678 Final Project Report

Haochen Pan

2022-12-05

Part1 Abstract

In the database recommended by Japanese animation, I found data about a large number of animation ratings, including numerical data such as animation scores, number of participants, and variables of more than 10 groups such as producer and source. This report will use multilevel modeling and other methods to analyze: 1. how popularity and number of community fans and other variables affect the rating of anime; 2. what variables cause high rating animes. The first 10 page will focus on multilevel modeling and other useful visulizations and analysis will in the Appendix part.

Part2 Introduction

The name of the data set from Kaggle is called "anime recommendation", which contains information on user preference data from 73,516 users on 12,294 anime. In the excel from the data, some useful variables can be used to find which variables can affect the rating of animes and how they exactly affect the rating. The following contents are the explanation of each variable.

Anime_id: myanimelist.net's unique id identifying an anime. Title: full name of anime. Type: Movie, TV, Special, etc. Producer: Different producer companies that produce the anime. Studio: The creator company of anime. Rating: Average rating out of 10 for this anime. ScoreBy: Number of people who rate the anime. Popularity: The popularity of the anime(the lower number means more famous). Members: Number of community fans that are in this anime's "group". Episodes: How many episodes in this show. (1 if movie). Source: The source of the anime, including Manga, original, etc. Aired: The date that the anime start to show.

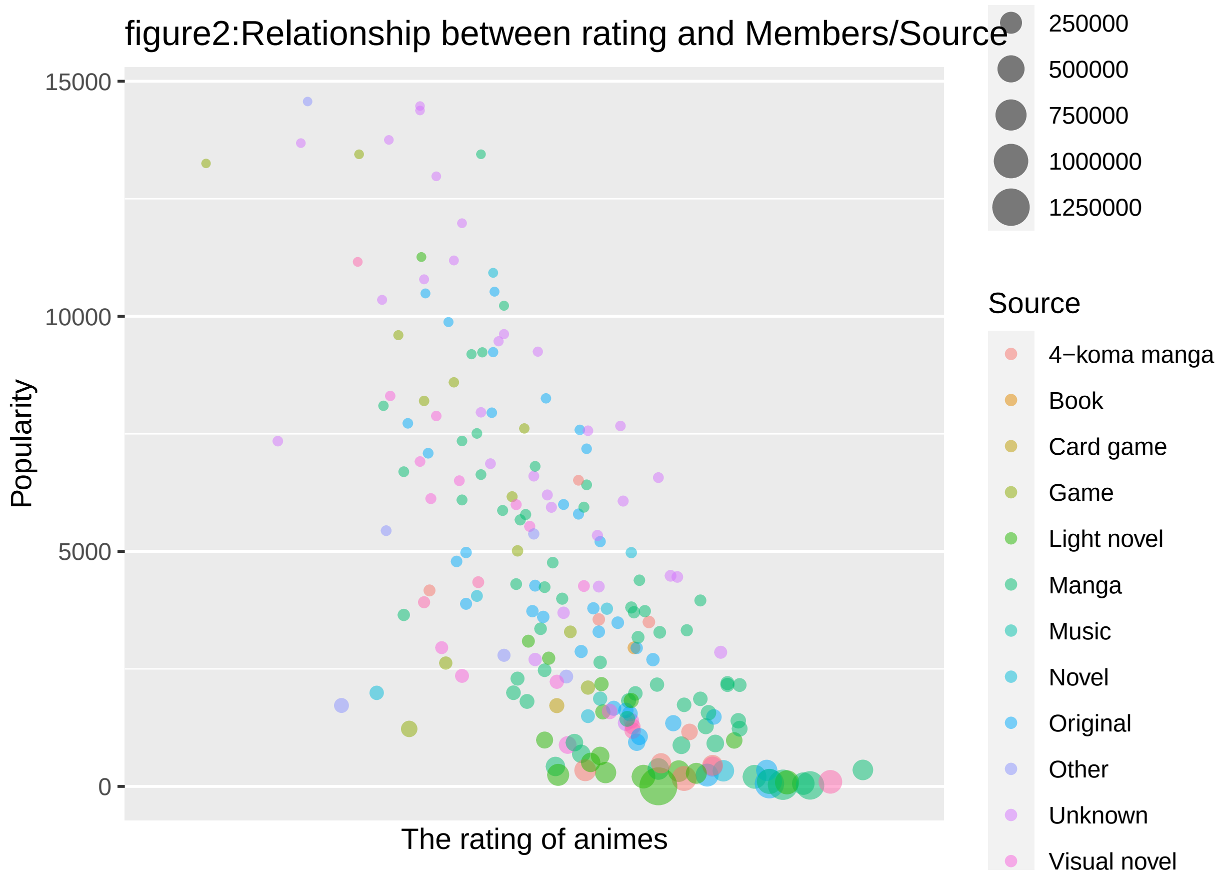
Part3 Methods

Data Cleaning

In the uploading part, I delete the missing data and NAs.Then I choose 2 dataframes, the raw data and the data ranked by top-50-rating.

EDA

Before the linear regression and modeling, I first use some variables mentioned above to make graphs to predict which methods are more useful in the next parts. 1 of the graphs is used more are in part 6.



In figure 1 to 3, I randomly select 100 of the animes from the data to see the changes of ratings and how other variables can affect them. In a short conclusion, some sources like manga, some types like TV, more popularity and more members seems to have a higher rating in the graph. While the scoring people will not have strong relationship, As a result, I plan to focus on this variables.

Linear Model

First, I use linear model to analysis how Popularity/Members affect Rating. The graphs/summary/conclusion are in part 6.

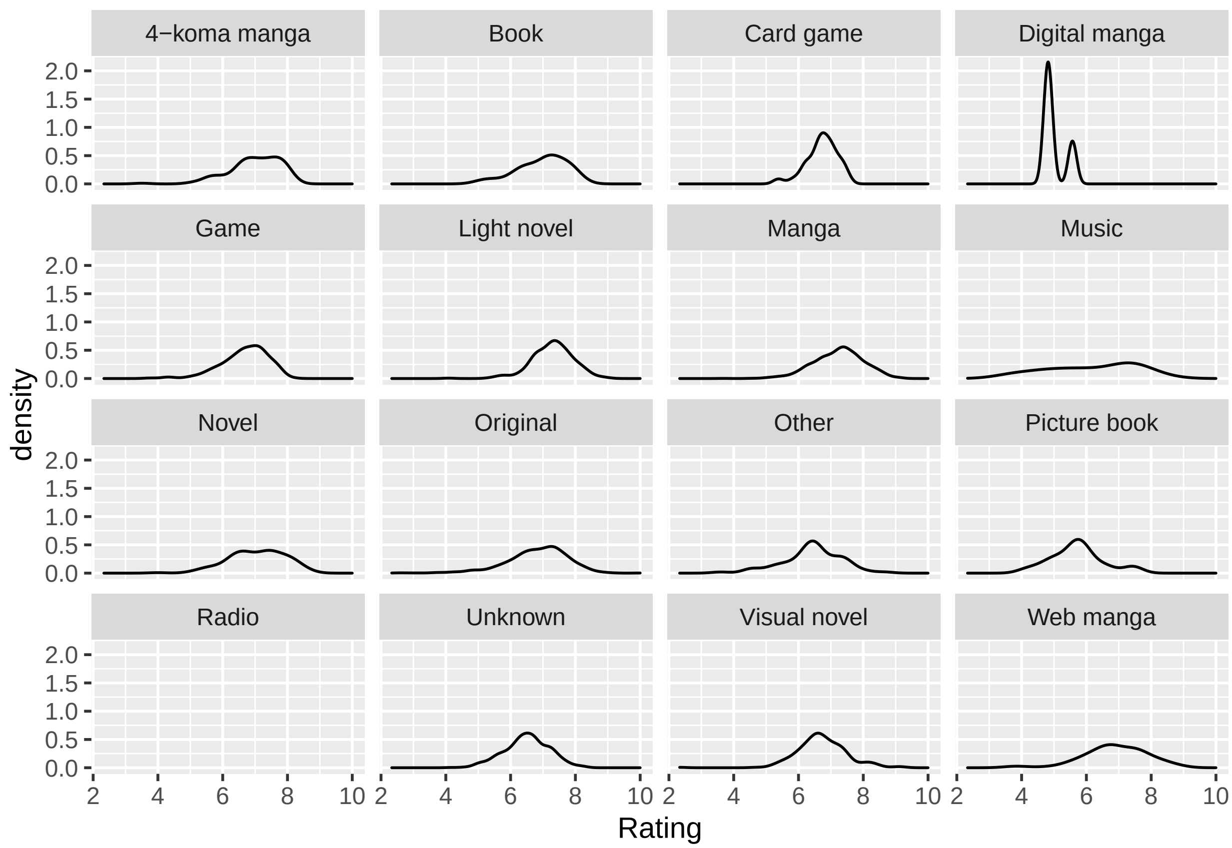
Multilevel Modeling

Then, based on the requirement, i use the multilevel model to analysis and predict the data.Due to the page limit, I will choose Source/Type for different model and visualization. Analysis about what are missing are in part 6.

First, I check the distributions by source.

```
## # A tibble: 16 x 4
##    Source          mean    SD  miss
##    <chr>          <dbl> <dbl> <dbl>
##  1 4-koma manga    7     7        0
##  2 Book            6.97  6.97     0
##  3 Card game       6.72  6.72     0
```

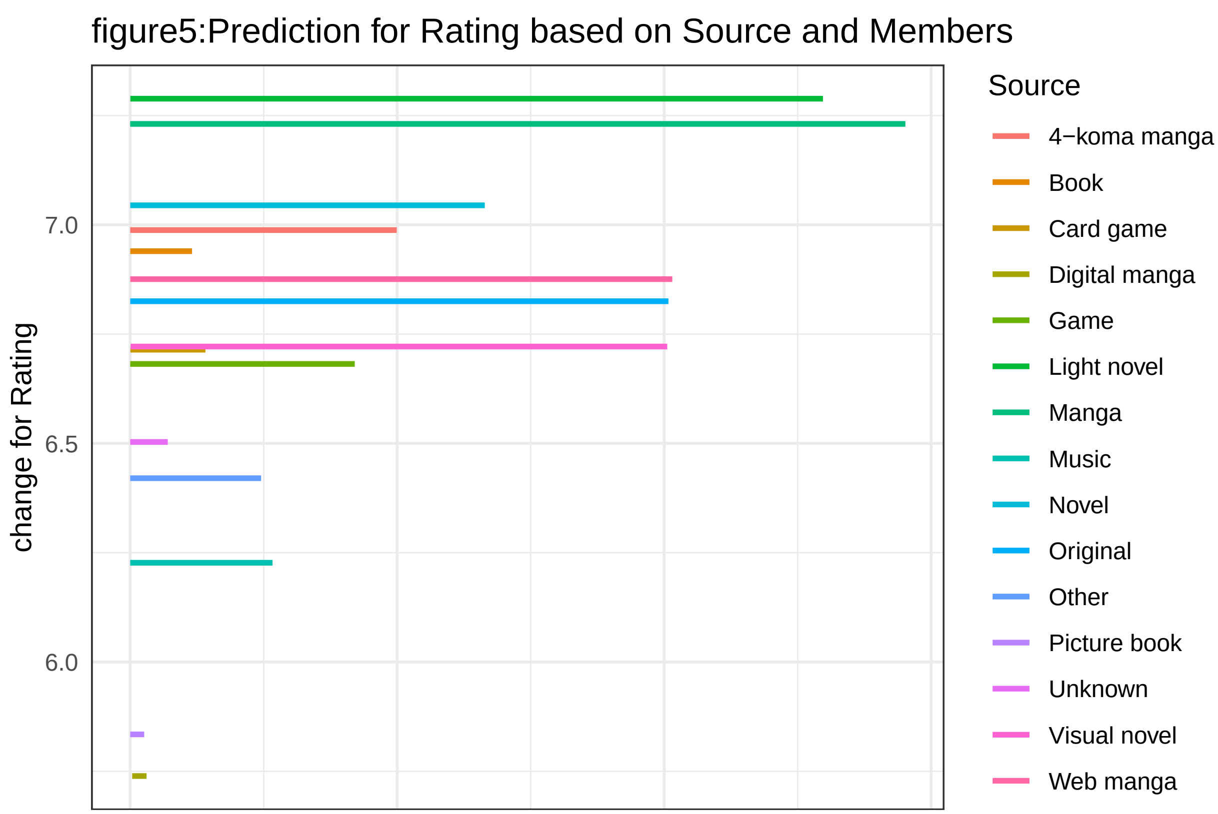
```
##   4 Digital manga  5.01  5.01     0
##   5 Game           6.68  6.68     0
##   6 Light novel    7.29  7.29     0
##   7 Manga          7.23  7.23     0
##   8 Music          6.19  6.19     0
##   9 Novel          7.05  7.05     0
##  10 Original       6.83  6.83     0
##  11 Other          6.41  6.41     0
##  12 Picture book   5.72  5.72     0
##  13 Radio          6.71  6.71     0
##  14 Unknown        6.5   6.5      0
##  15 Visual novel   6.72  6.72     0
##  16 Web manga      6.88  6.88     0
```



