Xi'an Jiaotong-Liverpool University

西交利物浦大学

Knowledge Discovery with Linear Mixed-Effects Model

使用线性混合模型的信息挖掘

Name: Zhiyi Chen

ID: 1508116

Supervisor: Shengxin Zhu

Financial Mathematics

Department of Mathematics Science

April 17 2019

Abstract

Knowledge discovery in databases possesses high popularity during the process of data analyzing in modern days because once certain patterns, regularities or potential knowledge which most people have not realized are digged out from tremendous data, to a strong degree it can bring about huge benefit for specific parties. In daily life, people are enjoying the merits from recommender systems (RS) which process knowledge discovery from users' data. For example, Internet movie database (IMDB) provides marks of movies which can be regarded as indicator to make decision about which movies are worth being watched. Accordingly, the methods for carrying out knowledge discovery are developing rapidly and efficiently. The linear mixed-effects model (LMM), the one avoiding the limitations of traditional linear model, has strong advantages to analyze real data. In this paper, we will briefly discuss several approaches in recommender system, the details of LMM as well as how to apply LMM for analyzing movie rating data.

Keyword:Knowledge discovery in databases(KDD), Linear-mixed effects model(LMM), recommender system (RS), R software.

在如今的数据处理过程中，数据库中的知识发现拥有极高的人气，因为一旦某些大部分人不知道的模式，规律或者潜在的知识从庞大的数据中被挖掘出来，在很大程度上，这能够为某些特定的群体带来极大的收益。在日常生活中，人们享受着推荐系统（RS）带来的好处，这些推荐系统能够使用用户的数据进行知识发掘。举一个例子，网络电影数据库（IMDB）提供了电影评分，这个可以被看做是一个指标去做决定关于哪些电影值得观看。当然，执行知识发掘的方法正在快速并高效地发展。线性混合模型（LMM）可以避免传统线性模型的局限性并且有着更强大的优势去分析实际数据。在这篇论文中，我们会简单地讨论几种推荐系统的方法，线性混合模型的具体细节以及如何去应用线性混合模型去分析电影数据。

Contents

1 Introduction

1.1 Background

With digital concept and analysis evolution dramatically, it is no exaggeration to define that human itself is the combination of data including the physical condition, such as the figures on the checkup report and even every behavior derived from individuals. For example, each grade in transcripts, the locations of places people often visit and their evaluations towards other people or goods are all data. With population increases significantly, the volume of data is bound to jump up exponentially[16]. If without deep analysis on data, these tremendous data sets only exist in its name which indirectly doubts the meaning of human's existence. Therefore, what we are interested in and pay much attention to is how to unearth useful information or knowledge from digital data that grows rapidly and the process is called knowledge discovery[16][9].

Knowledge discovery aims to discover high-level knowledge from raw or low data. Based on the fact that highly increasing data also contains the more fields with the order 10^2 or 10^3 involved in digital family , the final purpose obviously reveals that knowledge discovery is able to contribute to high pay-off in each domain, especially scientific and business areas[16][9]. The famous Kepler's law is derived from Kepler's conclusion by looking at the location of Mars from astronomer Tycho[16]. This achievement not only won him reputation but also lay important contribution for further study in astronomy, physics and mathematics. Moreover, In business application, inventory management, product planning, manufacturing, and recommending products to customers all demand to extract useful information from business data, from which firms are able to win profits. It mainly reflects in two ways that firms or companies facilitate efficiency by discovering the potential for the purpose of saving money and they earn more money through finding ways for more sale amounts to customers[1]. For example, firms can forecast customers' behaviors and obtain their preference by analyzing customer databases, which would be preserved as private information and help companies with more sales. Of course, knowledge discovery has different outcome depending on the methods used. Relying on huge number of trials in analyzing data, there are some excellent methods to analyze certain conditions. SKICAT , a system used by astronomers, can be applied to perform image analysis, classification and cataloging of sky objects[16]. The CASSIOPE troubleshooting system, becomes a significant tool applied by three major European airlines in order to make diagnosis of problems in the Boeing 737. However, knowledge discovery is a continuous field. Since the databases become much more complicated and huge, the methods for extracting pattern and knowledge from data must be outdated quickly. Discovering new methods for knowledge discovery will always be our interest and direction.

