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. 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.

In this project, my goal is to study the factors that may have an impact on the stars for restaurant,under different type of cuisines.

1.2 Data and Big Data Challenges

1.2.1 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 form 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 of dataset:

```
##              business_id                 business_name state stars
## 1 --6MefnULPED_I942VcFNA John's Chinese BBQ Restaurant    ON   3.0
## 2 --S62v0QgkqQaVUhFnNHrw                       Denny's    OH   2.0
## 3 --SrzpvFLwP_YFwB_Cetow          Keung Kee Restaurant    ON   3.5
## 4 --qJNlGWyvPJfBrqwp9c0w                   Bella Vista    BW   3.5
## 5 -01XupAWZEXbdNbxNg5mEg 18 Degrees Neighborhood Grill    AZ   3.0
##   noiselevel Delivery WiFi average_stars   Avuser TV Alcohol
## 1          1        0    0      3.233333 3.522000  0       1
## 2          2        0    0      2.000000 3.205000  1       0
## 3          2        0    2      3.631579 3.570526  0       0
```

```
## 4           1        0    0       3.266667 3.686333  0       1
## 5           3        0    2       2.888889 3.487500  1       2
##   OutdoorSeating Reservations Attire              category
## 1              0            1      1               Chinese
## 2              0            0      1 American (Traditional)
## 3              0            1      1               Chinese
## 4              0            1      1               Italian
## 5              1            0      1 American (Traditional)
```

1.2.2 Data explanation

Here is some explanation for crucial variables:

NoiseLevel: 4—very loud; 3—loud; 2—average; 1—quiet.

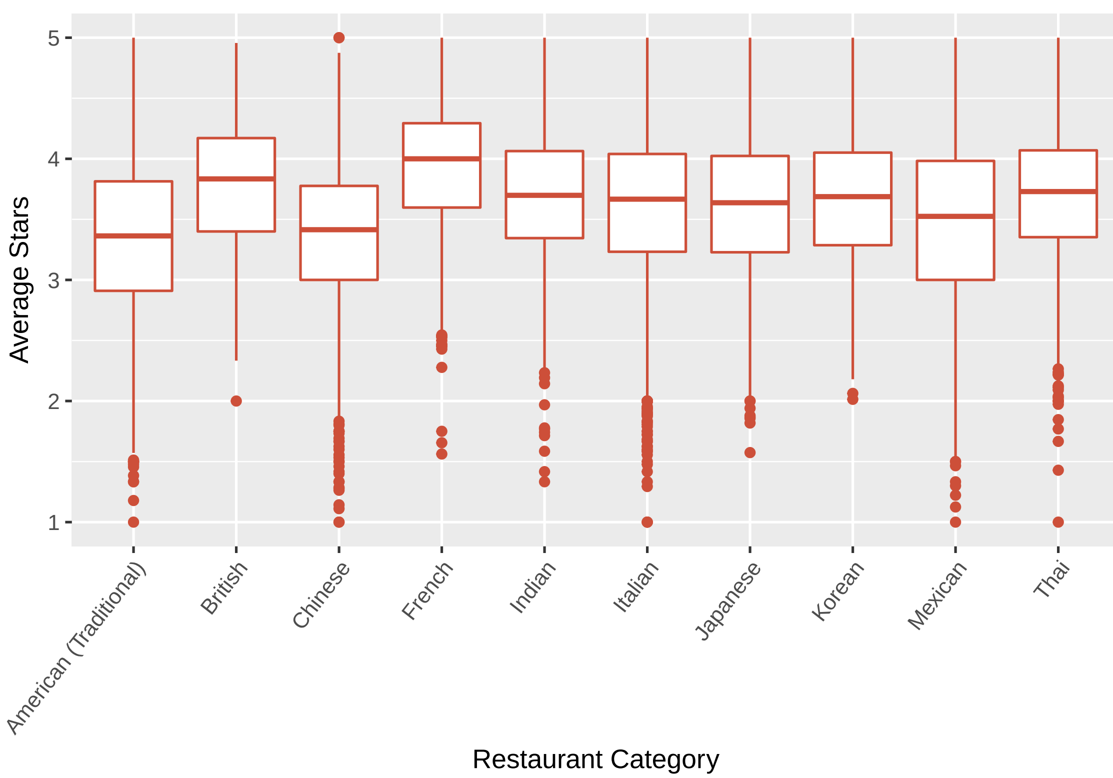
WiFi: 2—free; 1—paid; 0—no.

