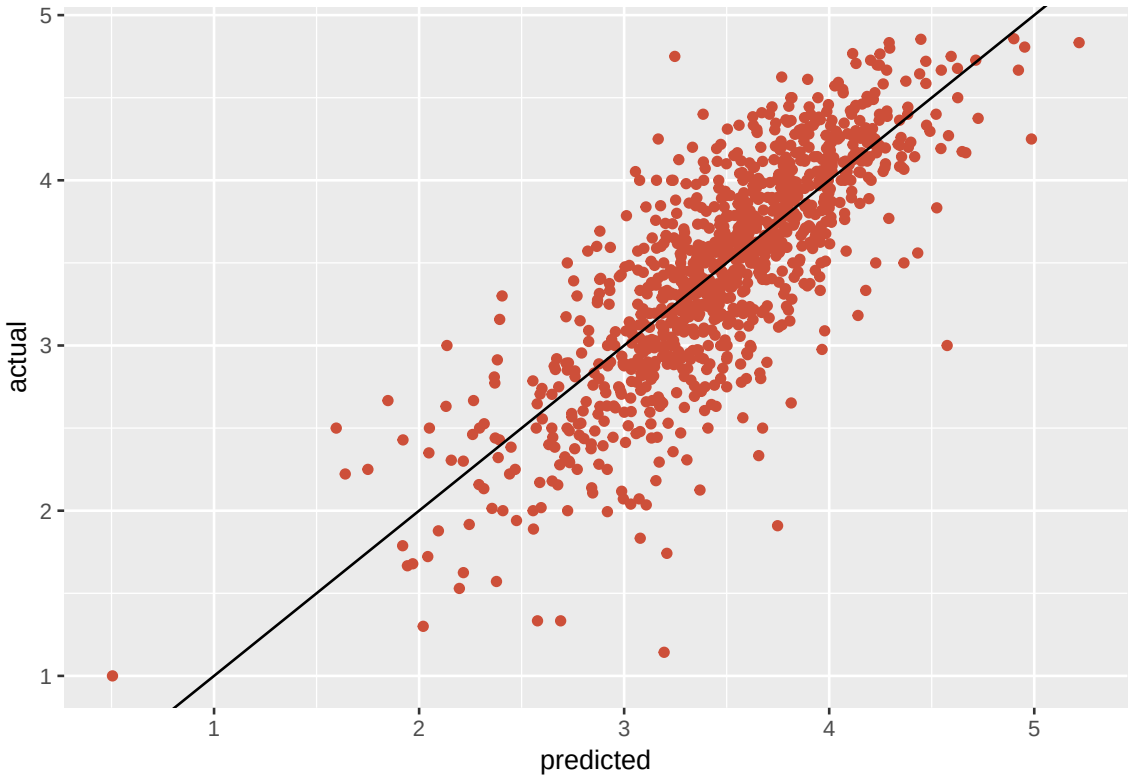
```
## refitting model(s) with ML (instead of REML)
```

```
## Data: yelp1
## Models:
## object: average_stars ~ Avuser + WiFi + Alcohol + OutdoorSeating + Reservations +
## object:     Attire + noiselevel + (1 | category)
## ..1: average_stars ~ Avuser + WiFi + Alcohol + OutdoorSeating + Reservations +
## ..1:     Attire + noiselevel + (1 + Avuser | category)
##        Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## object 15 14316 14431 -7143.2    14286
```

```
## ..1    17 14217 14346 -7091.4    14183 103.61      2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is the anova table for model fit2 and fit3. For the comparison of AIC, I notice that AIC decreases from 14316 to 14217. We prefer lower AICs so it means fit3 does get better.