Recommender system, which is a popular and updated method for knowledge discovery are receiving broad success in E-Commerce nowadays[1]. Taobao, a Chinese famous shopping application, has ability to learn from customers' behavior and shopping history record, based on which it can construct model and corresponding algorithm. The calculation from the model will make recommendation of products to the customers. From consumer's perspective, facing the overwhelming items in the websites, users can readily find and purchase items of interest by recommender system in Taobao[14]. Obviously, one of the key challenges of recommender system should be how to producing high quality recommendations[1]. Thinking in diverse situations, Algorithm optimization and repeat experiments will be unavoidable steps. However, in this paper, we will not go for that far.

1.2 Motivation

Movie becomes one of the important entertainments in people's daily life and increasingly individuals are posting evaluations about movies online[10]. In this circumstance, the database of movie rating is created and its volume is expanding rapidly. Relying on the recommendation system, it is no doubt that efforts on creating model and sophisticated algorithm is worthwhile. Netflix emphasized on the importance of better recommender system indicated by Netflix Price, which is a contest willing to provide 1 million dollars for the team giving the best method on recommender system improvement[10]. More simply, we should think about what the movie rating databases indicate and can we obtain useful information from simple knowledge discovery. Moon, S.(2010) did analysis about movie database exploring the relationship between movie genre and movie rating and it gives a conclusion that sequel movies can achieve lower rating than the original ones, which is resulted from the reason that viewers' satisfaction is declined[19][10]. It delivers the information that original and innovative ones can be more attractive so that it inspires movie firms to produce more original movies. However, this kind of conclusion is too general to recommend accurately since it does not consider the fact that every individual possesses diverse traits and viewers can be divided by numbers of different factors. Little research exploring users' influence on movie rating motivates us to consider database about movie rating and users' data including gender, age, occupation, etc. Given these data, we are willing to find useful pattern between users' information and their movie rating. The movie websites or firms can apply this kind of discoveries to make recommendations for certain users. It not only can satisfy people's demands, but also brings about good reputation and profits for companies. In order to do this knowledge discovery, we choose linear mixed-effects model as our method to do analyze.

2 Literature review

Deigned for solving the problem related to data overload, recommender system is becoming in great demand and a hot topic[21]. It experienced from simple one to quite complicated algorithm. Physically receiving information from surrounding environment is the most simple way to obtain recommendation. For instance, individuals can hear others talking about the quality of certain goods[14]. Their subjective feeling and judgment invisibly convey recommendation to people. Obviously it is irrational since just two data from talkers are roughly processed by listener. Although this kind of RS seems unreasonable, it leaves great idea for future RS improvement. Collaborative filtering (CF) inherits this idea, expands the range of data put into analyze and finally provides generalized recommendations via aggregating the assessments of the community at large[14]. However, its effectiveness is bound to expire once the scale of consumers and products reach at large. Many other recommender techniques arise including user-to-user correlation and nearest neighbor algorithm. The former one is mainly used to measure the similarity of two users and according to their correlation value, RS provides recommender suggestions. The latter one works as calculating the distance between users relying on their preference history[1]. These are both personalized methods and their influence on recommendation is restricted by different situations and data volume. Moreover, some complicated algrithms are provided to deal with data more accurately. Generalized linear mixed model is one of these techniques.

Generalized linear mixed (GLMix) model has widespread application in solving problems about many different type of response because we can choose different link function to fit our responses[20]. Here, movie recommendation is discussed so that we assume the response is rating 1 to 5 which demands the link function is linear regression. Using GLMix for recommender system, responses(rating), users' features(age, gender, occupation, etc), movie features(ID, genre, etc) should be considered. We denote m as users, j as movies, t as the rating website and time. Then y_{mjt} denotes the numeric responses of the rating given by user m for movie j at context t, $\mathbf{q}_m$ denotes the feature factor of user m, $\mathbf{s}_j$ denotes the feature factor of movie. The generalized linear mixed model can be formulated:

$$g(E[y_{mjt}]) = \mathbf{x}'_{mjt}\mathbf{b} + \mathbf{s}'_j\alpha_m + \mathbf{q}'_m\beta_j \tag{1}$$

where $g(E[y_{mjt}]) = E[y_{mjt}]$. $\mathbf{b}$ is the fixed effects coefficient, $\mathbf{x}_{mjt}$ is the design matrix. α_m and β_j both are random effects which shows the preference of user m to each feature of movie and the attractive of movie j to each feature of user respectively. If user m has no or little rating to many movies, then α_m tends to zero. When user m has many rating records, we can estimate parameter α_m accurately. As for β_j, if movie j did not receive many rating from users, it becomes zero. On the contrary, when movie j obtain numbers of responses from users, this parameter can be estimated accurately. The benefit of GLMix is that it considers the historical evaluation effect to future rating. If historical records are abundant, the user's preference can be reflected more clearly. Otherwise, the rating is only influenced by fixed effects.