Alcohol: 2—full_bar; 1—beer_and_wine; 0—none.

And there are also some binary variables that measure whether a restaurant has something(TV, Delivery): 0 for no and 1 for yes.

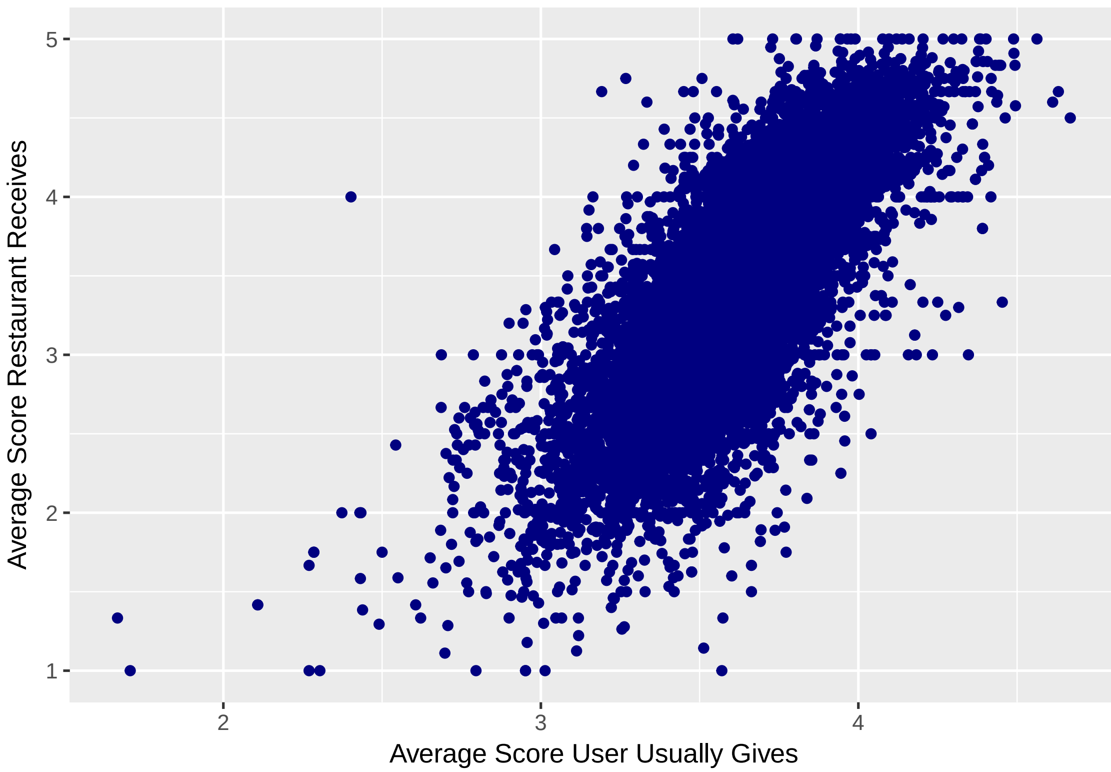
2. EDA

2.1 Boxplot of Average stars for each category



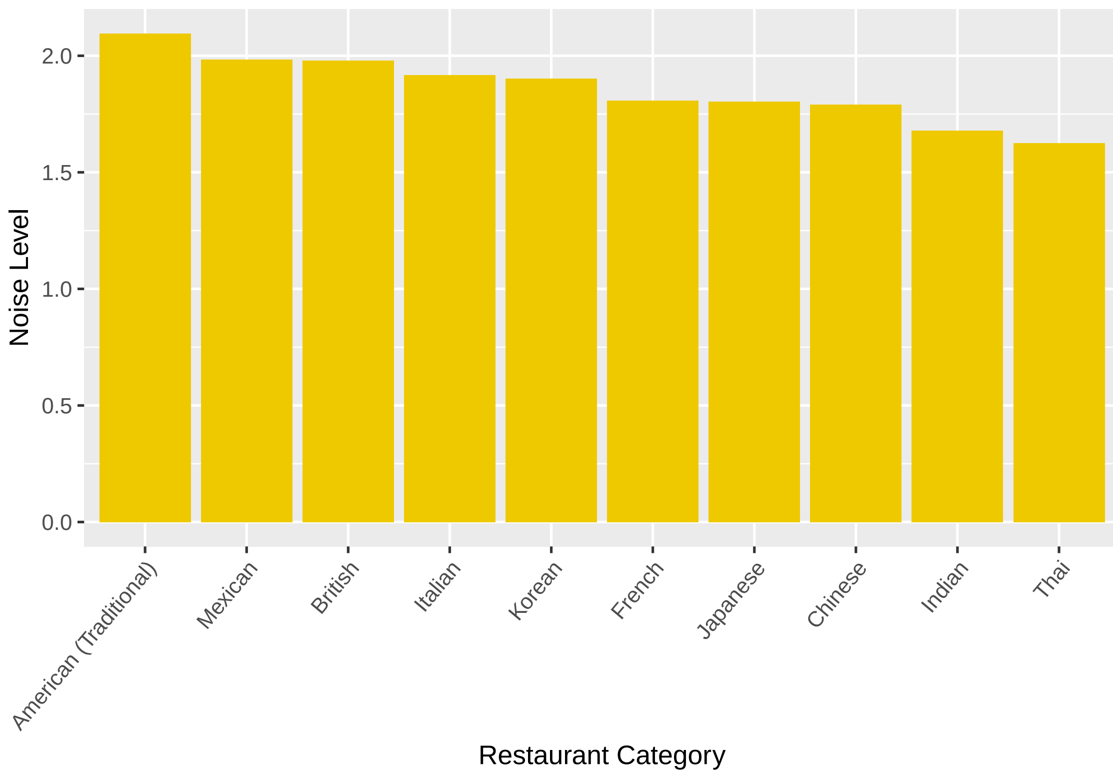
This boxplot shows the comprison of average stars that restaurant received for each cuisine category. From this plot we can find out that the average stars that each type of cuisine received are basically close to each other. Among these scores, French cuisine has a highest mean score, While American Tradisional cuisine and Chinese cuisine receive relatively lower mean scores. Also, there exist some outliers for each category, which means that those scores are numerically distant from the rest of data.

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, 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.

3. Model Selection

For the Model Selection part, I tested 3 models in total and try to find the best one.

3.1 First model —fit1, I treat category as a random effect and fit the model, which will fit the data into 10 regression models (10 categories) with different intercepts. And these are the summary, coeficients and residual plot of fit1:

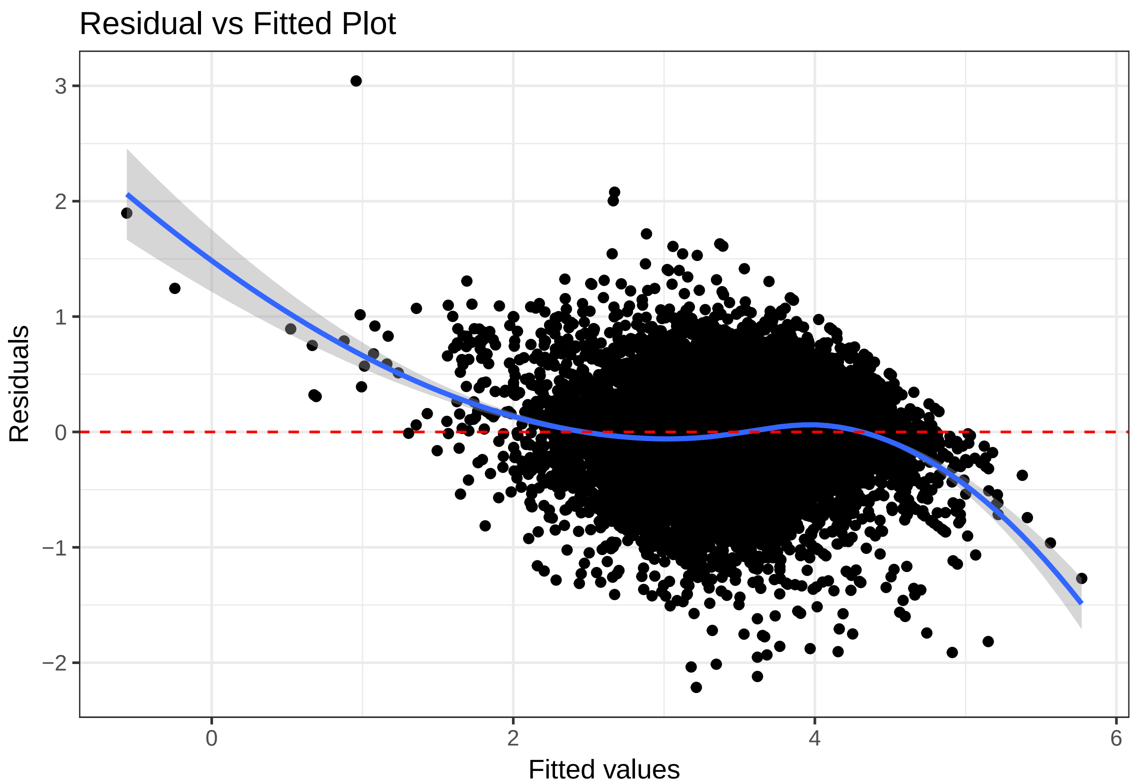
```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## average_stars ~ Avuser + Delivery + WiFi + TV + Alcohol + OutdoorSeating +
```

```
##     Reservations + Attire + noiselevel + (1 | category)
##    Data: yelp
##
## REML criterion at convergence: 15399.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.6792 -0.5834  0.0677  0.6380  7.8037
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  category (Intercept) 0.00454  0.06738
##  Residual             0.15197  0.38983
## Number of obs: 16016, groups:  category, 10
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)     -3.614511   0.054148  -66.75
## Avuser           1.971115   0.013272  148.51
## Delivery1        0.009538   0.008040    1.19
## WiFi1           -0.148282   0.040281   -3.68
## WiFi2           -0.017581   0.006835   -2.57
## TV1              0.001692   0.006892    0.25
## Alcohol1        -0.006569   0.009421   -0.70
## Alcohol2        -0.054612   0.008971   -6.09
## OutdoorSeating1 -0.043040   0.006984   -6.16
## Reservations1    0.032885   0.007606    4.32
## Attire2          0.149424   0.018901    7.91
## Attire3         -0.058778   0.085475   -0.69
## noiselevel2     -0.006898   0.008150   -0.85
## noiselevel3     -0.076717   0.013510   -5.68
## noiselevel4     -0.183636   0.022256   -8.25
```

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

```
## $category
##                        (Intercept)   Avuser   Delivery1      WiFi1
## American (Traditional)   -3.718730 1.971115 0.009538311 -0.1482825
## British                  -3.599791 1.971115 0.009538311 -0.1482825
## Chinese                  -3.667772 1.971115 0.009538311 -0.1482825
## French                   -3.517832 1.971115 0.009538311 -0.1482825
## Indian                   -3.524654 1.971115 0.009538311 -0.1482825
## Italian                  -3.606222 1.971115 0.009538311 -0.1482825
## Japanese                 -3.624829 1.971115 0.009538311 -0.1482825
## Korean                   -3.572768 1.971115 0.009538311 -0.1482825
## Mexican                  -3.690074 1.971115 0.009538311 -0.1482825
## Thai                     -3.622438 1.971115 0.009538311 -0.1482825
##                              WiFi2         TV1     Alcohol1    Alcohol2
## American (Traditional) -0.01758065 0.001692426 -0.006569049 -0.05461193
## British                -0.01758065 0.001692426 -0.006569049 -0.05461193
## Chinese                -0.01758065 0.001692426 -0.006569049 -0.05461193
## French                 -0.01758065 0.001692426 -0.006569049 -0.05461193
```