```
## Warning: executing %dopar% sequentially: no parallel backend registered
```



Finally, I manage to use my model to do prediction for new data. For those 1016 observations that I did not use before, this plot shows the relationship between predicted and actual data. We can see that nearly every point is around the abline and most of them are very close or even right on the line, which means our model is effective and does reflect the features for this new data set.

4. Conclusion

4.1 Results

Based on the analysis above, we have proved the model fit3 a good model, and we also can see that nearly all the variables that stay in model are significant. So conditions like noise, wifi, reservations, they all influence a restaurant's star for all types of restaurants, but in a relatively subtle way, not very obvious. On the contrary, the variable Avuser, which is the pattern of individual customers, has a huge impact on a restaurant's star. So apart from its food, how to find out consumers'chracteristics is also important for restaurants, since scores on Yelp could influence their business and profits.

4.2 Discussions

As I mentioned above in the project, it seems that customer's subjectivity and personal preference and habbits play an important part on restaurant's grading scores. In the future study, my priority will be finding that "what determines a person's subjectivity on grading?" and I also think of the case that, for example, a person in a loud noise restaurant might become whiny so he will give a lower score on Yelp in a foul mood, "Is there any relationship between outside conditions and a person's subjectivity?" In my opinion, these questions are all worth to study and talk about.

5. Appendix

Here are some information that you might be interested in:

5.1 Entire Data explanation:

business_id: This is the identification for each business(restaurant), which is to make every business unique;

business_name: The name of every restaurant;

State: The state location of every restaurant;

NoiseLevel: The noise level of the restaurant; 4—very loud; 3—loud; 2—average; 1—quiet.

Delivery: Whether a restaurant has delivery service; 0—no; 1—yes.

WiFi: The wifi type for restaurant; 2—free; 1—paid; 0—no.

Alcohol: The alcohol serving type in the restaurant: 2—full_bar; 1—beer_and_wine; 0—none.

average_stars: The average stars that customers give for the restaurant on Yelp given the data we have online.

Avuser: For a single business, the average score that its users(customers) usually give for commenting restaurants on their accounts.

TV: Whether a business has a TV; 0—no; 1—yes.

```
## 5.2.1 Model: Fit1
```

```
### summary for model: fit1
```

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula:
## average_stars ~ Avuser + Delivery + WiFi + TV + Alcohol + OutdoorSeating +
##     Reservations + Attire + noiselevel + (1 | category)
##    Data: yelp1
##
## REML criterion at convergence: 14386.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.6892 -0.5856  0.0684  0.6432  7.8054
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  category (Intercept) 0.004504 0.06711
##  Residual             0.151539 0.38928
## Number of obs: 15000, groups:  category, 10
##
## Fixed effects:
##                    Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)     -3.612e+00  5.569e-02  3.760e+02 -64.868  < 2e-16 ***
## Avuser           1.972e+00  1.373e-02  1.498e+04 143.591  < 2e-16 ***
## Delivery1        1.191e-02  8.295e-03  1.486e+04   1.436 0.150927
## WiFi1           -1.491e-01  4.156e-02  1.498e+04  -3.587 0.000335 ***
## WiFi2           -1.833e-02  7.047e-03  1.498e+04  -2.601 0.009307 **
## TV1              2.025e-03  7.113e-03  1.497e+04   0.285 0.775910
## Alcohol1        -7.253e-03  9.735e-03  1.498e+04  -0.745 0.456265
## Alcohol2        -5.796e-02  9.257e-03  1.498e+04  -6.261 3.92e-10 ***
## OutdoorSeating1 -4.317e-02  7.198e-03  1.494e+04  -5.998 2.04e-09 ***
## Reservations1    3.279e-02  7.841e-03  1.488e+04   4.182 2.91e-05 ***
## Attire2          1.590e-01  1.962e-02  1.491e+04   8.102 4.44e-16 ***
## Attire3         -1.386e-01  9.490e-02  1.498e+04  -1.460 0.144252
## noiselevel2     -8.648e-03  8.419e-03  1.498e+04  -1.027 0.304343
## noiselevel3     -7.261e-02  1.399e-02  1.498e+04  -5.191 2.12e-07 ***
## noiselevel4     -1.877e-01  2.292e-02  1.498e+04  -8.192 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation matrix not shown by default, as p = 15 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```