Then I tried lmer model with Source

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + (1 | Source)
##    Data: data2
##
##      AIC      BIC   logLik deviance df.resid
##  10939.2  10958.5  -5466.6  10933.2     4520
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -5.4525 -0.5952  0.0674  0.6696  3.9419
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Source   (Intercept) 0.2062   0.4541
##  Residual             0.6487   0.8054
## Number of obs: 4523, groups:  Source, 16
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   6.6696     0.1204   55.39
```




```
## $Source
```

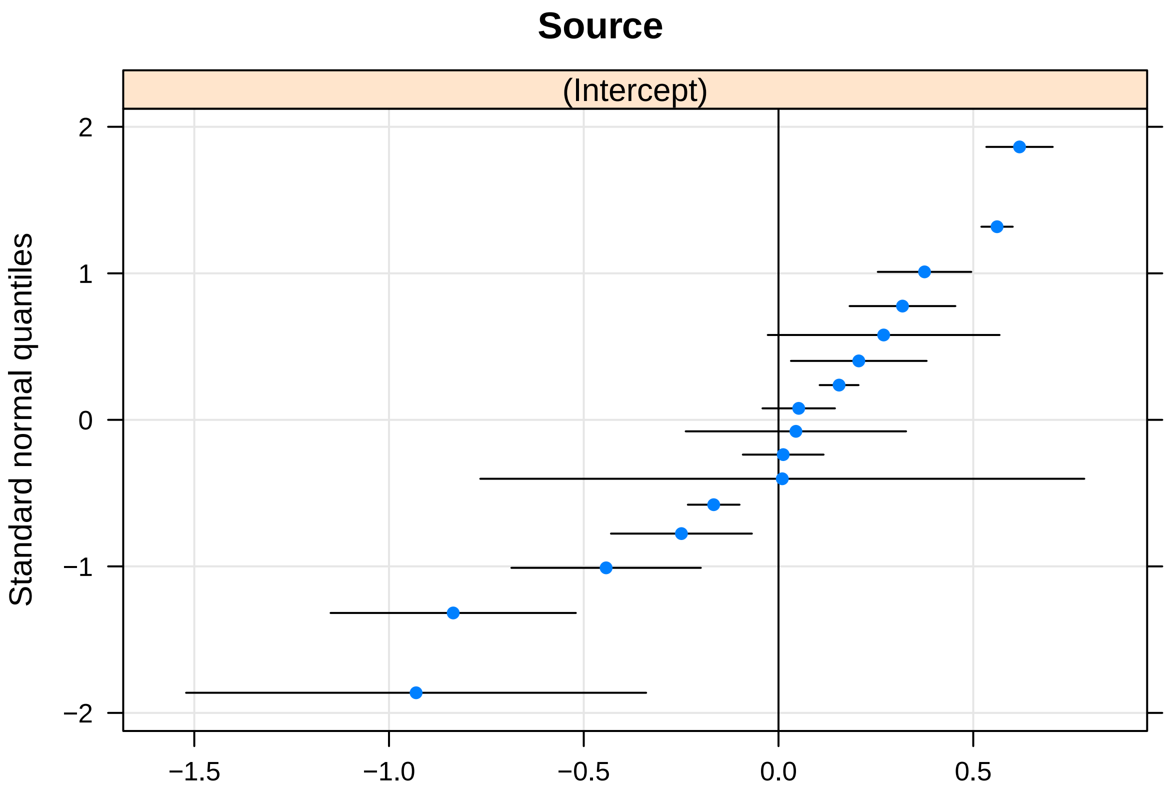
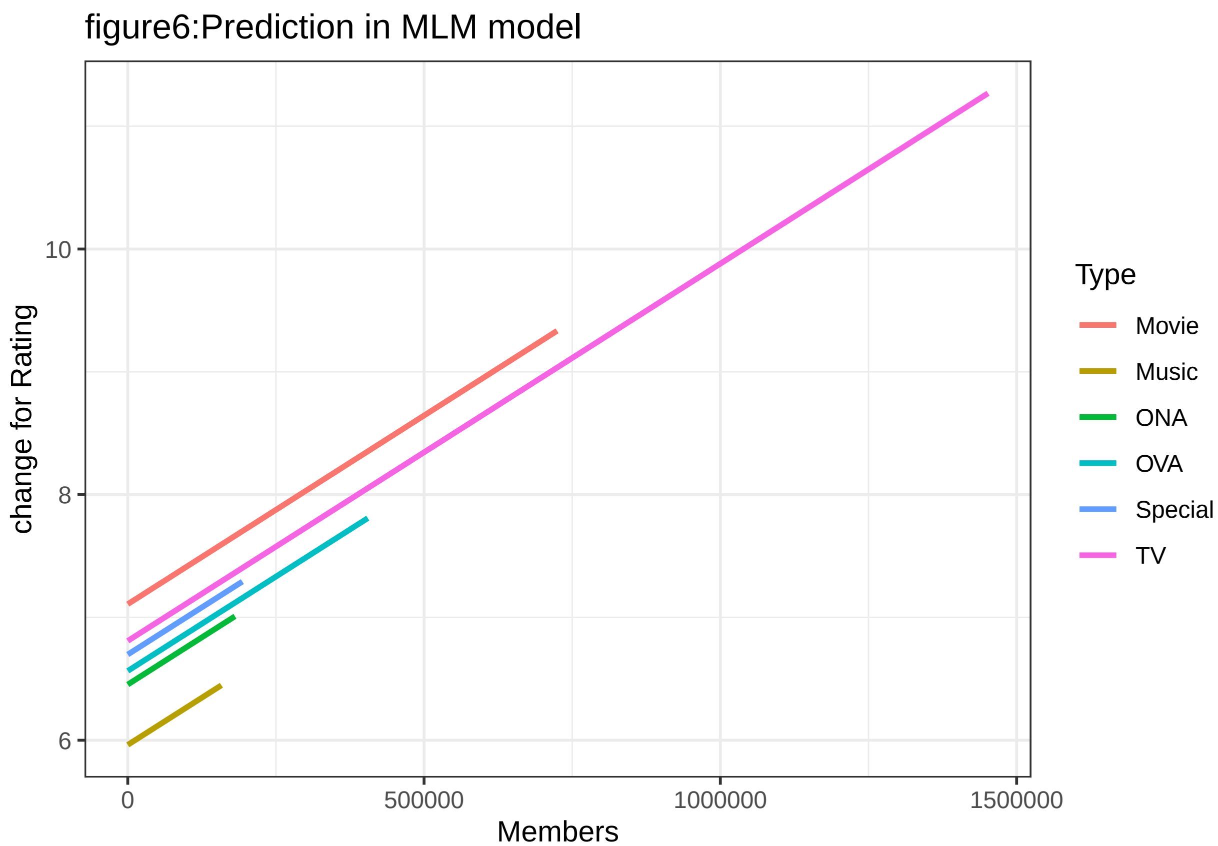


Figure 5 allows each source to have a different average change in rating through the random effect. I also visualize these random effects in a qqmath. In this graph each dot represents a source and the line around it is the confidence interval. The 0 on the x axis is the intercept or expected value. So the most source have values significantly different from that, again indicating that this is a relevant level for the analysis.

Then I try the prediction in MLM models for Type and Member

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + Members + (1 | Type)
##    Data: data2
##
##      AIC      BIC   logLik deviance df.resid
##  10170.3  10196.0  -5081.2  10162.3     4519
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.8502 -0.5443  0.0746  0.6552  4.7796
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Type     (Intercept) 0.1260   0.3549
##  Residual             0.5504   0.7419
## Number of obs: 4523, groups:  Type, 6
##
## Fixed effects:
##              Estimate Std. Error t value
```

```
## (Intercept) 6.599e+00  1.464e-01   45.08
## Members     3.072e-06  9.709e-08   31.64
##
## Correlation of Fixed Effects:
##         (Intr)
## Members -0.021
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

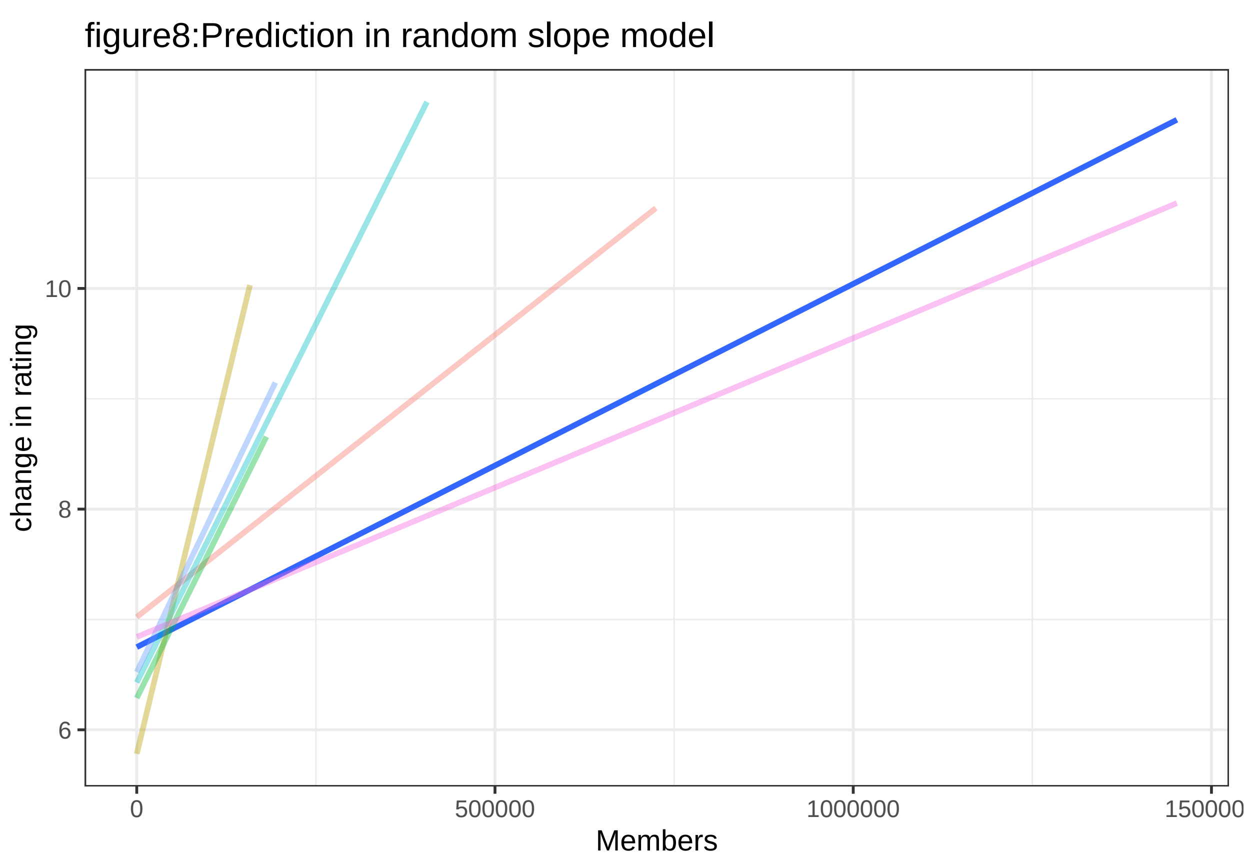


Here the result shows as member of fans increases by 1 the expected increase for rating is 3.072e-06. The intercept now is understood as the expected increase of rating when the independent variable is 0. I visualize the relationship between the two variables as implied by our model in figure 6.

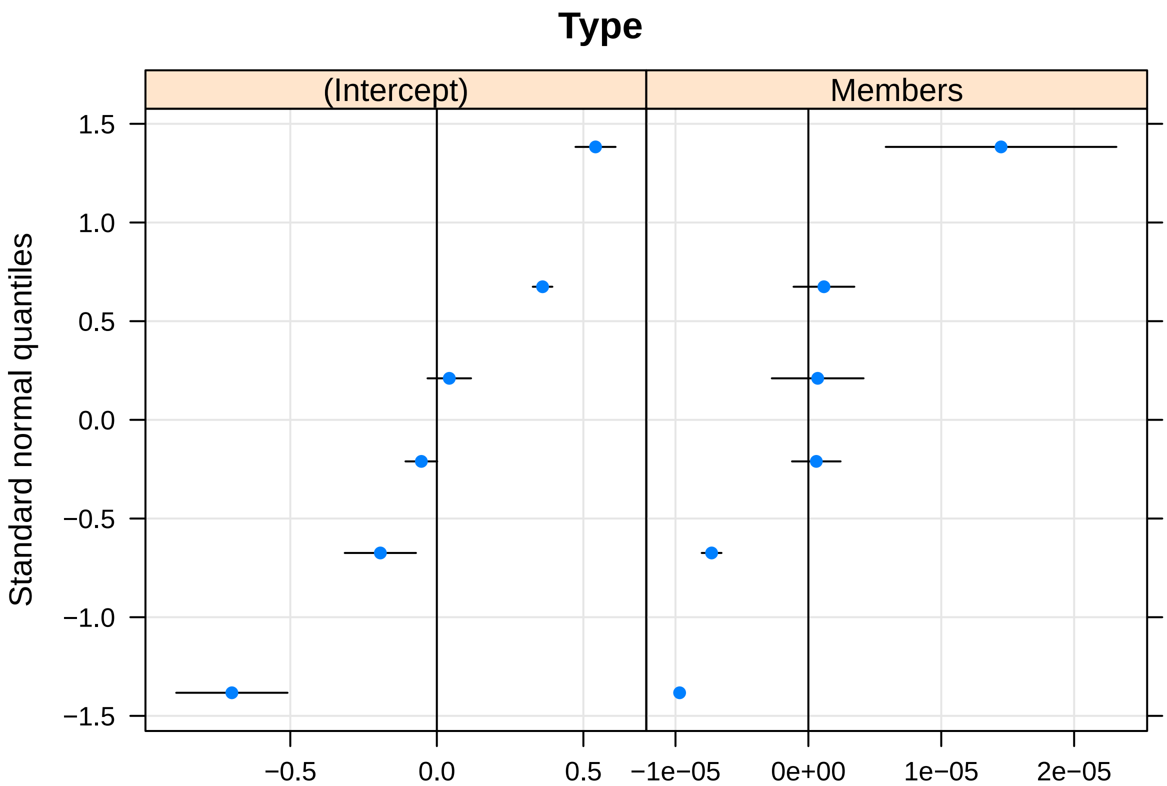
In the next step we try random slope model for Type/Member.

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + Members + (1 + Members | Type)
##    Data: data2
##
##      AIC      BIC   logLik deviance df.resid
##   9941.0   9979.5  -4964.5   9929.0     4517
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.3276 -0.5329  0.0830  0.6547  5.1630
```