```
## Indian                 -0.01758065 0.001692426 -0.006569049 -0.05461193
## Italian                -0.01758065 0.001692426 -0.006569049 -0.05461193
## Japanese               -0.01758065 0.001692426 -0.006569049 -0.05461193
## Korean                 -0.01758065 0.001692426 -0.006569049 -0.05461193
## Mexican                -0.01758065 0.001692426 -0.006569049 -0.05461193
## Thai                   -0.01758065 0.001692426 -0.006569049 -0.05461193
##                        OutdoorSeating1 Reservations1  Attire2     Attire3
## American (Traditional)      -0.0430398    0.03288493 0.149424 -0.05877775
## British                     -0.0430398    0.03288493 0.149424 -0.05877775
## Chinese                     -0.0430398    0.03288493 0.149424 -0.05877775
## French                      -0.0430398    0.03288493 0.149424 -0.05877775
## Indian                      -0.0430398    0.03288493 0.149424 -0.05877775
## Italian                     -0.0430398    0.03288493 0.149424 -0.05877775
## Japanese                    -0.0430398    0.03288493 0.149424 -0.05877775
## Korean                      -0.0430398    0.03288493 0.149424 -0.05877775
## Mexican                     -0.0430398    0.03288493 0.149424 -0.05877775
## Thai                        -0.0430398    0.03288493 0.149424 -0.05877775
##                         noiselevel2 noiselevel3 noiselevel4
## American (Traditional) -0.006898223 -0.07671707  -0.1836364
## British                -0.006898223 -0.07671707  -0.1836364
## Chinese                -0.006898223 -0.07671707  -0.1836364
## French                 -0.006898223 -0.07671707  -0.1836364
## Indian                 -0.006898223 -0.07671707  -0.1836364
## Italian                -0.006898223 -0.07671707  -0.1836364
## Japanese               -0.006898223 -0.07671707  -0.1836364
## Korean                 -0.006898223 -0.07671707  -0.1836364
## Mexican                -0.006898223 -0.07671707  -0.1836364
## Thai                   -0.006898223 -0.07671707  -0.1836364
##
## attr(,"class")
## [1] "coef.mer"
```



