Generalised Linear Model
Final Project

Bolin Wu

Department of Statistics, Uppsala University

December 19, 2019

1 Introduction

In the era of big data, people are more and more believing in the power of data. The movie industry is one of the industries that huge amount of data are sprung out. In this project, I would like to predict the rating of movies by generalised additive model and generalised mixed model and compare their prediction accuracy by k-fold mean squared error.

2 Data

2.1 Data Description

In the original data set, there are 45457 observations and 26 variables. The details of the 11 variables are listed below. The variables are divided into 3 types, categorical data, numeric data and date data because knowing the class of data is essential for further analysis. Also here the ranging of numeric data and the level of categorical data are mentioned in order to have a clue whether there's outlier in a specific numeric variable or whether it is possible to include a specific categorical variable into modelling process.
Categorical Data:

- belongs_to_collection: If the movie belongs to a collection (series movie) or not. A binary data, 1 = Yes, 0 = No.
- genres: The genre of the movie. A categorical variable containing 1295 levels.
- original_language: The original language of the movie. Containing 92 levels.
- production_companies: The names of production companies. Containing 22668 levels.
- production_countries: The names of production countries. Containing 2389 levels.
- status: The status of the movie in minutes. Containing 6 levels. E.g "Released","Rumored",etc.
- Title: The title of the movie. Containing 42267 levels.

- Actor: The names of the main cast. Containing 42639 levels.
- Director: The name of the director for this movie. Containing 19043 levels.

Numeric Data:

- budget: The budget of the movie in US dollars. A numeric variable ranging from 1 to 380,000,000.
- popularity: The total number of audiences who watched the movie in cinemas. A numeric data ranging from 0 to 547.4 883.
- revenue: The revenue of the movie in US dollars. A numeric data ranging from 0 to 2,788,000,000.
- runtime: The length of the movie in minutes. A numeric data ranging from 0 to 1256.
- vote_average: The average rating vote for the movie in **website 1**. Ranging from 0 to 10.
- vote_count: Total number of votes of a given movie from website 1. Ranging from 0 to 14075.
- vote_another_site1: Total number of 1 star vote of a given movie from **website 2**. Like wise, I have vote_another_site_2 to vote_another_site_10. All of them are ranging from 1.

Date Data:

- release_date: The data when the movie is released. The date is ranging from 1/1/1900 to 9/9/2016.

2.2 Data Cleaning

In the data cleaning process, the missing values and outliers are being mainly focused on. First, I will deal with the missing value first. The missing value proportion of each variables is listed in the table 1. Here we assume the missing mechanism to be missing at random instead of missing completely at random, that is, the missing value does not depend on the unobserved data, but might have relationship with the other observed data. There are two reasons why I make such assumption. First, for our data set, it is possible that the missing value are missing because of other variables. For example, some production companies may tend not to publish their budget, or some high-budget movies may tend to hide their revenue if they fail to meet the expectation. Secondly,I can see that from table 1, the missing proportion of some variables is so high that if I treat missing values as MCAR and eliminate the observations with missing values, only 3017 observations will be left, which means 93% of the information are lost.

Table 1: Proportion of Missing Value

Variable	Missing Proportion	Variable	Missing Proportion
budget	0.8044	belongs_to_collection	0
vote_another_site3	0.7903	vote_another_site2	0.7834
vote_another_site1	0.7808	vote_another_site10	0.7678
vote_another_site9	0.7586070352	vote_another_site4	0.7366
vote_another_site5	0.7152	vote_another_site8	0.7007
vote_another_site7	0.6913	vote_another_site6	0.6760
production_companies	0.2612	genres	0.2262
production_countries	0.1383	runtime	0.0057
status	0.0019	release_date	0.0018
actor	0.0007	director	0.0007
original_language	0.0002	popularity	0.0001
revenue	0.0001	title	0.0001
vote_average	0.0001	vote_count	0.0001

Further, when trying to use multiple imputation(MI) to deal with the missing value of the original data set, there are some problems that particularly exist in our case that prevent the MI process from going on. First, as I can see from data description, there are so many categories in the categorical data that R fails to handle them because of computational power limitation. MI with R can not include factors with lots of levels, saying higher than 50, otherwise the R session would either be aborted automatically or an error will show up. Second, if there are some linearly dependent columns exist in the imputation model, computationally singular error will show up. The solution to the problems above is to remove some variables from the imputation model. As a result, some variables need to be removed from the imputation model, that is, doing variable selection before MI.