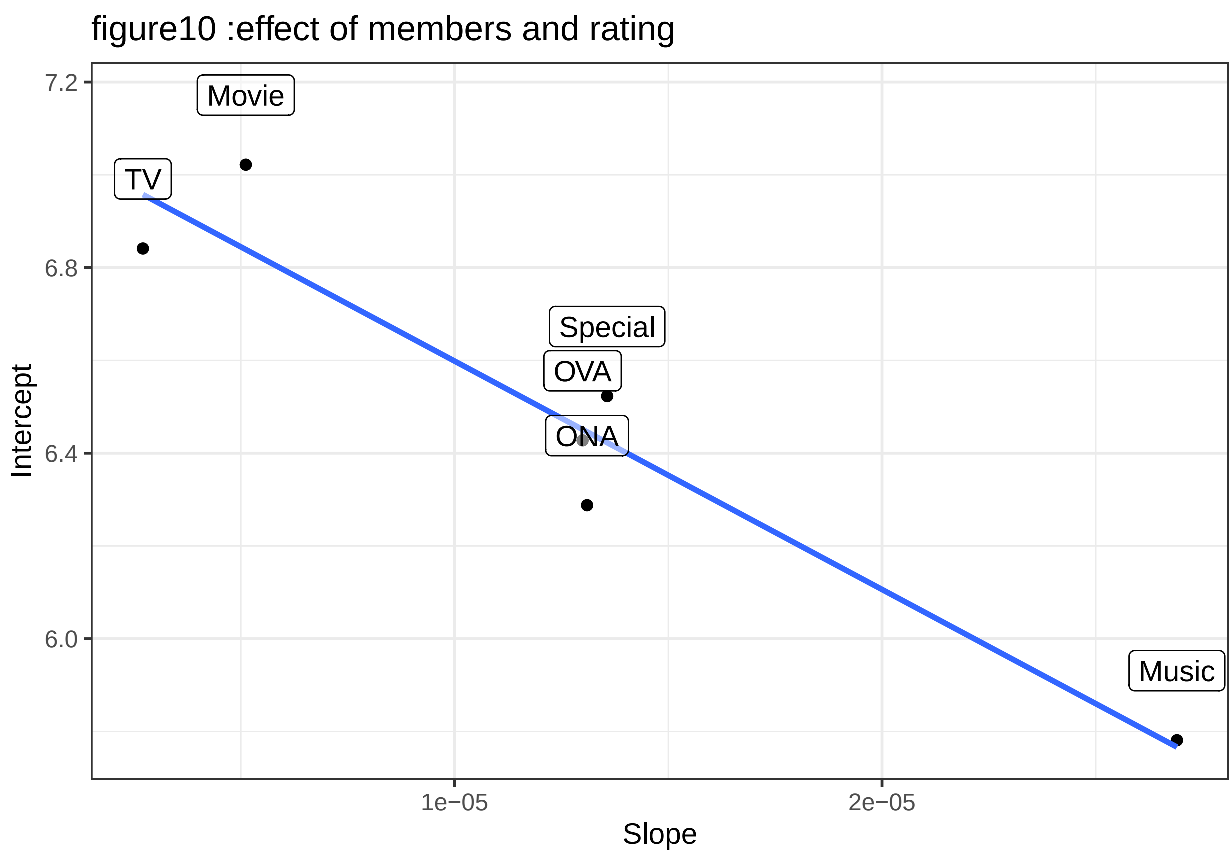
```
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.  Corr
##  Type     (Intercept) 3.151e-01 5.614e-01
##           Members     6.624e-09 8.139e-05 -0.66
##  Residual             5.161e-01 7.184e-01
## Number of obs: 4523, groups:  Type, 6
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 6.480e+00  2.302e-01  28.150
## Members     1.240e-05  3.324e-05   0.373
##
## Correlation of Fixed Effects:
##         (Intr)
## Members -0.658
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```




```
## $Type
```



In the model, the Member variable is slightly smaller and that we have a new coefficient in the random part of the model. The variance of the random slope for member is 1.67e-08.This coefficient is hard to interpret on its own. So I draw another graph which is fig8. Some types' members is more important than others. The dot plot also makes sense.



In figure10 The graphs shows that we can divide the types into 3 parts. TV and Movie have most members and high increasing in rating but low effect of members to outcome. The Music has low support in rating but members for music has strong effect.

K-means algorithm

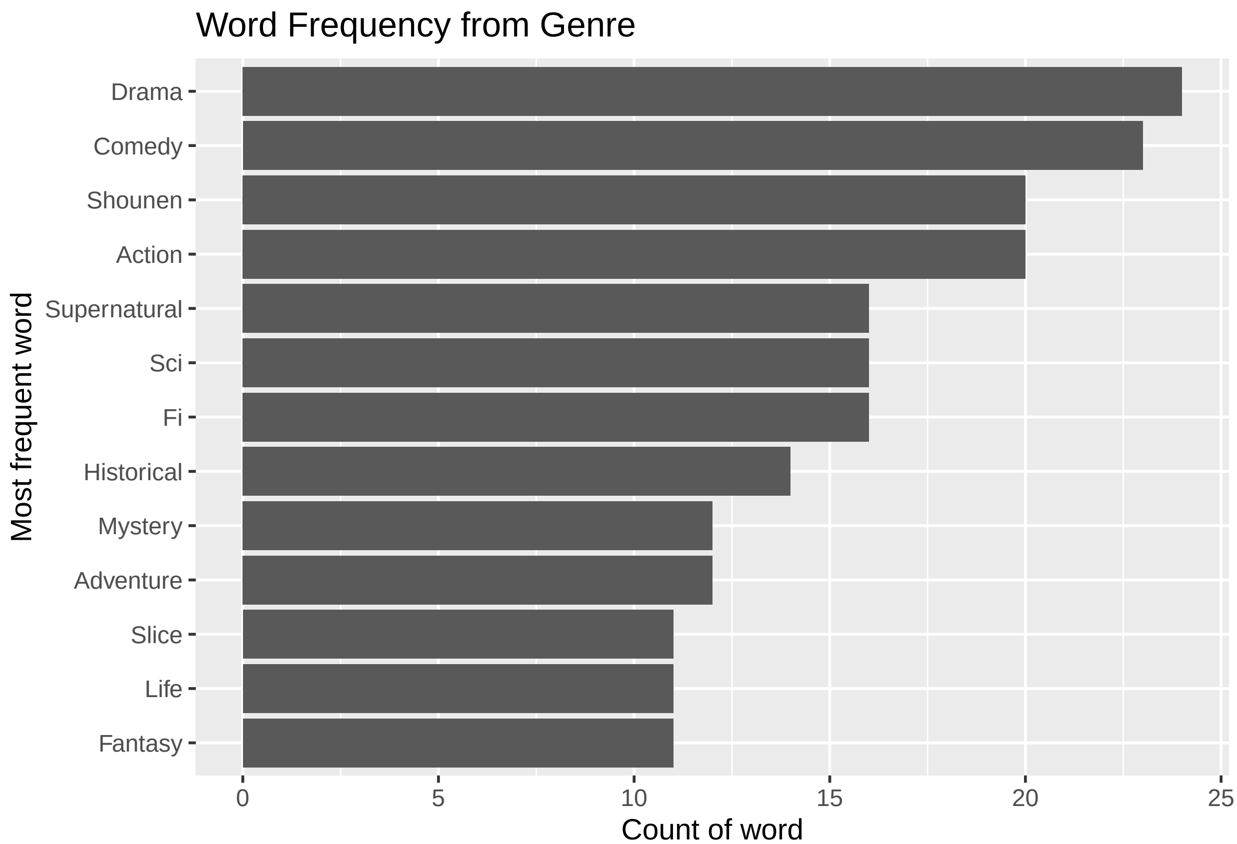
For this part, I want to use k-means to cluster observations and want observations in the same group to be similar and observations in different groups to be dissimilar. The visulization is in part 6.

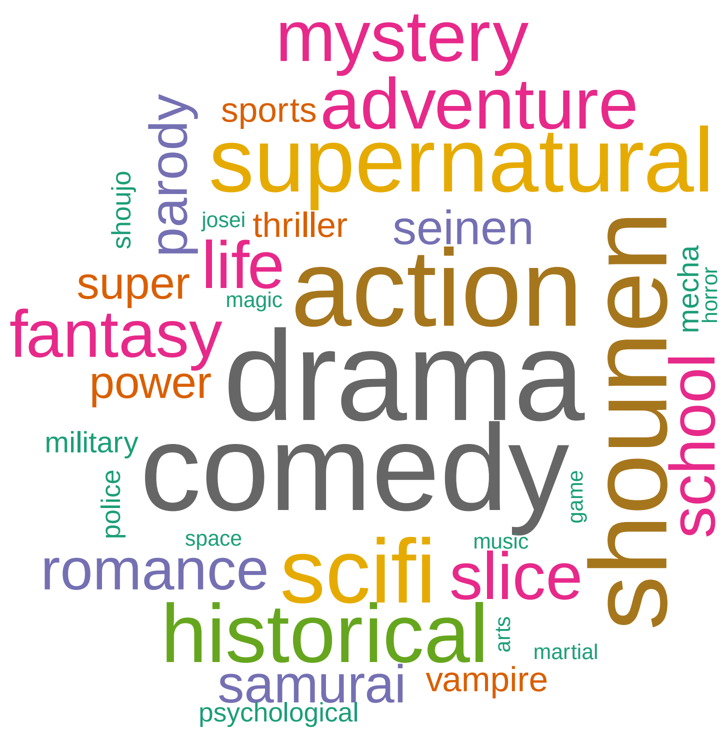
Word Could and Tpoic Modeling

Finally, I tried what I learn in 615 to analysis the non-numerical data.The word cloud shows the hot topics for data_top50. More details are in part 6.