From the information above, we can see that for varaible TV, Estimate coefficient is 0.001692 which is smaller than 2 times of standard deviation 0.006892;and for variable Delivery, Estimate coefficient is 0.009538 which is smaller than 2 times of standard deviation 0.008040. It means that these variables are not significant enough, so we will probably remove them from the model. For other varaibles in the model, although the coefficient are relatively small but siginificant.

Those coefficients stay the same while intercepts vary for different types of restaurants, but those intercepts only vary a litttle and they are close to each other.

For the residual plot, although there are some points with relatively large variances, most of the points are clustered around the zero line, which means the model we used is 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 these are the summary, coeficients and residual plot of fit2:

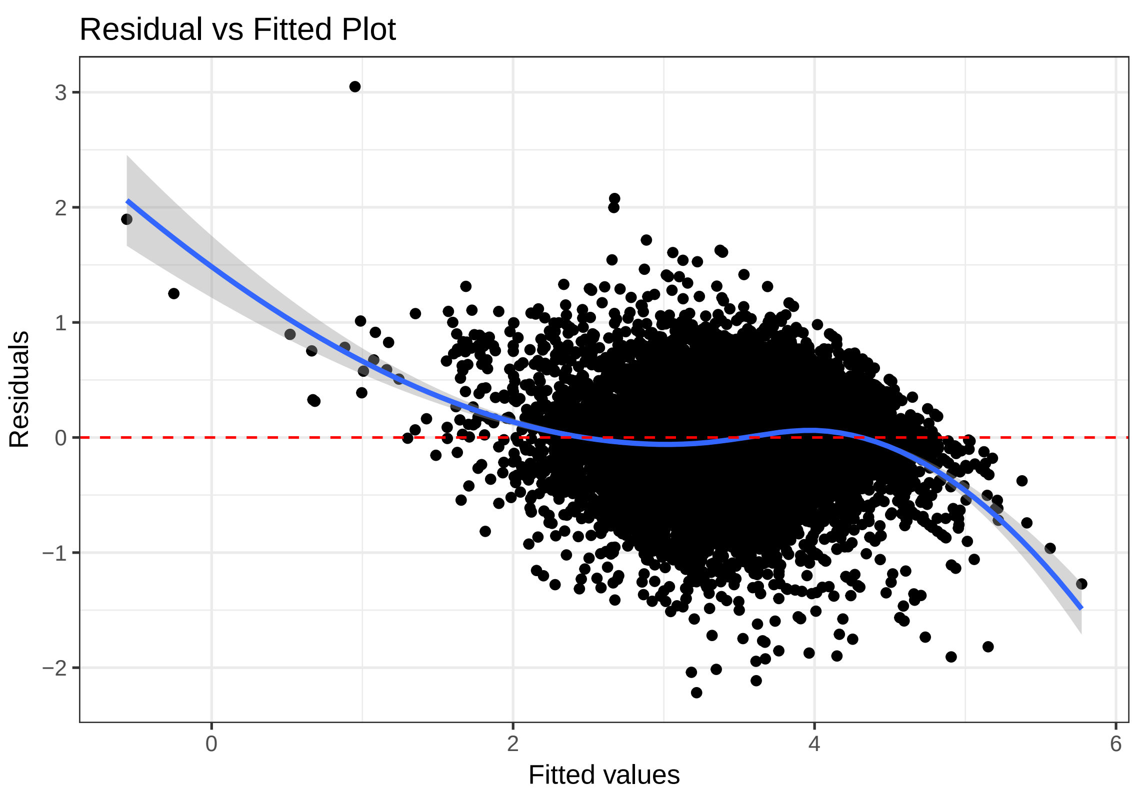
```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## average_stars ~ Avuser + WiFi + Alcohol + OutdoorSeating + Reservations +
##     Attire + noiselevel + (1 | category)
##    Data: yelp
##
## REML criterion at convergence: 15384.9
##
## Scaled residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -5.6884 -0.5864  0.0692  0.6392  7.8207
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  category (Intercept) 0.004552 0.06747
##  Residual             0.151961 0.38982
## Number of obs: 16016, groups:  category, 10
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)     -3.608855   0.053948  -66.89
## Avuser           1.970571   0.013250  148.72
## WiFi1           -0.148564   0.040279   -3.69
## WiFi2           -0.016655   0.006726   -2.48
## Alcohol1        -0.006245   0.009364   -0.67
## Alcohol2        -0.054584   0.008651   -6.31
## OutdoorSeating1 -0.043254   0.006978   -6.20
## Reservations1    0.032964   0.007599    4.34
## Attire2          0.147292   0.018723    7.87
## Attire3         -0.059312   0.085472   -0.69
## noiselevel2     -0.008015   0.008089   -0.99
## noiselevel3     -0.078429   0.013433   -5.84
## noiselevel4     -0.185213   0.022206   -8.34

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