Before introducing what variables to include in the imputation model, it is better to do variable selection for the prediction model. Since there more than 70 percent of the rating data of website 2 are missing whereas only 0.0001 missing proportion in website 1, and they share the similar explaining power, so that all the variables related to website 2 are excluded. Moreover, "release_date" are excluded because by observing the rating on IMDB, there seems no clear correlation between the release year and rating. For example,there are high-rating movies produced throughout every year period like in the 1970s,The Godfather (1972), in the 1990s, The Shawshank Redemption (1994), in the 2010s, Jocker(2019),etc. Similarly, there are low-rating movies in every year period as well. "original_language" and "production_countries" are excluded because now I have subtitles for movies in various languages so that the language and production location may not hinder us to understand and rate movies.

After several tests, the imputation model includes"runtime","budget","popularity" and "vote_count". Figure 1 shows that the imputation is good because there is no clear trend in the mixing markov chains.

From the data description, it is obvious that there are unreasonable data like revenue=0, runtime=0. So that I set the filter to be runtime between 40 to 210 (American Film Institute definition), popularity>0 and revenue>0 and omit other observations with missing value in other variables. One thing worth mentioning is that there are 21341 zero values in "revenue". I had been struggling with choosing between excluding the variable to obtain more observations or keeping this variable to enable the prediction model to have more explanation power. Finally I chose the latter because I believe "revenue" is an indispensable factor and get a clean data set with 12 variables and 6230 observations.

3 Modelling

In the modelling part, I will choose generalised linear mixed model(GLMM) and generalised mixed model(GAM) to model our data. Since the predicted variable, rating_average, is a continuous data, it could only be assumed to be Gaussian, gamma or inverse Gaussian distributed. Further analysis finds

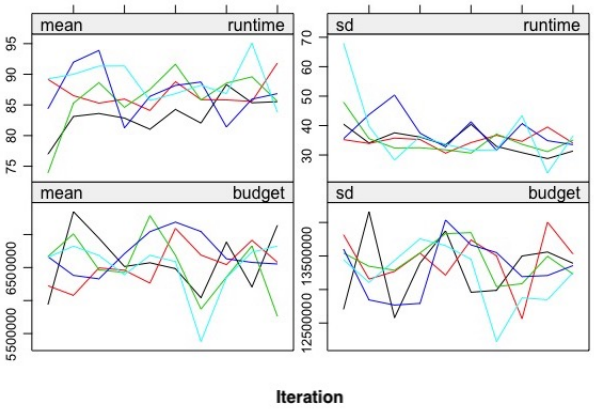


Figure 1: Mixing Markov Chains

that when assuming dependent variable to be gamma or inverse Gaussian, convergence failure occurs. So that in the following analysis, dependent variable is assumed to be Gaussian distributed only.
Alternative possible models are hierarchical generalised linear model(HGLM), generalised linear model(GLM), etc. The former is not chosen because when I run HGLM in R with large categorical data, it is aborted automatically. The later is not chosen to avoid the report being too long.

3.1 Model Description

(1) GLMM
Because of limited computational power, the mixed model with only random intercept is considered. In practice, I can add random coefficients for β as well. GLMM with a random intercept can be expressed as follows:

$$g(E[y_{ij}|v_i^{(1)}]) = \alpha + v_i^{(1)} + \beta_{ij}x_{ij} \tag{1}$$

Here g() is the link function, α is the fixed intercept, β_{ij} is the fixed coefficients. $v_i^{(1)}$ is the random intercept, i is the number of clusters and j is the number of observations in each cluster.
There are some important assumptions for GLMM. First is that dependent variable belongs to a exponential family. Second is that the random effects $v_i^{(1)}$ conform to N(0, σ^2). Third is that within each cluster, observations are

correlated. It is because $cov(\alpha + v_{i1}^{(1)} + \beta_{i1}x_{i1} + e_{i1}, \alpha + v_{i2}^{(1)} + \beta_{i2}x_{i2} + e_{i2}) = cov(v_i, v_i) = \sigma^2$. Similar to GLM, the normality analysis of residual is needed with the model

In R, package "lme4"[1] can be used to perform GLMM.

(2) GAM

The model of GAM can be espressed as follows:

$$g(\mu_i) = f_1(x_{1i}) + f_2(x_{2i}) + ... + f_p(x_{pi}) \tag{2}$$

Where g() is the link function and f() is unknown function, estimated by a number of basis functions.

$$f(x) = \sum_i \beta_i b_i(x) + e \tag{3}$$

Usually we have 4 commonly used basis function $b(x)$, that are "cubic regression spline basis", "cubic B-spline basis", "fourier basis" and "this plate spline basis". The advantage of "cubic regression spline basis" is that it is easy to simulate, but it usually has poor estimation performance. The "fourier basis" is usually used to model the seasonal data. In the following analysis I would use "this plate spline basis" and "cubic B-spline basis". Basic assumptions include the dependent variable belong to the exponential family, and the residual should be normally distributed. GAM can be done in R by package "mgcv"[2].