```
###coefficients for model: fit1
```

```
## $category
##                         (Intercept)   Avuser  Delivery1      WiFi1
## American (Traditional)   -3.715872 1.971538 0.01191438 -0.1490912
## British                  -3.589387 1.971538 0.01191438 -0.1490912
## Chinese                  -3.665977 1.971538 0.01191438 -0.1490912
```

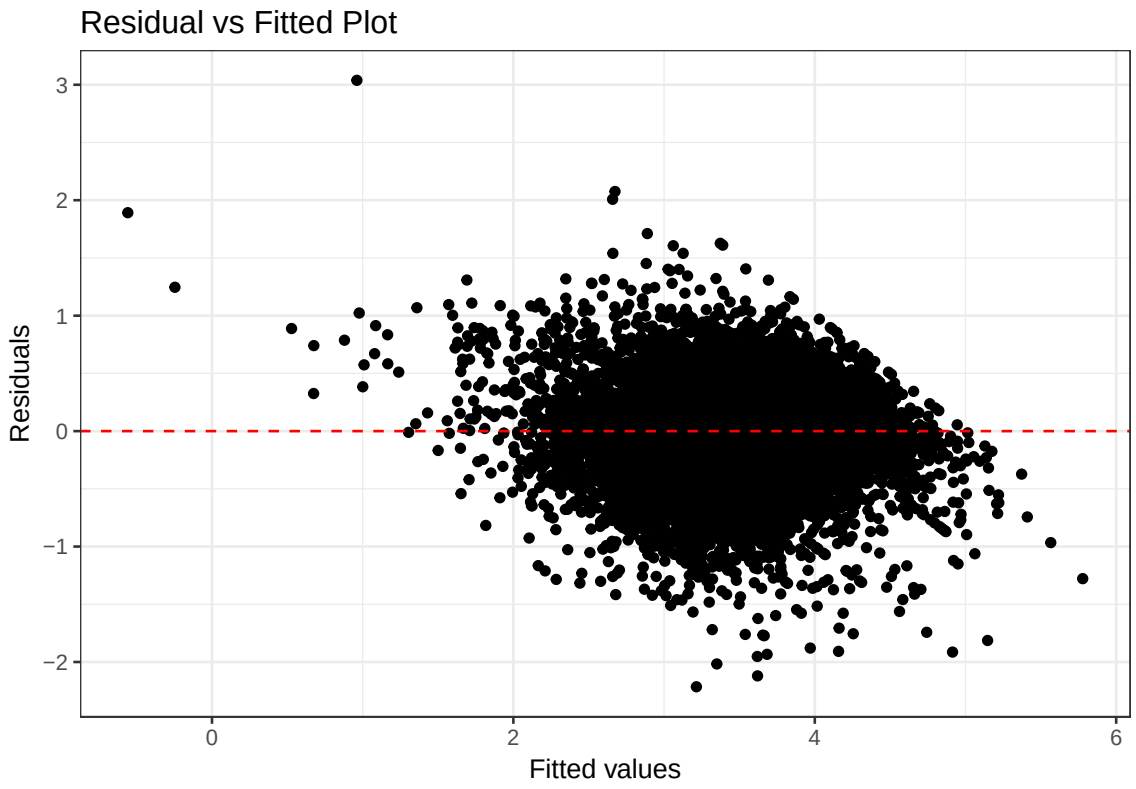
```
## French                   -3.521173 1.971538 0.01191438 -0.1490912
## Indian                   -3.518181 1.971538 0.01191438 -0.1490912
## Italian                  -3.608772 1.971538 0.01191438 -0.1490912
## Japanese                 -3.624559 1.971538 0.01191438 -0.1490912
## Korean                   -3.572435 1.971538 0.01191438 -0.1490912
## Mexican                  -3.686171 1.971538 0.01191438 -0.1490912
## Thai                     -3.620332 1.971538 0.01191438 -0.1490912
##                             WiFi2         TV1     Alcohol1    Alcohol2
## American (Traditional) -0.01832781 0.002024637 -0.007253037 -0.05796126
## British                -0.01832781 0.002024637 -0.007253037 -0.05796126
## Chinese                -0.01832781 0.002024637 -0.007253037 -0.05796126
## French                 -0.01832781 0.002024637 -0.007253037 -0.05796126
## Indian                 -0.01832781 0.002024637 -0.007253037 -0.05796126
## Italian                -0.01832781 0.002024637 -0.007253037 -0.05796126
## Japanese               -0.01832781 0.002024637 -0.007253037 -0.05796126
## Korean                 -0.01832781 0.002024637 -0.007253037 -0.05796126
## Mexican                -0.01832781 0.002024637 -0.007253037 -0.05796126
## Thai                   -0.01832781 0.002024637 -0.007253037 -0.05796126
##                        OutdoorSeating1 Reservations1   Attire2    Attire3
## American (Traditional)     -0.04317415     0.0327889 0.1589752 -0.1385721
## British                    -0.04317415     0.0327889 0.1589752 -0.1385721
## Chinese                    -0.04317415     0.0327889 0.1589752 -0.1385721
## French                     -0.04317415     0.0327889 0.1589752 -0.1385721
## Indian                     -0.04317415     0.0327889 0.1589752 -0.1385721
## Italian                    -0.04317415     0.0327889 0.1589752 -0.1385721
## Japanese                   -0.04317415     0.0327889 0.1589752 -0.1385721
## Korean                     -0.04317415     0.0327889 0.1589752 -0.1385721
## Mexican                    -0.04317415     0.0327889 0.1589752 -0.1385721
## Thai                       -0.04317415     0.0327889 0.1589752 -0.1385721
##                         noiselevel2 noiselevel3 noiselevel4
## American (Traditional) -0.008647579 -0.07261125   -0.187725
## British                -0.008647579 -0.07261125   -0.187725
## Chinese                -0.008647579 -0.07261125   -0.187725
## French                 -0.008647579 -0.07261125   -0.187725
## Indian                 -0.008647579 -0.07261125   -0.187725
## Italian                -0.008647579 -0.07261125   -0.187725
## Japanese               -0.008647579 -0.07261125   -0.187725
## Korean                 -0.008647579 -0.07261125   -0.187725
## Mexican                -0.008647579 -0.07261125   -0.187725
## Thai                   -0.008647579 -0.07261125   -0.187725
##
## attr(,"class")
## [1] "coef.mer"
```