## $category
##                         (Intercept)   Avuser      WiFi1       WiFi2
## American (Traditional)   -3.714063 1.970571 -0.1485641 -0.01665459
## British                  -3.595519 1.970571 -0.1485641 -0.01665459
## Chinese                  -3.660887 1.970571 -0.1485641 -0.01665459
## French                   -3.513879 1.970571 -0.1485641 -0.01665459
## Indian                   -3.517222 1.970571 -0.1485641 -0.01665459
## Italian                  -3.599408 1.970571 -0.1485641 -0.01665459
## Japanese                 -3.619252 1.970571 -0.1485641 -0.01665459
## Korean                   -3.567938 1.970571 -0.1485641 -0.01665459
## Mexican                  -3.685507 1.970571 -0.1485641 -0.01665459
## Thai                     -3.614877 1.970571 -0.1485641 -0.01665459
##                            Alcohol1    Alcohol2 OutdoorSeating1
## American (Traditional) -0.006245477 -0.05458407     -0.04325392
## British                -0.006245477 -0.05458407     -0.04325392
## Chinese                -0.006245477 -0.05458407     -0.04325392
## French                 -0.006245477 -0.05458407     -0.04325392
## Indian                 -0.006245477 -0.05458407     -0.04325392
## Italian                -0.006245477 -0.05458407     -0.04325392
## Japanese               -0.006245477 -0.05458407     -0.04325392
## Korean                 -0.006245477 -0.05458407     -0.04325392
## Mexican                -0.006245477 -0.05458407     -0.04325392
## Thai                   -0.006245477 -0.05458407     -0.04325392
##                        Reservations1   Attire2     Attire3  noiselevel2
## American (Traditional)    0.03296439 0.1472922 -0.05931195 -0.008015247
```

```
## British                    0.03296439 0.1472922 -0.05931195 -0.008015247
## Chinese                    0.03296439 0.1472922 -0.05931195 -0.008015247
## French                     0.03296439 0.1472922 -0.05931195 -0.008015247
## Indian                     0.03296439 0.1472922 -0.05931195 -0.008015247
## Italian                    0.03296439 0.1472922 -0.05931195 -0.008015247
## Japanese                   0.03296439 0.1472922 -0.05931195 -0.008015247
## Korean                     0.03296439 0.1472922 -0.05931195 -0.008015247
## Mexican                    0.03296439 0.1472922 -0.05931195 -0.008015247
## Thai                       0.03296439 0.1472922 -0.05931195 -0.008015247
##                         noiselevel3 noiselevel4
## American (Traditional) -0.07842861  -0.1852134
## British                -0.07842861  -0.1852134
## Chinese                -0.07842861  -0.1852134
## French                 -0.07842861  -0.1852134
## Indian                 -0.07842861  -0.1852134
## Italian                -0.07842861  -0.1852134
## Japanese               -0.07842861  -0.1852134
## Korean                 -0.07842861  -0.1852134
## Mexican                -0.07842861  -0.1852134
## Thai                   -0.07842861  -0.1852134
##
## attr(,"class")
## [1] "coef.mer"
```