```
## # A tibble: 36 x 2
##    word             n
##    <chr>        <int>
##  1 Drama           24
##  2 Comedy          23
##  3 Action          20
##  4 Shounen         20
##  5 Fi              16
##  6 Sci             16
##  7 Supernatural    16
##  8 Historical      14
```

```
##  9 Adventure       12
## 10 Mystery         12
## # ... with 26 more rows
```





If we compare the topics, we can see what kind of topics in genre are more likely to be shown in the top rating animes.

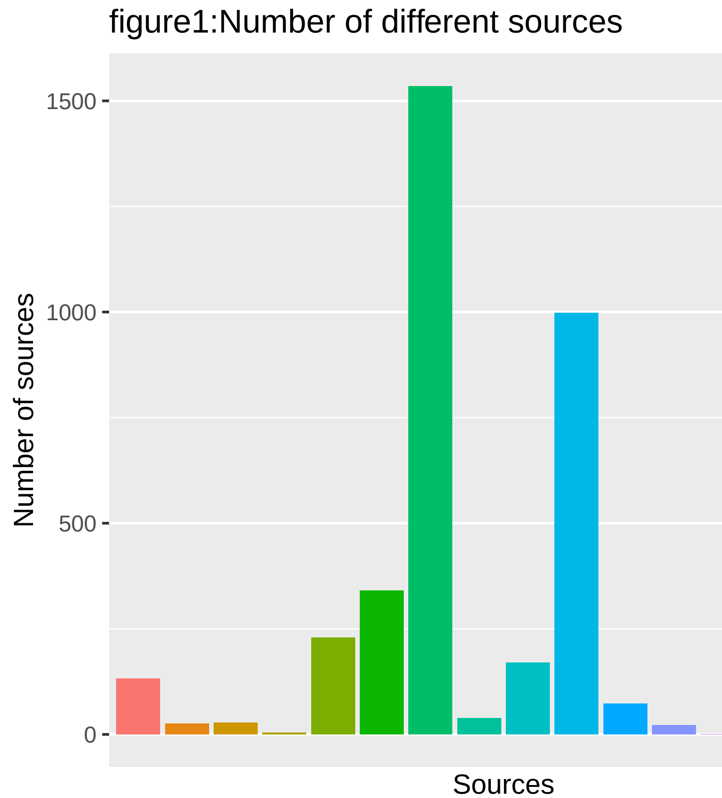
Part4 Results

Based on the modeling above, there results are: 1. More Popularity and More members of fan groups are more likely to increase the rating for animes, while the scoring people do not have strong relationship. The different type and source also make sense. 2. Multilevel modeling distribution works well on type/source, for the members value as the x-label, some specific type/source are more effective in rating, like OVA, Special/ Manga and game.The random effects are changing for different model. This indicates that number of members for group is explaining some variation. 3. There are some words that show high frequency in the genre part, which means high rating animes are often related to some of the words.

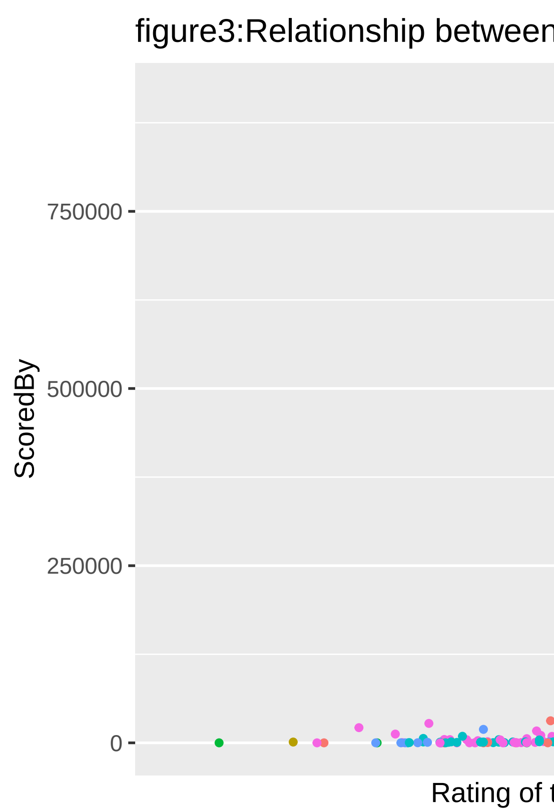
Part5 Discussion