3.2 Model Evaluation

To evaluate the each model, the analysis of residual is needed which could be done by looking at the residual plots. Moreover, for GLMM, there should not be strong correlation between the variants of fixed effects. For the GAM, the number of additive basis, k, should be chosen to the point where the plots of the model could be stabilized. To compare the prediction performance of different models, k-fold cross validation is used instead of leave one out cross-validation (LOOCV) because LOOCV splits the data in a total of n times and fits model for n times which is challenging for the case when n is large.

(1) GLMM

There are three categorical variables available to be the cluster in model which are "genres", "production_companies" and "production_countries". Here

"genres" is chosen as cluster for the reason that further analysis shows choosing "genres" as cluster enables model to have the smallest MSE. The results are shown in table 2.
The variable "run_time" is excluded because it produces strong correlation(-0.958) of other fixed effects. With such strong correlation, the model will be unreliable. Also, the independent variables are on very different scale, like revenue can be more than 100000 whereas run time are smaller than 210, therefore I first form a model by scaling the variants.

Table 2: MSE of Models with Different Clusters

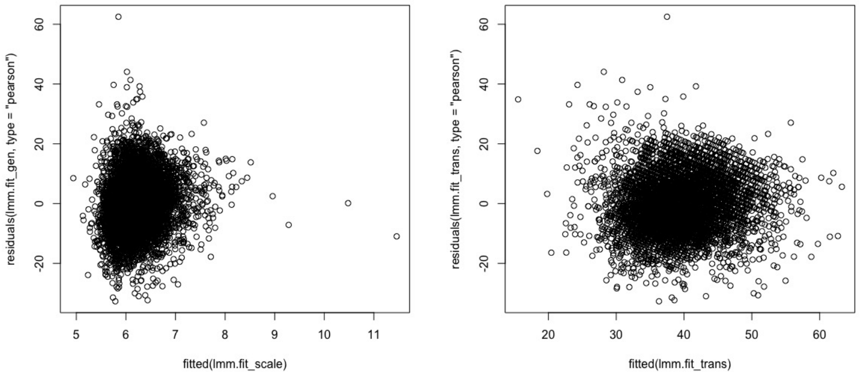
Cluster	MSE
genres	0.676
production_countries	0.733
production_companies	0.704

However, the left residual plot in figure 2 shows that there are heteroscedasticity problem in the model. The usual remedies of heteroscedasticity for OLS model like using generalised lease square method may not applicable in GLMM so that I try to solve this by transforming the variables. Then I find transforming the dependent variable to the power of two, and taking logarithm of each independent variables is a good remedy. The residual plot of further transformed model is shown as the right picture in figure 2. The figure 3 shows that the normality assumption is satisfied. The model can be expressed as equation 4, and the estimated coefficients of fixed effects and the variance of v_i, residual u_i are listed in table 3:

$$\begin{aligned} E[vote\_average_{ij}|v_i^{(1)}]^2 = \alpha + v_i^{(1)} &+ \beta_{i1}(belongs\_to\_collection = 0) \\ &+ \beta_{i2}log(budget) + \beta_{i3}log(popularity) \\ &+ \beta_{i4}log(revenue) \end{aligned} \tag{4}$$

Table 3: Estimation Output of GLMM

Variable	Coefficients	Variable	Variance
intercept (fixed)	37.739	v_i	18.69
belongs_to_collection=0	0.91288	residual	94.73
log(budget)	-0.712		
log(popularity)	3.331		
log(revenue)	00.385		



(a) The Model of Scaled Data (b) The Model of further Transformed Data

Figure 2: Residual Plot versus Fitted Value

(2) GAM

To better compare with GLMM, "run_time" is excluded as well in this model. Similar to the previous GLMM, there is also heteroscedasticity problem with fitted GAM before data transformation, so that I solve this problem by transforming dependent variable to the power of three, and standardizing each independent variables. When setting the number of additive basis k to be 10, the plots are stable. The plots of estimated variables are in figure 4. In the plots, the shade are not supposed to center around the horizontal line of y = 0, otherwise it is not significant in the model. The residual plots in figure 5 show that there is no clear trend in residual versus fitted value and qq plot seems fine. The k-fold MSE is 0.725.