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. 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. In this way, my next step will be allowing it to vary with group, which leads to my third model.

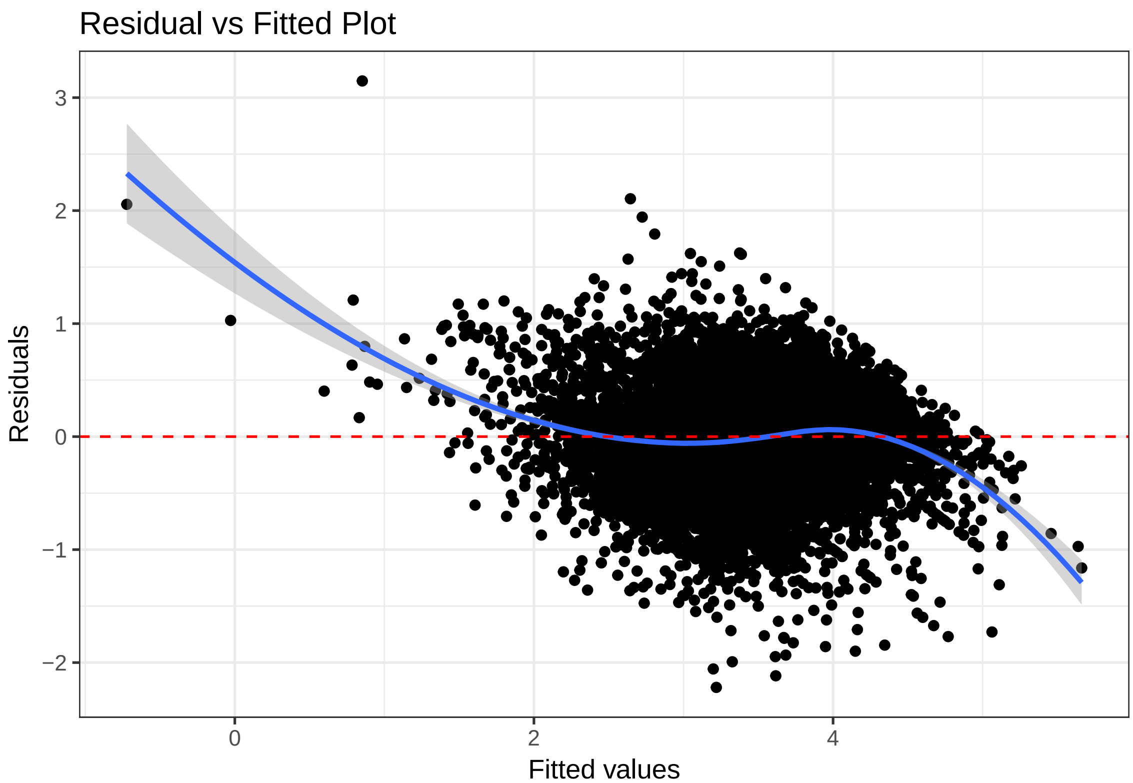
3.3 Third model —fit3, I treat category with Avuser as random effect and fit the model, which will fit the data into 10 regression models (10 categories) with different intercepts AND DIFFERENT COEFFICIENTS for the variable of average score of user. And these are the summary, coeficients and residual plot of fit3:

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## average_stars ~ Avuser + WiFi + Alcohol + OutdoorSeating + Reservations +
##     Attire + noiselevel + (1 + Avuser | category)
##    Data: yelp
##
## REML criterion at convergence: 15271.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.7171 -0.5802  0.0679  0.6388  8.1072
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  category (Intercept) 0.3287   0.5733
##           Avuser      0.0208   0.1442   -1.00
##  Residual             0.1507   0.3882
## Number of obs: 16016, groups:  category, 10
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)     -3.448201   0.195817  -17.61
## Avuser           1.927685   0.049653   38.82
## WiFi1           -0.150260   0.040122   -3.75
## WiFi2           -0.016026   0.006706   -2.39
## Alcohol1        -0.007580   0.009335   -0.81
## Alcohol2        -0.055804   0.008625   -6.47
## OutdoorSeating1 -0.042701   0.006952   -6.14
## Reservations1    0.031313   0.007575    4.13
## Attire2          0.148887   0.018656    7.98
## Attire3         -0.039896   0.085270   -0.47
## noiselevel2     -0.010553   0.008063   -1.31
## noiselevel3     -0.078718   0.013387   -5.88
## noiselevel4     -0.188221   0.022123   -8.51

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

## $category
##                         (Intercept)   Avuser      WiFi1       WiFi2
```

```
## American (Traditional)   -4.008723 2.052659 -0.1502597 -0.01602557
## British                  -3.502457 1.946151 -0.1502597 -0.01602557
## Chinese                  -3.019219 1.792607 -0.1502597 -0.01602557
## French                   -3.053588 1.849031 -0.1502597 -0.01602557
## Indian                   -2.831469 1.783849 -0.1502597 -0.01602557
## Italian                  -3.180053 1.857497 -0.1502597 -0.01602557
## Japanese                 -3.598623 1.965853 -0.1502597 -0.01602557
## Korean                   -3.074972 1.837744 -0.1502597 -0.01602557
## Mexican                  -4.566839 2.211838 -0.1502597 -0.01602557
## Thai                     -3.646062 1.979616 -0.1502597 -0.01602557
##                            Alcohol1   Alcohol2 OutdoorSeating1
## American (Traditional) -0.007579976 -0.0558041       -0.042701
## British                -0.007579976 -0.0558041       -0.042701
## Chinese                -0.007579976 -0.0558041       -0.042701
## French                 -0.007579976 -0.0558041       -0.042701
## Indian                 -0.007579976 -0.0558041       -0.042701
## Italian                -0.007579976 -0.0558041       -0.042701
## Japanese               -0.007579976 -0.0558041       -0.042701
## Korean                 -0.007579976 -0.0558041       -0.042701
## Mexican                -0.007579976 -0.0558041       -0.042701
## Thai                   -0.007579976 -0.0558041       -0.042701
##                        Reservations1   Attire2     Attire3 noiselevel2
## American (Traditional)    0.03131302 0.1488873 -0.03989624 -0.01055276
## British                   0.03131302 0.1488873 -0.03989624 -0.01055276
## Chinese                   0.03131302 0.1488873 -0.03989624 -0.01055276
## French                    0.03131302 0.1488873 -0.03989624 -0.01055276
## Indian                    0.03131302 0.1488873 -0.03989624 -0.01055276
## Italian                   0.03131302 0.1488873 -0.03989624 -0.01055276
## Japanese                  0.03131302 0.1488873 -0.03989624 -0.01055276
## Korean                    0.03131302 0.1488873 -0.03989624 -0.01055276
## Mexican                   0.03131302 0.1488873 -0.03989624 -0.01055276
## Thai                      0.03131302 0.1488873 -0.03989624 -0.01055276
##                        noiselevel3 noiselevel4
## American (Traditional) -0.07871807  -0.1882209
## British                -0.07871807  -0.1882209
## Chinese                -0.07871807  -0.1882209
## French                 -0.07871807  -0.1882209
## Indian                 -0.07871807  -0.1882209
## Italian                -0.07871807  -0.1882209
## Japanese               -0.07871807  -0.1882209
## Korean                 -0.07871807  -0.1882209
## Mexican                -0.07871807  -0.1882209
## Thai                   -0.07871807  -0.1882209
##
## attr(,"class")
## [1] "coef.mer"
```



From the results of model fit3, compare to last two model, the significance of the coefficient of noiselevel 2 has increased. Also from the coefficient table we can see, the intercepts vary more between different categories and the coefficient of Avuser also varies. 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, compare to other types of restaurants.

4. Conclusion

4.1 Results

Based on the analysis above, we 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, especially the variable Avuser, which has a huge impact on a restaurant's star.

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.