Based on the results, especially for the prediction part, the members and popularity are the most 2 important numerical variables that affect the rating. More popularized and discussed anime are easier to get rating higher than average. However, the groups of type and source may be different in different situations. For example, TV type of animes have more members to discuss, but the effect to increase the rating is not that important. In the future, the type of animes will still be TV-mained, but if more people start to become members of other types, there rating will increase faster than TV. Also, it will be harder to get higher rating when evaluating the date of anime(no space to show in the pdf) and about the genre, animes about comedy, action and advanture will still be the main topics for anime.

Part6 Appendix and more things



1. EDA:The geom_bar for variable "Source" and this count

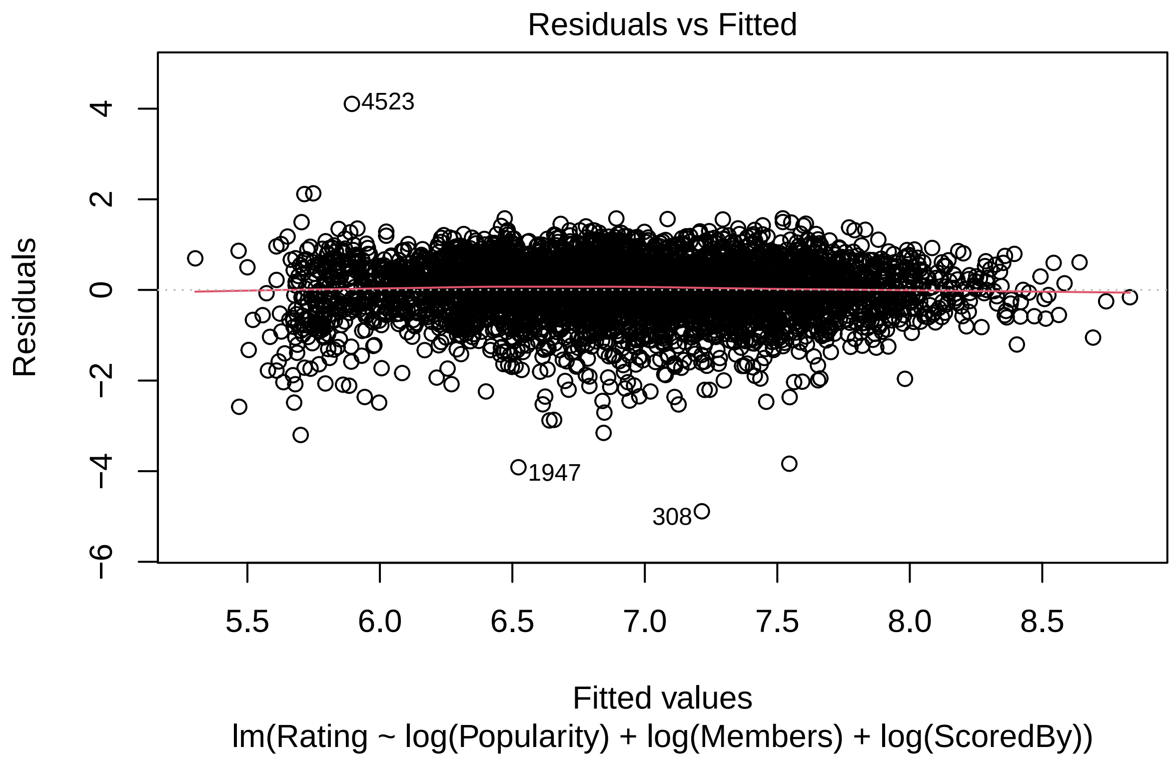


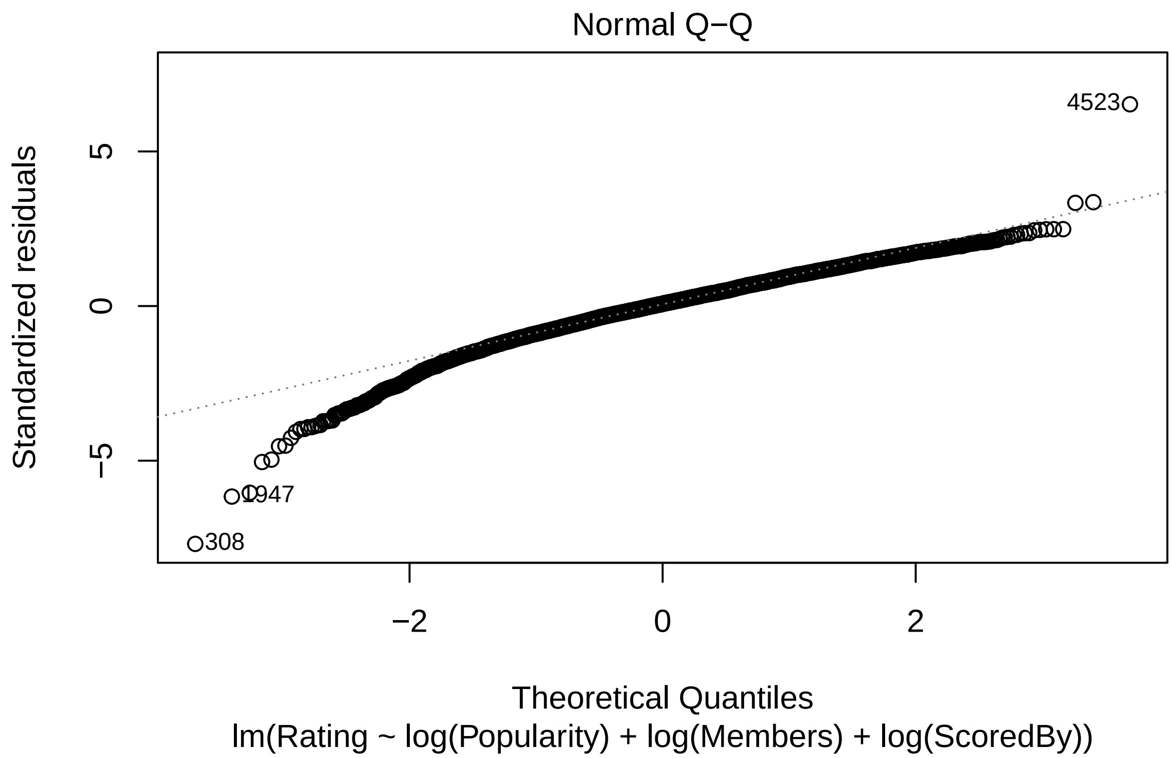
2. EDA:The geom_point for variable "Type" and "ScoredBy" with Rating

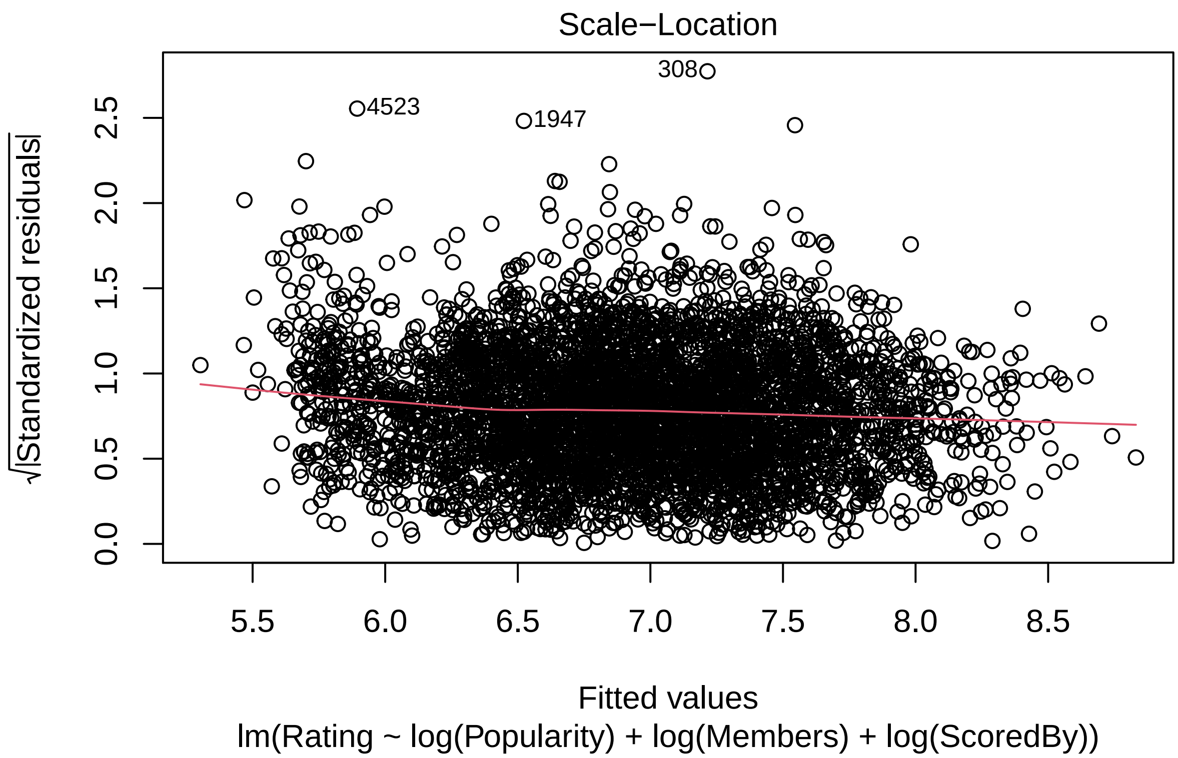
3. Linear Model

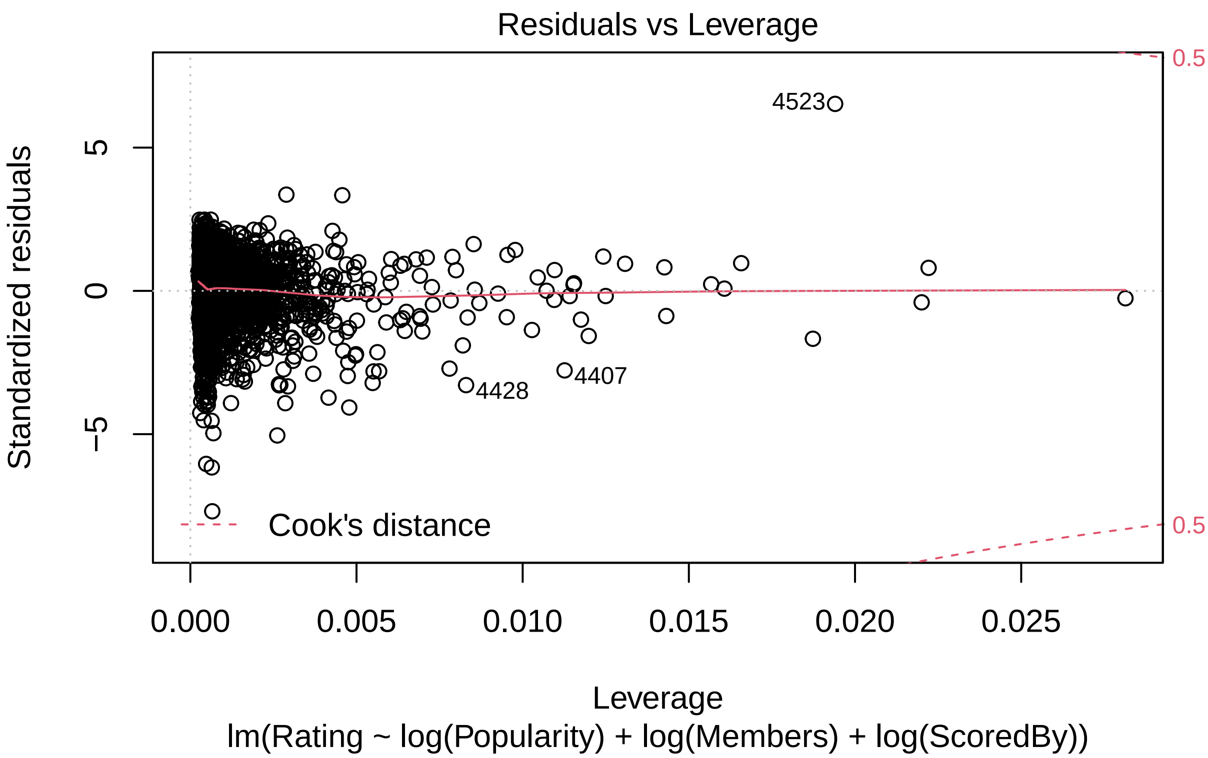
```
##
## Call:
## lm(formula = Rating ~ log(Popularity) + log(Members) + log(ScoredBy),
##     data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8852 -0.3553  0.0464  0.4265  4.1057
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.76323    0.27090  21.275  < 2e-16 ***
## log(Popularity) -0.10496    0.01950  -5.383  7.7e-08 ***
## log(Members)     0.19692    0.03027   6.505  8.6e-11 ***
## log(ScoredBy)    0.01980    0.02280   0.868    0.385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6353 on 4519 degrees of freedom
## Multiple R-squared:  0.4538, Adjusted R-squared:  0.4535
## F-statistic:  1252 on 3 and 4519 DF,  p-value: < 2.2e-16
```