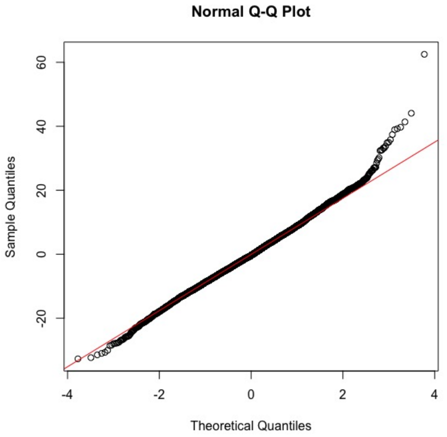


Figure 3: Q-Q Plot of Estimated Model

4 Conclusion

In this project, I used generalised linear mixed model and generalised linear additive model to model make prediction of the rating of movies. If assuming the rating to be Gaussian distributed and using the four dependent variables mentioned above, MSE of GLMM and GAM are 0.676 and 0.725 perspectively. As a result, GLMM is more accurate than GAM is our case.

There are several improvements could be implemented for future analysis. First, if the computational power allows, I would try to use HGLM to fix the heteroscedasticity problem. Also, I would include other categorical variables in the multiple imputation model and mixed model to enlarge the clean data set and enrich the prediction model. Second, I would use try other exponential assumptions for rating, like inverse Gaussian and gamma distribution. Third, generalised liner models could be involved to the comparison.

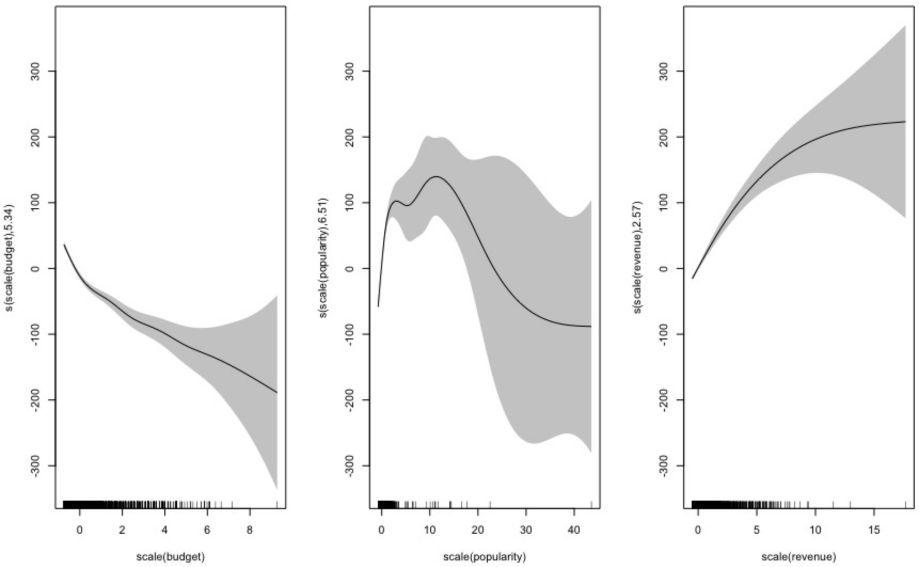


Figure 4: Variables of GAM

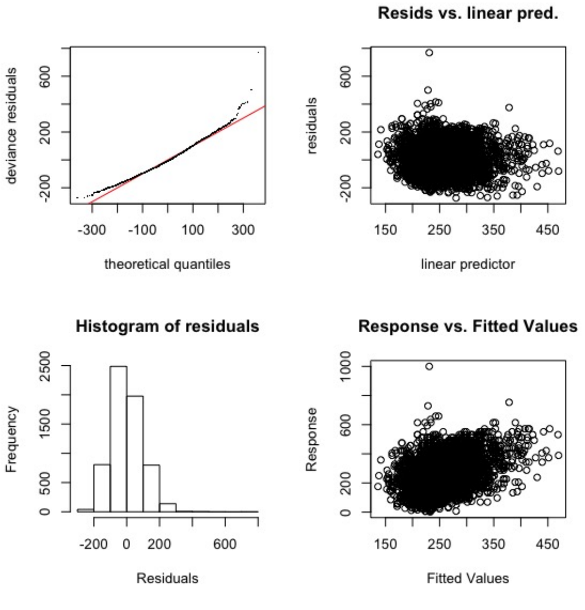


Figure 5: Residual Plots of GAM

References

[1] Mächler M. Bolker B. Bates, D. and S. Walker. *Fitting linear mixed-effects models using lme4.* Journal of Statistical Software, 67(1):1–48., 2015.

[2] S. Wood. *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC, 2 edition., 2017.