```
###ANOVA calculation  for model: fit1
```

```
## Analysis of Variance Table of type III  with  Satterthwaite
## approximation for degrees of freedom
##                 Sum Sq Mean Sq NumDF DenDF F.value    Pr(>F)
## Avuser         3124.47 3124.47     1 14982 20618.2 < 2.2e-16 ***
## Delivery          0.31    0.31     1 14863     2.1 0.1509271
## WiFi              2.78    1.39     2 14977     9.2 0.0001037 ***
## TV                0.01    0.01     1 14968     0.1 0.7759105
## Alcohol           7.02    3.51     2 14948    23.2 9.072e-11 ***
## OutdoorSeating    5.45    5.45     1 14943    36.0 2.042e-09 ***
```

```
## Reservations      2.65    2.65     1 14885    17.5 2.909e-05 ***
## Attire           10.37    5.19     2 14944    34.2 1.554e-15 ***
## noiselevel       14.18    4.73     3 14982    31.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##Residual plot for model: fit1
```




```
## 5.2.2 Model: Fit2
```

```
### summary for model: fit2
```

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula:
## average_stars ~ Avuser + WiFi + Alcohol + OutdoorSeating + Reservations +
##     Attire + noiselevel + (1 | category)
##    Data: yelp1
##
## REML criterion at convergence: 14372.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.7010 -0.5877  0.0675  0.6429  7.8265
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  category (Intercept) 0.004522 0.06725
##  Residual             0.151541 0.38928
## Number of obs: 15000, groups:  category, 10
##
```

```
## Fixed effects:
##                    Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)     -3.605e+00  5.547e-02  3.680e+02 -64.987  < 2e-16 ***
## Avuser           1.971e+00  1.371e-02  1.498e+04 143.774  < 2e-16 ***
## WiFi1           -1.492e-01  4.156e-02  1.498e+04  -3.591  0.00033 ***
## WiFi2           -1.719e-02  6.934e-03  1.498e+04  -2.479  0.01317 *
## Alcohol1        -6.891e-03  9.678e-03  1.498e+04  -0.712  0.47644
## Alcohol2        -5.795e-02  8.925e-03  1.497e+04  -6.493 8.69e-11 ***
## OutdoorSeating1 -4.345e-02  7.192e-03  1.494e+04  -6.041 1.56e-09 ***
## Reservations1    3.287e-02  7.835e-03  1.488e+04   4.195 2.75e-05 ***
## Attire2          1.563e-01  1.944e-02  1.487e+04   8.042 8.88e-16 ***
## Attire3         -1.401e-01  9.489e-02  1.499e+04  -1.476  0.13999
## noiselevel2     -1.010e-02  8.350e-03  1.498e+04  -1.210  0.22645
## noiselevel3     -7.482e-02  1.390e-02  1.499e+04  -5.382 7.49e-08 ***
## noiselevel4     -1.898e-01  2.286e-02  1.498e+04  -8.303  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation matrix not shown by default, as p = 13 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```

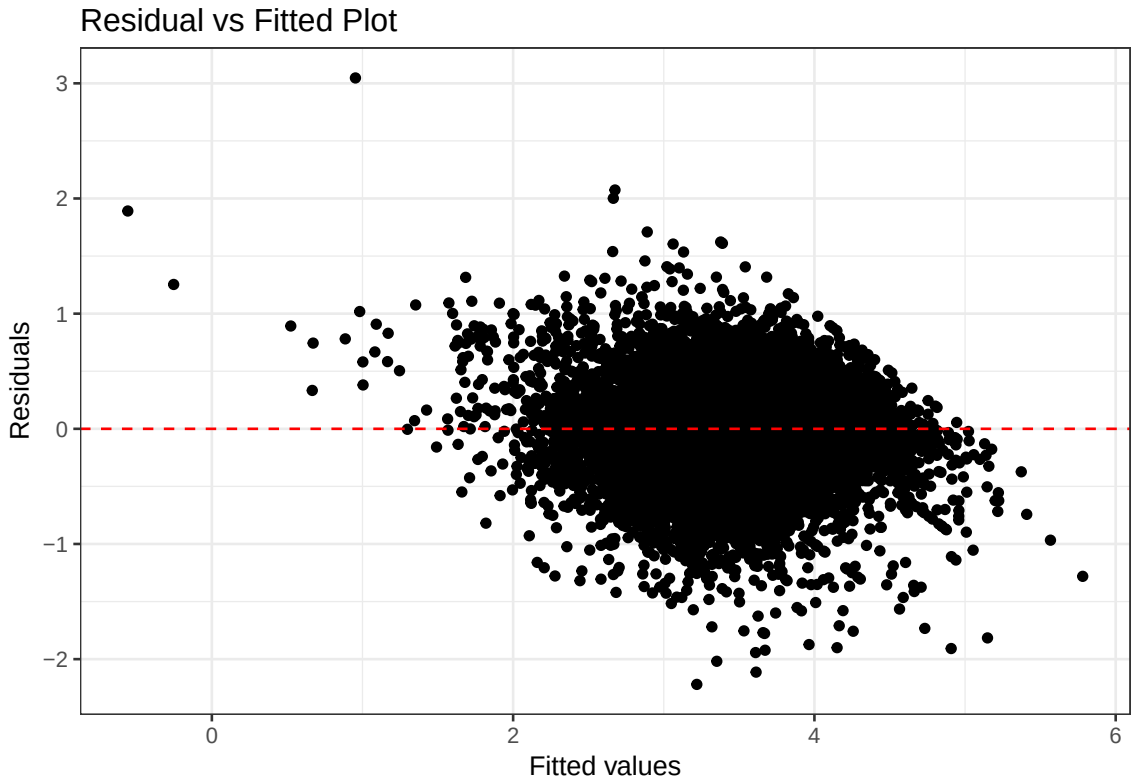
```
###coefficients for model: fit2