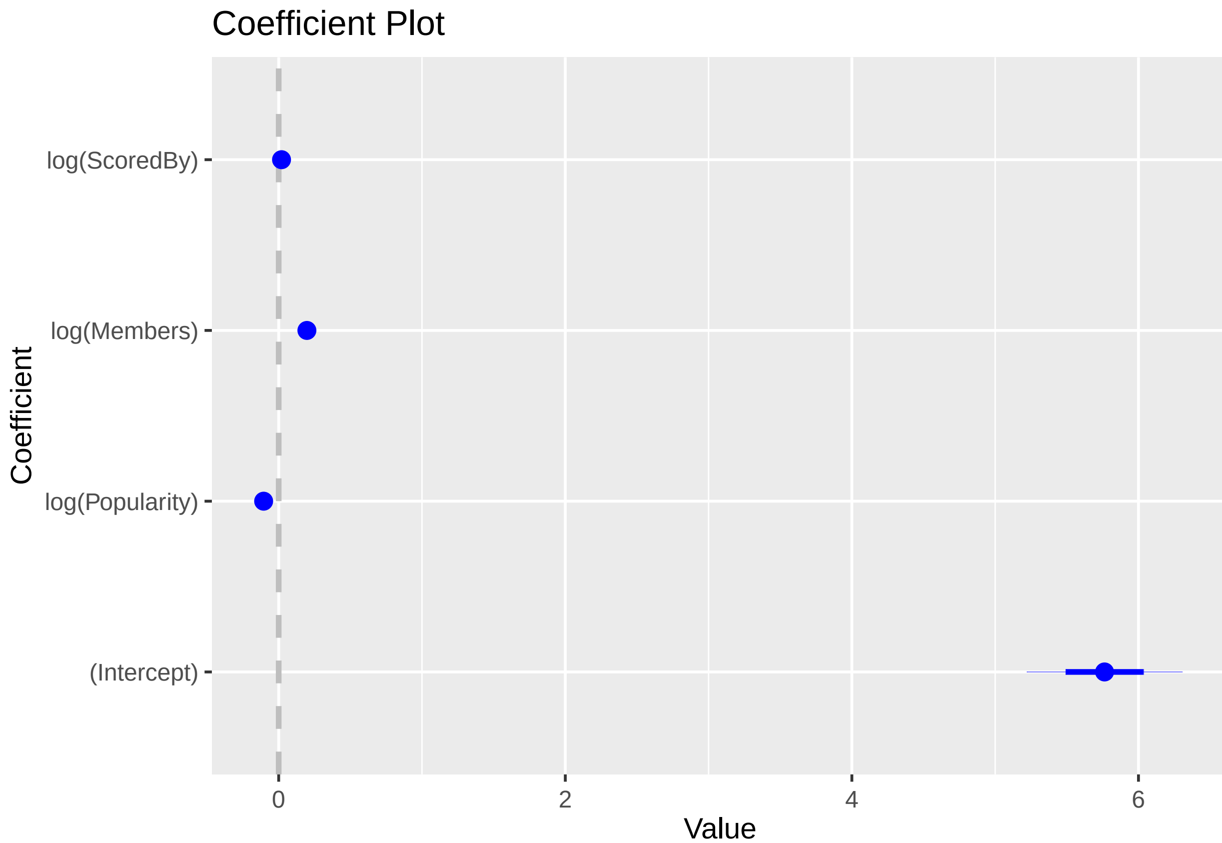
	2.5 %	97.5 %
(Intercept)	5.23	6.29
log(Popularity)	-0.14	-0.07
log(Members)	0.14	0.26
log(ScoredBy)	-0.02	0.06


Residuals vs Fitted
Residuals
Fitted values
lm(Rating ~ log(Popularity) + log(Members) + log(ScoredBy))
4523
1947
308
4
2
0
−2
−4
−6
5.5
6.0
6.5
7.0
7.5
8.0
8.5


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Rating ~ log(Popularity) + log(Members) + log(ScoredBy))
4523
1947
308
5
0
−5
−2
0
2


Scale−Location
4523
1947
308
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
5.5
6.0
6.5
7.0
7.5
8.0
8.5
Fitted values
lm(Rating ~ log(Popularity) + log(Members) + log(ScoredBy))

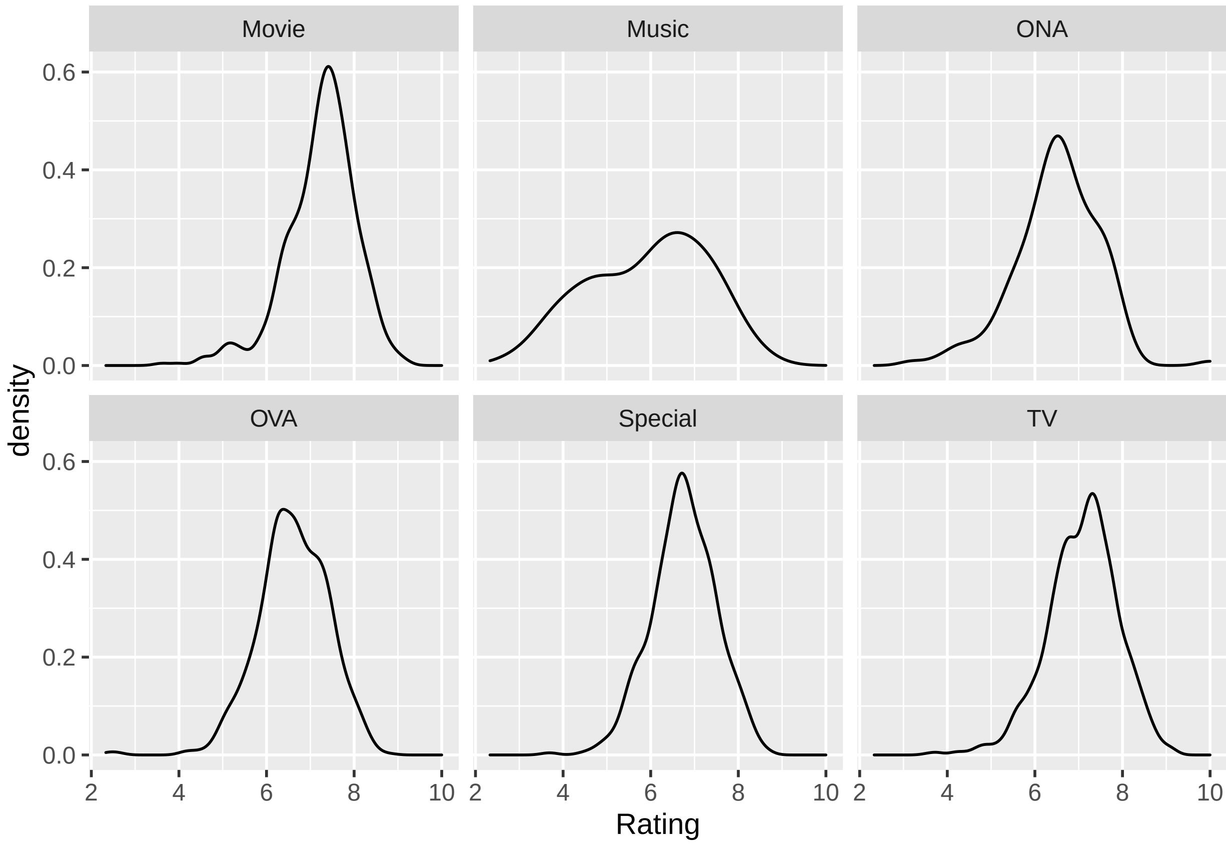

Residuals vs Leverage
Standardized residuals
5
0
−5
0.5
0.5
4523
4428
4407
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
Leverage
lm(Rating ~ log(Popularity) + log(Members) + log(ScoredBy))



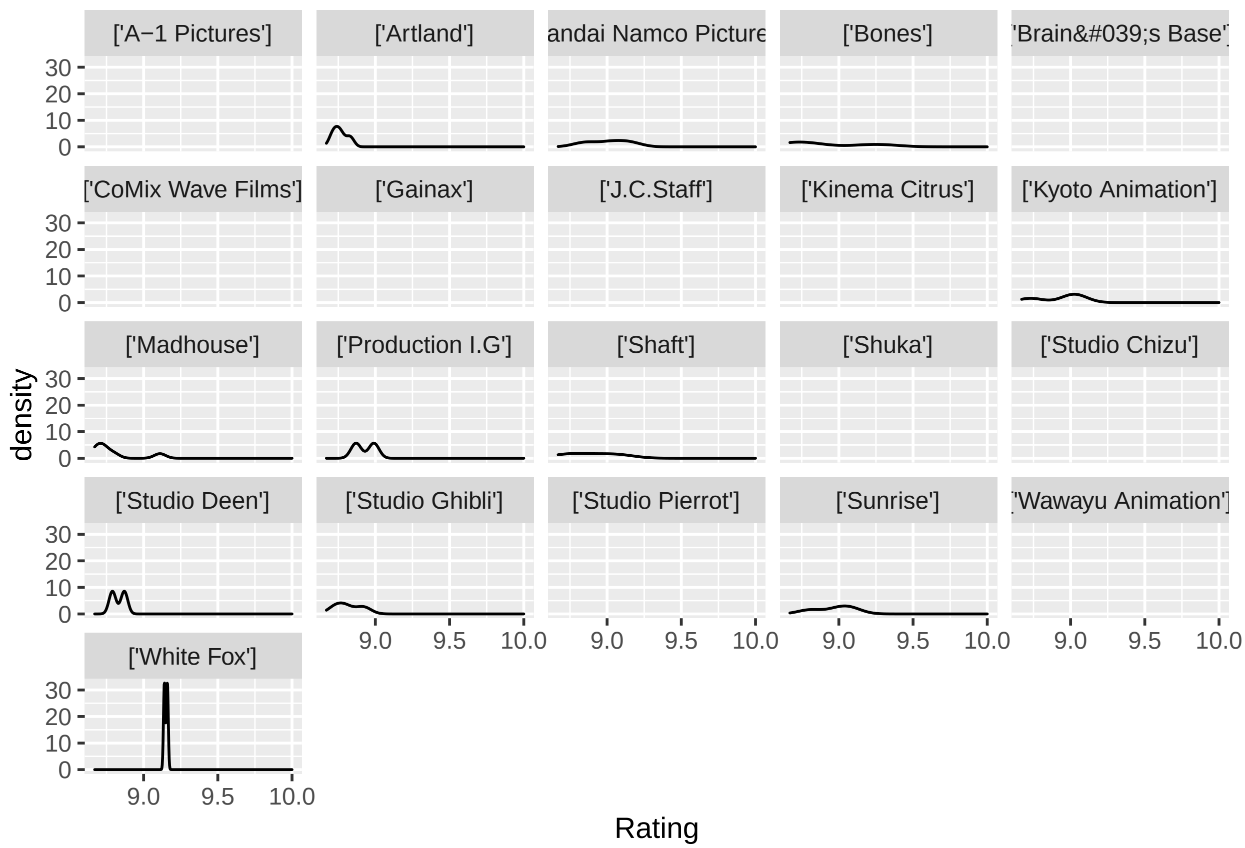
In simple linear model, the model complies with the assumptions of normality and constant variance, so there is no issue about violating the model assumptions. More popularity score(less popular) will decress the rating with 0.10502 , more community fans and scored people will increase the rating with 0.19808 and 0.01965.

4. Multilevel Modeling distribution about Type/ Studio/Producer