3 Methodology

3.1 Background

The traditional linear model is the model we have learned most to express the distribution of data and to analyze influence the fixed effects have on the response which has the expression showed below

$$Y = X\beta + \varepsilon \tag{2}$$

where Y is the $n \times 1$ vector of responses, β is the $p \times 1$ vector of fixed effects, X is the $n \times p$ matrix of fixed effects, ε is the $n \times 1$ vector of random errors. There are three important assumptions on traditional linear model that: 1) Normality, which means responses Y follow normal distribution from population. 2) Independence, which means responses are independent so that their correlation coefficient is zero. 3) Homogeneity of variance, which means every response has the same variance. However, when we analyze the actual data in the reality, it is not hard to find the high limitation of traditional linear model because the true condition of data sets mostly violates the assumptions of traditional linear model.

Firstly, the assumption of normal distribution only can be applied to the responses in the same level of factor whereas the actual data sets normally have complex hierarchy and show several different levels of a factor. For instance, a certain data sets contains one factor, gender, with two levels, male and female within which the observations sharing the same gender can follow the same normal distribution. It demonstrates that the normal distributions from different levels are not consistent. In this circumstance, it is not reasonable to assume Y have the coincident normal distribution from population. Secondly, Y are impossible to have complete independence among the actual data sets. Observations are collected from different background and levels. It indicates that the observations with similar or the same background or level are able to have non-zero or even high correlation. Therefore, we cannot guarantee independence of all observations. Thirdly, in the actual data sets, the responses measured have variance containing two types, one is random errors during analyzing process such as operation error and equipment mistake and the other one is a number of unstable and unknown random effects from different levels. For example, the data sets from survey of happiness index can be divided by age, region, occupation, etc. Therefore, when we consider the correlation between two observations, not only the individual difference (instinctive difference) should be contained but also age, region, occupation difference, that is, the factors under consideration need to be paid attention to. In this way, not all observations can share the same variance.

Under this situation, linear mixed-effect model which suffices to taking account the correlation of observations contained in a data set can be applied to solve traditional linear model's limitations.

3.2 Classical linear mixed-effects model

Linear mixed-effects model (LMM) is a more useful and realistic model to analyze real data sets. It also can be called Hierarchical Linear model and as its name implies, this model divides data sets into levels according to certain grouping factors[4]. For multilevel data, we are able to show the expression of the classical linear mixed-effects at a given level of a grouping factor as following:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i \tag{3}$$

where $\mathbf{y}_i$ is a $n_i \times 1$ vector of responses, $\mathbf{X}_i$ is a $n_i \times p$ design matrix of fixed effects, β is a $p \times 1$ vector of fixed effects, $\mathbf{b}_i$ is a $q \times 1$ vector of random effects in level i, $\mathbf{Z}_i$ is a $n_i \times q$ matrix of covariates which shows the correlation between responses $\mathbf{y}_i$ and random effects b_i, ε_i is the vector of residual errors for level i.

What should be emphasized is that $\mathbf{Z}_i$ contains known values of q covariates corresponding to q random effects chosen from its distribution[4]:

$$\mathbf{Z}_i = (\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}, \cdots, \mathbf{z}_i^{(q)}) \tag{4}$$

Moreover, $\mathbf{b}_i$ is unobservable. It delivers the information that random effect lacks pattern, causing difficulty being calculated.

In LMM, observations are considered not necessarily independent and have heteroscedasticity. The correlation between observations in the same level are reflected in the distribution of $\mathbf{b}_i$ and ε_i. Since they are in the same level, they are able to follow bivariate normal distribution:

$$\mathbf{b}_i \sim N(0, \mathbf{G}) \qquad \varepsilon_i \sim N(0, \mathbf{R}_i) \tag{5}$$

where $\mathbf{b}_i$ is independent of ε_i. Moreover, $\mathbf{G}$ and $\mathbf{R}_i$ can be specified that:

$$\mathbf{G} = \sigma^2 G \qquad \mathbf{R}_i = \sigma^2 R_i \tag{6}$$

where G and R_i both are variance function which represents the weight of observation's variance decided by parameter θ_G and θ_{R_i} respectively. Therefore, when random effects $\mathbf{b}_i$ is known, then the conditional distribution of $\mathbf{y}_i$ can be formulated:

$$E[\mathbf{y}_i|\mathbf{b}_i] = \mu_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i \tag{7}$$

$$Var[\mathbf{y}_i|\mathbf{b}_i] = \sigma^2 R_i \tag{8}$$

When $\mathbf{b}_i$ is not given, the unconditional distribution of $\mathbf{y}_i$ can be defined as:

$$E[\mathbf{y}_i] = \mathbf{X}_i\beta$$
$$Var[\mathbf{y}_i] = \sigma^2[\mathbf{Z}_i G \mathbf{Z}_i' + R_i]$$

Combing data sets of all levels of a grouping factor, we can get the classical formula of LMM for all data:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \varepsilon \tag{9}$$

where $\mathbf{Y} = (\mathbf{y}_1', \mathbf{y}_2', \cdots, \mathbf{y}_N')'$ is the $n \times 1$ vector of responses, where $n_1 + n_2 + \cdots + n_N = n$, β is the $p \times 1$ vector of fixed effects, $\mathbf{X}$ is the $n \times p$ design matrix for fixed effects, $\mathbf{Z}$ is the $n \times Nq$ matrix of random effects, $\mathbf{b}$ is the $Nq \times 1$ vector of random effects, where $\mathrm{b} = (\mathbf{b}_1', \mathbf{b}_2', \cdots \mathbf{b}_N')$, ε is the $n \times 1$ vector of errors, $\varepsilon = (\varepsilon_1', \varepsilon_2', \cdots, \varepsilon_N')'$. Therefore, the unconditional distribution and conditional distribution can be expressed respectively as

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2(\mathbf{Z}G\mathbf{Z}' + R) \qquad \mathbf{Y}|_{\mathbf{b}} \sim N(\mathbf{X}\beta + \mathbf{Z}\mathbf{b}, \sigma^2 R) \tag{10}$$

The benefit of linear mixed-effects model is that as I mention before that actual data set $\mathbf{Y}$ cannot all follow normal distribution because of the level difference, this model can take level influence as random effects so that $\mathbf{Y}$ can be expressed in multivariate normal distribution form. On the other hand, according to $\mathbf{Y}$'s variance representing variance-covariance form, we can know that $\mathbf{Y}$ are not independent and error terms are also divided into different levels, better considering real data's features.

Linear mixed-effects model has wide use in software, such as SAS, SPSS, Matlab as well as R. This article will focus on linear mixed-effects models using R and the lme4 package giving lmer() function to construct linear mixed-effects models for discover knowledge.

3.3 Estimation

In order to know how linear mixed-effects model is obtained, we need to figure out the parameters' estimation. Two common ways to estimate parameters in linear mixed-effects model are maximum likelihood (ML) estimation and restricted maximum likelihood (REML) estimation[4]. The conditional distribution of $\mathbf{y}_i$ given $\mathbf{b}_i$ are not appropriate for constructing the likelihood function because we don't know the real value of random effects $\mathbf{b}_i$. Therefore, marginal distribution of $\mathbf{y}_i$ is applied to build up ML and REML function.

3.3.1 Maximum likelihood estimation

Summarizing the parameters contained in linear mixed-effects model above, we get three types of parameters, $\beta = (\beta_1, \beta_2, \cdot \cdot \cdot, \beta_p)$, σ^2, $\theta = (\theta_G, \theta_R)$. Estimators of them can be obtained by simultaneously maximizing the log-likelihood function with respect to these parameters. However, it is a numerically complex work which needs to find an optimum in a multidimensional parameters space. Fortunately it can be simplified by profile likelihood technique.

With parameters β, σ^2, θ, we have likelihood expression given that:

$$\begin{aligned} L_{ML}(\beta, \sigma^2, \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma\sqrt{det(V_i)}} exp\left\{-\frac{1}{2}\frac{(y_i - X_i\beta)^2}{\sigma^2 det(V_i)}\right\} \\ &= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\prod_{i=1}^{n} exp\left\{-\frac{1}{2}\frac{(y_i - X_i\beta)^2}{\sigma^2 det(V_i)}\right\} \end{aligned} \tag{11}$$

where $V_i = Z_i G(\theta_G) Z_i' + R(\theta_R)$. Ignore constant part and take log operation, we get log-likelihood function that:

$$\begin{aligned} l_{ML}(\beta, \sigma^2, \theta) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\log[\det(V_i)] \\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)' V_i^{-1}(y_i - X_i\beta) \end{aligned} \tag{12}$$