## $category
##                         (Intercept)   Avuser      WiFi1       WiFi2
## American (Traditional)   -3.709843 1.970807 -0.1492408 -0.01719227
## British                  -3.583673 1.970807 -0.1492408 -0.01719227
## Chinese                  -3.657141 1.970807 -0.1492408 -0.01719227
## French                   -3.515894 1.970807 -0.1492408 -0.01719227
## Indian                   -3.508721 1.970807 -0.1492408 -0.01719227
## Italian                  -3.599985 1.970807 -0.1492408 -0.01719227
## Japanese                 -3.617362 1.970807 -0.1492408 -0.01719227
## Korean                   -3.566105 1.970807 -0.1492408 -0.01719227
## Mexican                  -3.680227 1.970807 -0.1492408 -0.01719227
## Thai                     -3.610708 1.970807 -0.1492408 -0.01719227
##                           Alcohol1    Alcohol2 OutdoorSeating1
## American (Traditional) -0.006891459 -0.05795021     -0.04345183
## British                -0.006891459 -0.05795021     -0.04345183
## Chinese                -0.006891459 -0.05795021     -0.04345183
## French                 -0.006891459 -0.05795021     -0.04345183
## Indian                 -0.006891459 -0.05795021     -0.04345183
## Italian                -0.006891459 -0.05795021     -0.04345183
## Japanese               -0.006891459 -0.05795021     -0.04345183
## Korean                 -0.006891459 -0.05795021     -0.04345183
## Mexican                -0.006891459 -0.05795021     -0.04345183
## Thai                   -0.006891459 -0.05795021     -0.04345183
##                        Reservations1   Attire2    Attire3 noiselevel2
## American (Traditional)    0.03286566 0.1563225 -0.1400547 -0.01010041
## British                   0.03286566 0.1563225 -0.1400547 -0.01010041
## Chinese                   0.03286566 0.1563225 -0.1400547 -0.01010041
## French                    0.03286566 0.1563225 -0.1400547 -0.01010041
## Indian                    0.03286566 0.1563225 -0.1400547 -0.01010041
## Italian                   0.03286566 0.1563225 -0.1400547 -0.01010041
```

```
## Japanese                   0.03286566 0.1563225 -0.1400547 -0.01010041
## Korean                     0.03286566 0.1563225 -0.1400547 -0.01010041
## Mexican                    0.03286566 0.1563225 -0.1400547 -0.01010041
## Thai                       0.03286566 0.1563225 -0.1400547 -0.01010041
##                          noiselevel3 noiselevel4
## American (Traditional) -0.07482242  -0.1898033
## British                -0.07482242  -0.1898033
## Chinese                -0.07482242  -0.1898033
## French                 -0.07482242  -0.1898033
## Indian                 -0.07482242  -0.1898033
## Italian                -0.07482242  -0.1898033
## Japanese               -0.07482242  -0.1898033
## Korean                 -0.07482242  -0.1898033
## Mexican                -0.07482242  -0.1898033
## Thai                   -0.07482242  -0.1898033
##
## attr(,"class")
## [1] "coef.mer"
```

```
###ANOVA calculation  for model: fit2
```

```
## Analysis of Variance Table of type III  with  Satterthwaite
## approximation for degrees of freedom
##                  Sum Sq Mean Sq NumDF DenDF F.value    Pr(>F)
## Avuser         3132.50 3132.50     1 14984 20671.0 < 2.2e-16 ***
## WiFi              2.70    1.35     2 14981     8.9 0.0001366 ***
## Alcohol           7.54    3.77     2 14942    24.9 1.613e-11 ***
## OutdoorSeating    5.53    5.53     1 14940    36.5 1.564e-09 ***
## Reservations      2.67    2.67     1 14876    17.6 2.749e-05 ***
## Attire           10.23    5.11     2 14920    33.7 2.442e-15 ***
## noiselevel       14.53    4.84     3 14985    32.0 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##Residual plot for model: fit2
```




```
## 5.2.3 Model: Fit3
```

```
##summary for model: fit3
summary(fit3,correlation=T,maxsum=50)
```

```
## Linear mixed model fit by REML t-tests use Satterthwaite approximations
##   to degrees of freedom [lmerMod]
## Formula:
## average_stars ~ Avuser + WiFi + Alcohol + OutdoorSeating + Reservations +
##     Attire + noiselevel + (1 + Avuser | category)
##    Data: yelp1
##
## REML criterion at convergence: 14266.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.7318 -0.5811  0.0686  0.6401  8.1124
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  category (Intercept) 0.33686  0.5804
##           Avuser      0.02137  0.1462   -1.00
##  Residual             0.15031  0.3877
## Number of obs: 15000, groups:  category, 10
##
## Fixed effects:
##                   Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)     -3.439e+00  1.991e-01  1.000e+01 -17.276 1.20e-08 ***
## Avuser           1.926e+00  5.056e-02  1.000e+01  38.099 5.51e-12 ***
## WiFi1           -1.505e-01  4.140e-02  1.497e+04  -3.636 0.000278 ***
```