```
## # A tibble: 6 x 4
##   Type     mean    SD  miss
##   <chr>   <dbl> <dbl> <dbl>
## 1 Movie    7.24  7.24     0
## 2 Music    5.94  5.94     0
## 3 ONA      6.5   6.5      0
## 4 OVA      6.61  6.61     0
## 5 Special  6.75  6.75     0
## 6 TV       7.08  7.08     0
```



Movie
Music
ONA
OVA
Special
TV
density
0.0
0.2
0.4
0.6
2
4
6
8
10
Rating



5. Multilevel Modeling about Type

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + (1 | Type)
##    Data: data2
##
##      AIC      BIC   logLik deviance df.resid
##  11072.7  11092.0  -5533.4  11066.7     4520
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.2166 -0.5894  0.0571  0.6714  4.2634
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Type     (Intercept) 0.1698   0.4121
##  Residual             0.6721   0.8198
## Number of obs: 4523, groups:  Type, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   6.6947     0.1698   39.43
```

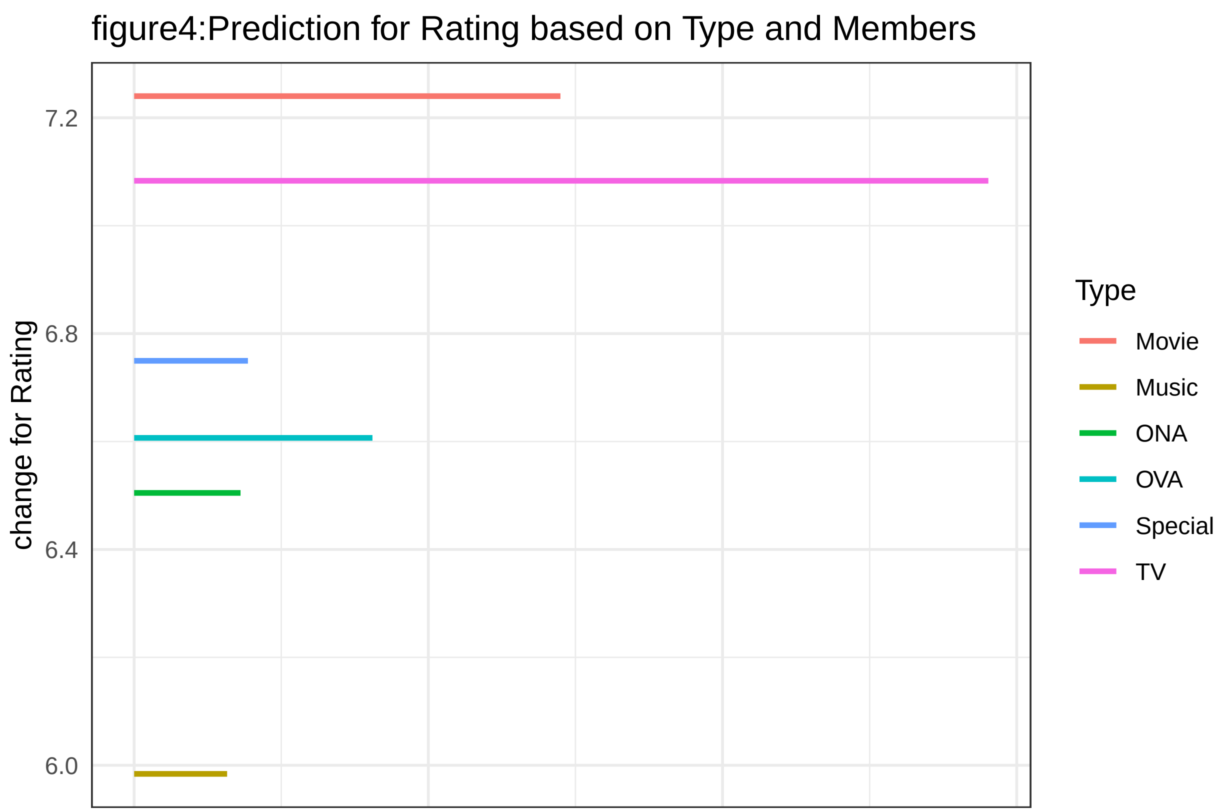
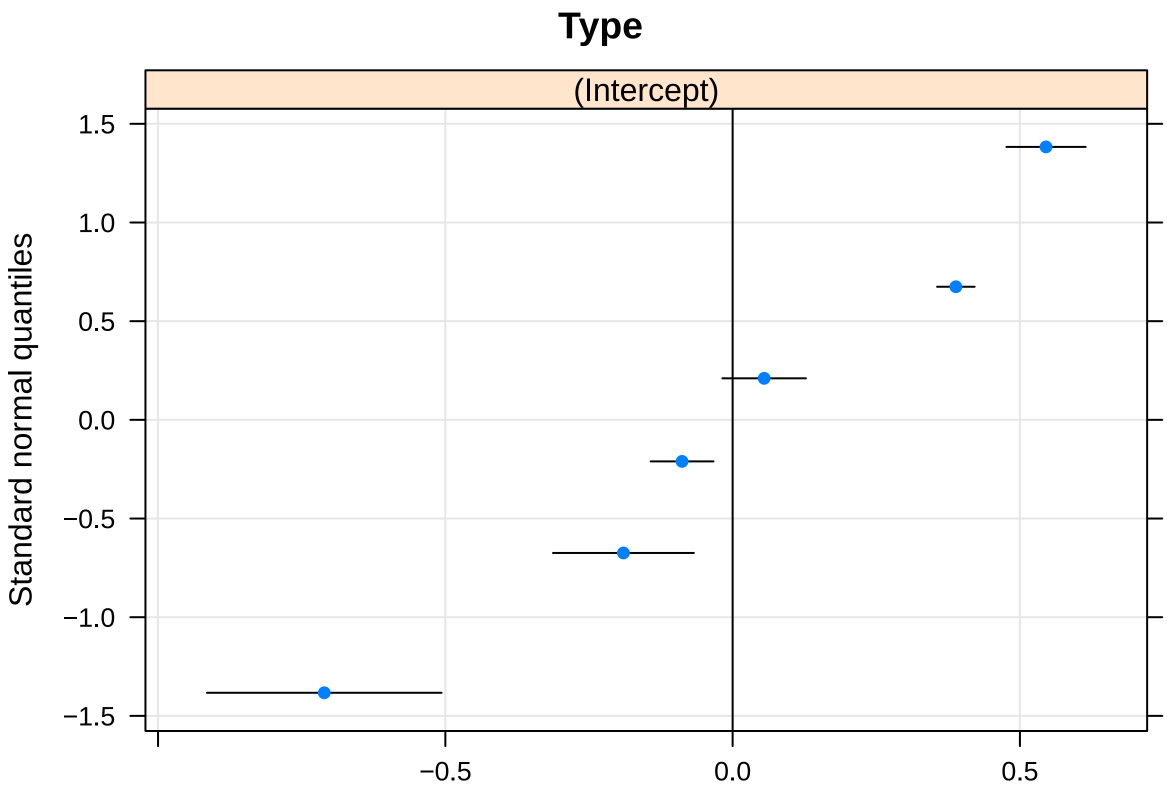


figure4:Prediction for Rating based on Type and Members
change for Rating
7.2
6.8
6.4
6.0
Type
Movie
Music
ONA
OVA
Special
TV

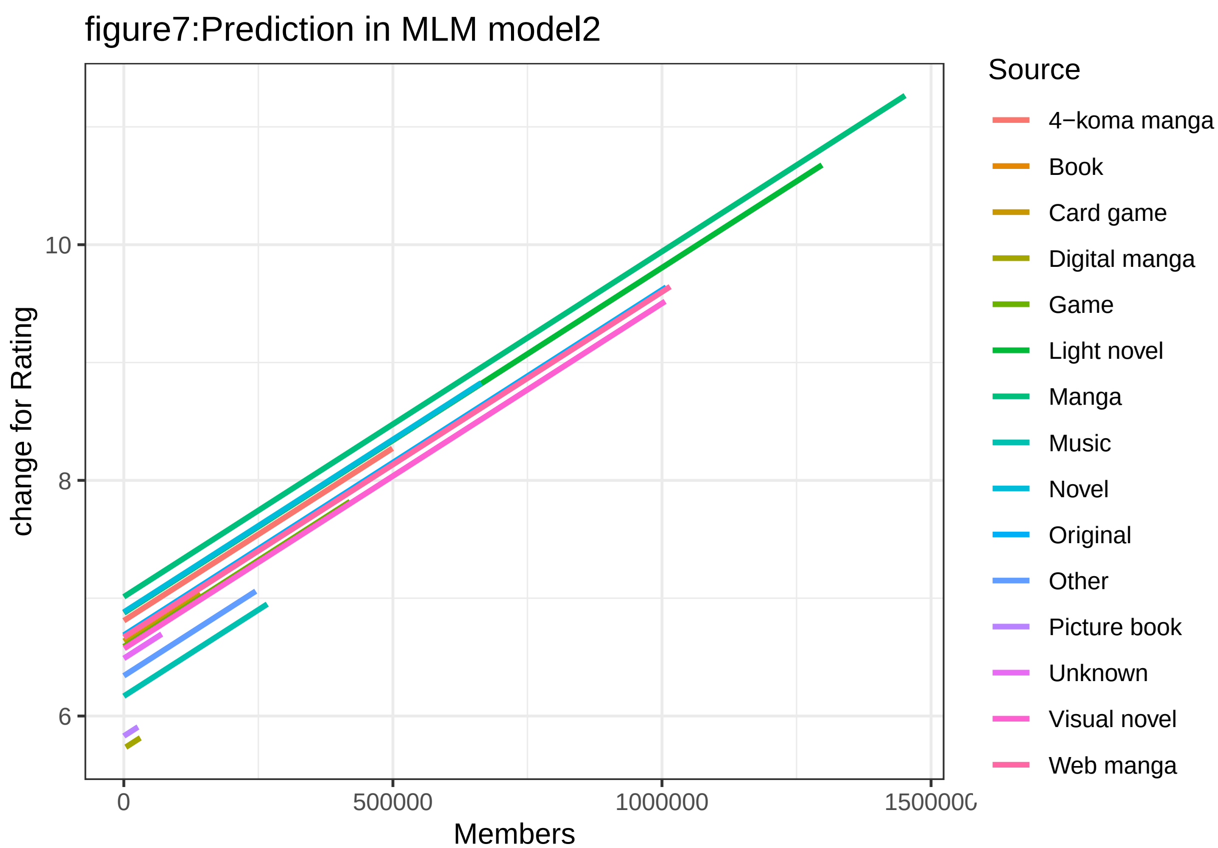
```
## $Type
```



6. the prediction in MLM models for Source and Member

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + Members + (1 | Source)
##    Data: data2
##
##      AIC      BIC   logLik deviance df.resid
##  10086.3  10112.0  -5039.1  10078.3     4519
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.9243 -0.5642  0.0865  0.6857  4.5205
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Source   (Intercept) 0.1449   0.3807
##  Residual             0.5372   0.7330
## Number of obs: 4523, groups:  Source, 16
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 6.546e+00  1.016e-01   64.42
## Members     2.931e-06  9.552e-08   30.68
##
## Correlation of Fixed Effects:
```

```
##          (Intr)
## Members -0.042
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

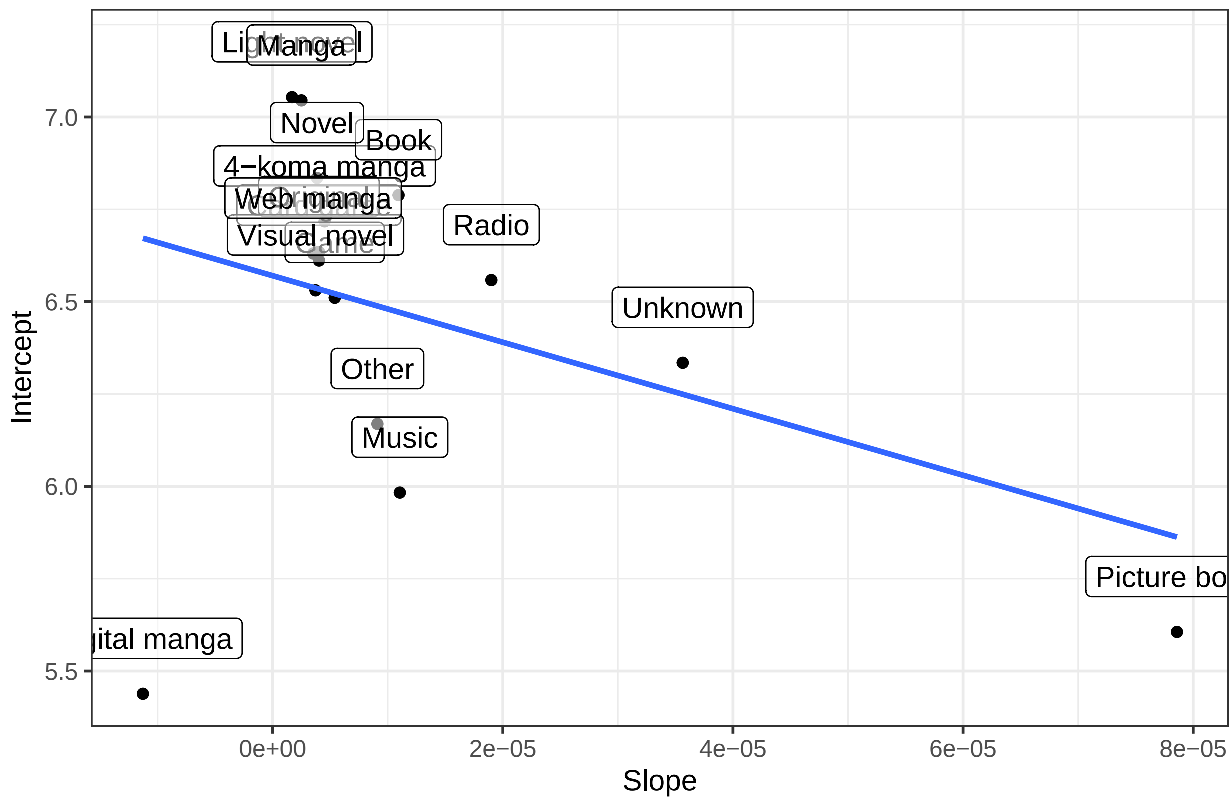


7. Random slope model for Source/Member.

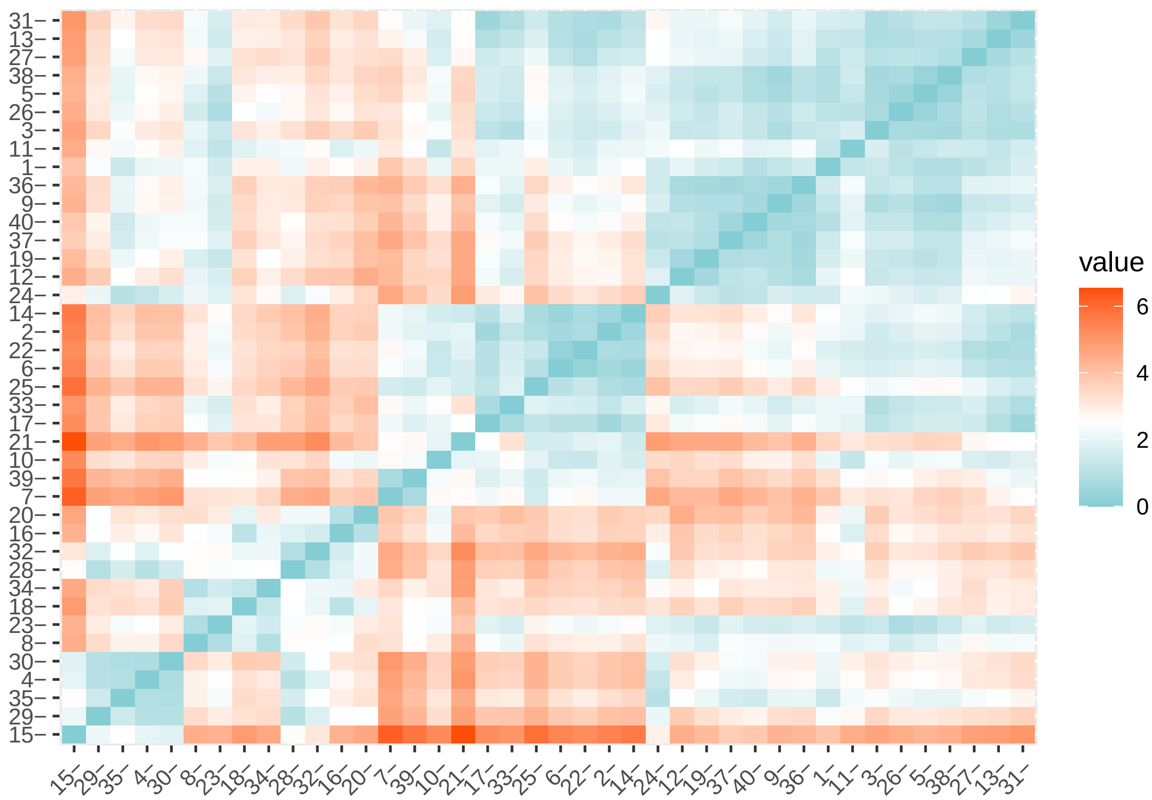
```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + Members + (1 + Members | Source)
##    Data: data2
##
##      AIC      BIC   logLik deviance df.resid
##  10024.7  10063.2  -5006.3  10012.7     4517
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.0447 -0.5670  0.1018  0.6754  4.6977
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.  Corr
##  Source   (Intercept) 5.115e-01 7.152e-01
##           Members     7.971e-09 8.928e-05 0.26
##  Residual             5.129e-01 7.162e-01
## Number of obs: 4523, groups:  Source, 16
##
```

```
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept) 6.465e+00  1.854e-01  34.867
## Members     1.164e-05  2.304e-05   0.505
##
## Correlation of Fixed Effects:
##         (Intr)
## Members 0.228
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## optimizer (nloptwrap) convergence code: 0 (OK)
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

figure10 :effect of members and rating



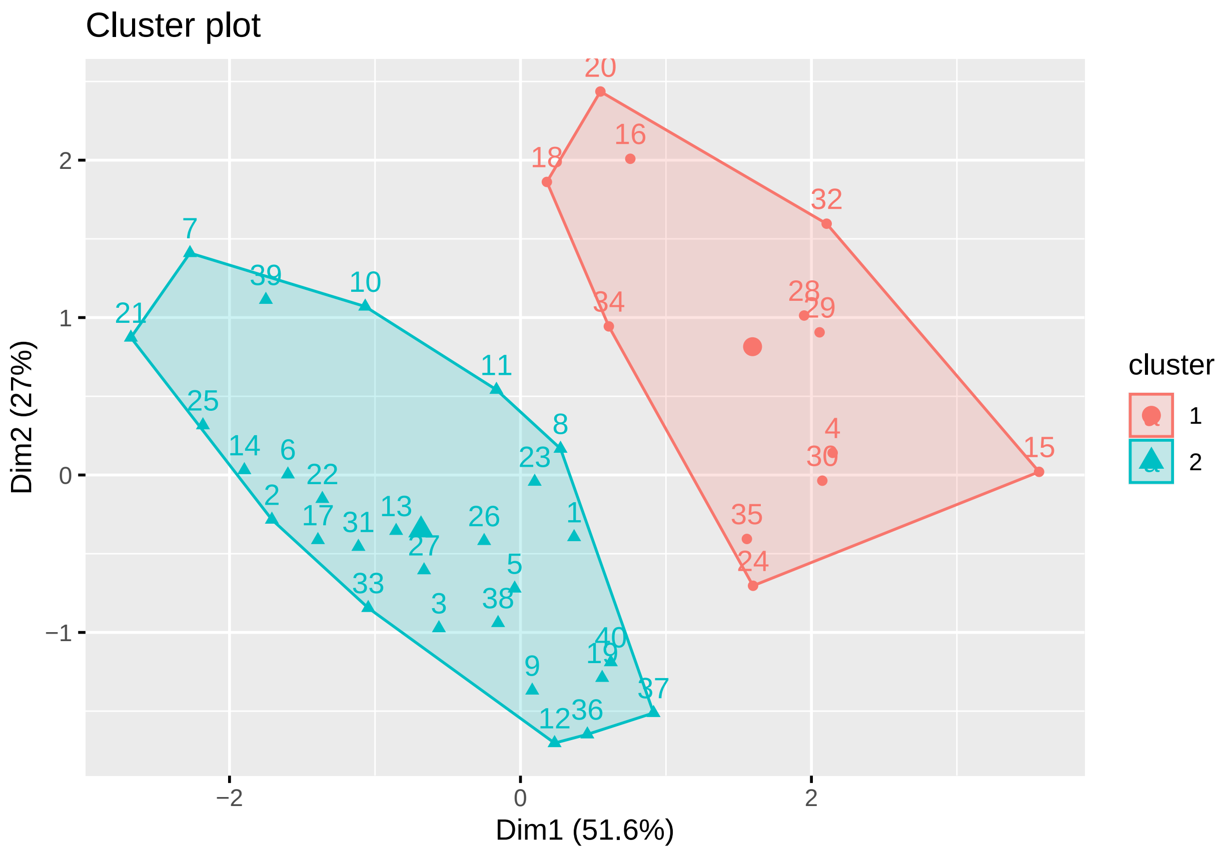
In the Source part, the random slope model is not useful in plotting the prediction. But the Coef plot is useful to analysis.



8. k-mean

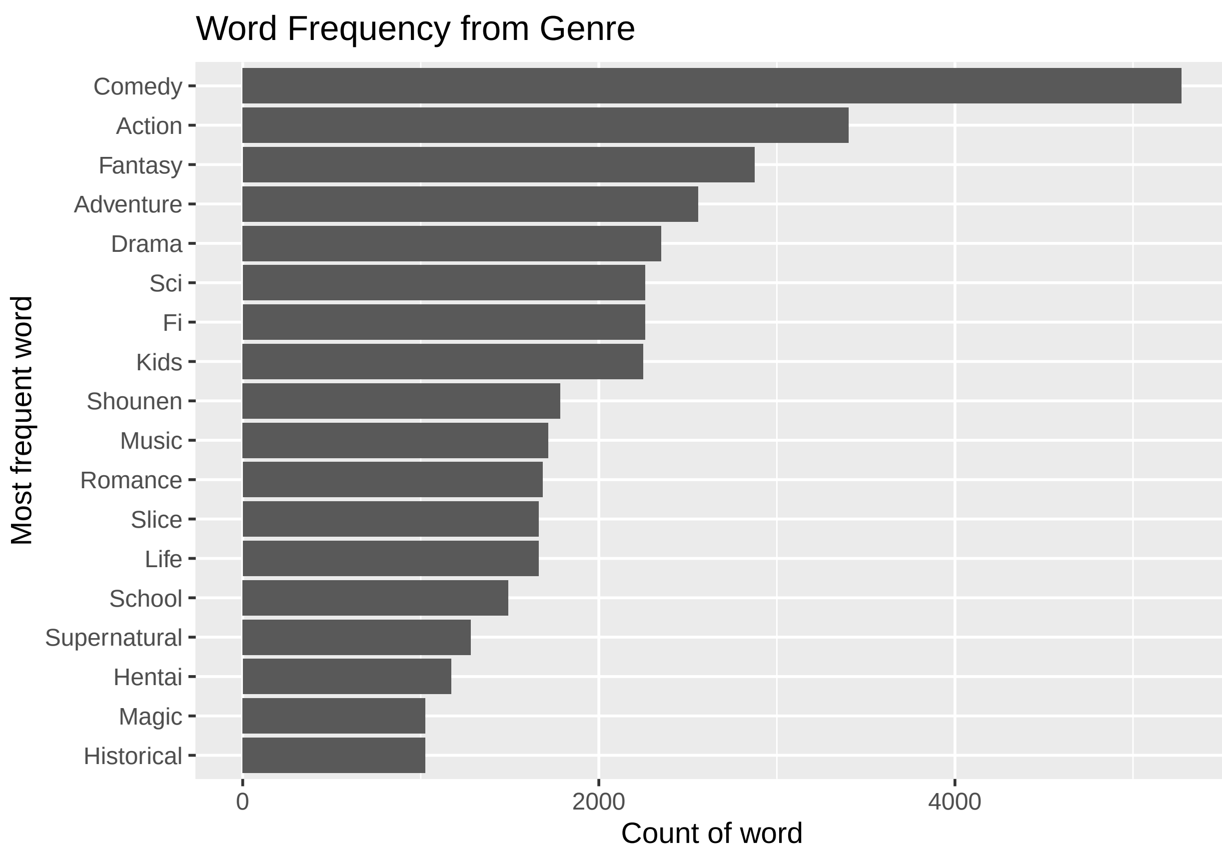
```
## List of 9
##  $ cluster     : int [1:40] 2 2 2 1 2 2 2 2 2 2 ...
##  $ centers     : num [1:2, 1:4] 1.0247 -0.4392 0.1919 -0.0822 1.1373 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:4] "Anime_id" "Rating" "Members" "Popularity"
##  $ totss       : num 156
##  $ withinss    : num [1:2] 36.4 64.4
##  $ tot.withinss: num 101
##  $ betweenss   : num 55.2
##  $ size        : int [1:2] 12 28
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"

## K-means clustering with 2 clusters of sizes 12, 28
##
## Cluster means:
##     Anime_id      Rating    Members Popularity
## 1  1.0247103  0.19185589  1.1372529 -0.9163143
## 2 -0.4391616 -0.08222395 -0.4873941  0.3927061
##
## Clustering vector:
##  [1] 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 1 2 1 2 1 2 2 2 1 2 2 2 1 1 1 2 1 2 1 1 2 2 2
## [39] 2 2
##
## Within cluster sum of squares by cluster:
```

```
## [1] 36.37676 64.42640
##  (between_SS / total_SS =  35.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```



9. Topic modeling Finally, I tried what I learn in 615 to analysis the non-numerical data.

```
## # A tibble: 46 x 2
##    word          n
##    <chr>     <int>
##  1 Comedy     5272
##  2 Action     3404
##  3 Fantasy    2874
##  4 Adventure  2558
##  5 Drama      2350
##  6 Fi         2259
##  7 Sci        2259
##  8 Kids       2249
##  9 Shounen    1784
## 10 Music      1717
## # ... with 36 more rows
```


Word Frequency from Genre
Comedy
Action
Fantasy
Adventure
Drama
Sci
Fi
Kids
Shounen
Music
Romance
Slice
Life
School
Supernatural
Hentai
Magic
Historical
Most frequent word
0
2000
4000
Count of word


adventure
samurai
dementia
action
music
space
harem
mecha
magic
school
life
martial
military
shounen
scifi
super
historical
power
yuri
comedy
demons
mystery
game
thriller
horror
psychological
fantasy
hentai
vampire
ecchi
kids
romance
shoujo
drama
parody
cars
sports
arts
seinen
slice
yaoi
supernatural
police
josei