```
## WiFi2           -1.653e-02  6.914e-03  1.498e+04  -2.391 0.016812 *
## Alcohol1        -8.036e-03  9.648e-03  1.498e+04  -0.833 0.404899
## Alcohol2        -5.884e-02  8.897e-03  1.495e+04  -6.613 3.89e-11 ***
## OutdoorSeating1 -4.287e-02  7.165e-03  1.486e+04  -5.983 2.23e-09 ***
## Reservations1    3.117e-02  7.811e-03  1.481e+04   3.991 6.62e-05 ***
## Attire2          1.575e-01  1.937e-02  1.476e+04   8.134 4.44e-16 ***
## Attire3         -1.213e-01  9.462e-02  1.494e+04  -1.282 0.199835
## noiselevel2     -1.238e-02  8.322e-03  1.492e+04  -1.488 0.136751
## noiselevel3     -7.502e-02  1.385e-02  1.498e+04  -5.416 6.20e-08 ***
## noiselevel4     -1.927e-01  2.277e-02  1.498e+04  -8.462  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 13 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```

```
##coefficients for model: fit3
coef(fit3)
```

```
## $category
##                         (Intercept)   Avuser     WiFi1       WiFi2
## American (Traditional)   -4.003082 2.052355 -0.150528 -0.01653294
## British                  -3.510494 1.951229 -0.150528 -0.01653294
## Chinese                  -3.027895 1.796351 -0.150528 -0.01653294
## French                   -3.092161 1.858762 -0.150528 -0.01653294
## Indian                   -2.743704 1.762259 -0.150528 -0.01653294
## Italian                  -3.195883 1.861783 -0.150528 -0.01653294
## Japanese                 -3.595911 1.965763 -0.150528 -0.01653294
## Korean                   -3.050581 1.831834 -0.150528 -0.01653294
## Mexican                  -4.552461 2.209479 -0.150528 -0.01653294
## Thai                     -3.621215 1.974191 -0.150528 -0.01653294
##                            Alcohol1    Alcohol2 OutdoorSeating1
## American (Traditional) -0.008035672 -0.05884102     -0.04287399
## British                -0.008035672 -0.05884102     -0.04287399
## Chinese                -0.008035672 -0.05884102     -0.04287399
## French                 -0.008035672 -0.05884102     -0.04287399
## Indian                 -0.008035672 -0.05884102     -0.04287399
## Italian                -0.008035672 -0.05884102     -0.04287399
## Japanese               -0.008035672 -0.05884102     -0.04287399
## Korean                 -0.008035672 -0.05884102     -0.04287399
## Mexican                -0.008035672 -0.05884102     -0.04287399
## Thai                   -0.008035672 -0.05884102     -0.04287399
##                        Reservations1  Attire2    Attire3 noiselevel2
## American (Traditional)    0.03117226 0.157541 -0.1213045 -0.01238321
## British                   0.03117226 0.157541 -0.1213045 -0.01238321
## Chinese                   0.03117226 0.157541 -0.1213045 -0.01238321
## French                    0.03117226 0.157541 -0.1213045 -0.01238321
## Indian                    0.03117226 0.157541 -0.1213045 -0.01238321
## Italian                   0.03117226 0.157541 -0.1213045 -0.01238321
## Japanese                  0.03117226 0.157541 -0.1213045 -0.01238321
## Korean                    0.03117226 0.157541 -0.1213045 -0.01238321
## Mexican                   0.03117226 0.157541 -0.1213045 -0.01238321
## Thai                      0.03117226 0.157541 -0.1213045 -0.01238321
```

```
##                         noiselevel3 noiselevel4
## American (Traditional) -0.07502204  -0.1927058
## British                -0.07502204  -0.1927058
## Chinese                -0.07502204  -0.1927058
## French                 -0.07502204  -0.1927058
## Indian                 -0.07502204  -0.1927058
## Italian                -0.07502204  -0.1927058
## Japanese               -0.07502204  -0.1927058
## Korean                 -0.07502204  -0.1927058
## Mexican                -0.07502204  -0.1927058
## Thai                   -0.07502204  -0.1927058
##
## attr(,"class")
## [1] "coef.mer"
```