Assume that θ is known, then maximizing (12) with respect to β for every value of θ leads to an estimation of β given by

$$\hat{\beta}(\theta, \sigma^2) = (\sum_{i=1}^{n} X_i' V_i^{-1} X_i)^{-1} \sum_{i=1}^{n} X_i' V_i^{-1} y_i \tag{13}$$

By plugging (13) into (12), we gain the log-profile likelihood function:

$$\begin{aligned} l^*_{ML}(\sigma^2, \theta) &= l_{ML}(\hat{\beta}(\theta), \sigma^2, \theta) \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log[\det(V_i)] - \frac{1}{2\sigma^2}\sum_{i=1}^{n} r_i' V_i^{-1} r_i \end{aligned} \tag{14}$$

where $r_i = r_i(\theta) = y_i - X_i\beta\hat{(\theta)}$.

In this way, the function does not depend on β which means the parameter space has lower dimension than previous one. Then use the same method, maximizing $l^*_{ML}(\sigma^2, \theta)$ with respect to σ^2 for every known value of θ leads to the estimation of σ^2:

$$\hat{\sigma}^2_{ML}(\theta) = \sum_{i=1}^{n} r_i' V_i^{-1} r_i / n \tag{15}$$

By plugging (15) into (14), we get a log-profile likelihood function for θ:

$$l^*_{ML}(\theta) = l^*_{ML}(\hat{\sigma^2}_{ML}, \theta) = -\frac{n}{2}\log(\hat{\sigma^2}_{ML}) - \frac{1}{2}\sum_{i=1}^{n}\log[\det(V_i)] - \frac{n}{2} \tag{16}$$

Therefore, there are less parameters in the parameter space again. Then maximization of $l^*_{ML}(\theta)$ can yield an estimator $\hat{\theta}_{ML}$ of θ. Plugging $\hat{\theta}_{ML}$ into (13) and (15) produces estimator $\hat{\beta}_{ML}$ of β and $\hat{\sigma}^2_{ML}$ of σ^2 that:

$$\hat{\beta}_{ML} = \hat{\beta}(\hat{\theta}_{ML}) = (\sum_{i=1}^{n} X_i'\hat{V_i}^{-1}X_i)^{-1}\sum_{i=1}^{n} X_i'\hat{V_i}^{-1}y_i \tag{17}$$

$$\hat{\sigma^2}_{ML} = \hat{\sigma^2}_{ML}(\hat{\theta}_{ML}) = \sum_{i=1}^{n} r_i'\hat{V_i}^{-1}r_i/n \tag{18}$$

However, there is a significant limitation on maximum likelihood estimation that ML estimators $\hat{\sigma^2}_{ML}$ and $\hat{\theta}_{ML}$ are both biased because they don't adjust for the uncertainty in estimation of β which means $\hat{\sigma^2}_{ML}$ and $\hat{\theta}_{ML}$ will change value following β's altering and cannot estimate σ^2 and θ accurately. However, σ^2 and β can be better estimated by restricted maximum likelihood estimation which will be discussed at next part.

3.3.2 Restricted maximum estimation

In order to obtain unbiased estimates of σ^2 and θ, we should use an estimation approach that is orthogonal to the estimation of β which means using a way to make estimates of σ^2 and θ independent to estimation of β[19]. We can consider the likelihood function of a set of $n - p$ independent contrasts of y to do it where p is the dimension of β. After obtaining $\hat{\beta}(\theta)$, the log-restricted-likelihood function is given by:

$$\begin{aligned} l^*_{REML}(\sigma^2, \theta) = -\frac{n-p}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log[\det(V_i)] - \frac{1}{2\sigma^2}\sum_{i=1}^{n} r_i'V_i^{-1}r_i \\ -\frac{1}{2}\log[\det(\sum_{i=1}^{n} X_i'V_i^{-1}X_i)] \end{aligned} \tag{19}$$

From this function, maximizing of $l^*_{REML}(\sigma^2, \theta)$ with respect to σ^2 leads to estimator of σ^2 that:

$$\hat{\sigma^2}_{REML} = \sum_{i=1}^{n} r_i'V_i^{-1}r_i/(n-p) \tag{20}$$

Plugging (20) into (19), we get a function with respect to θ only:

$$\begin{aligned} l^*_{REML}(\theta) = -\frac{n-p}{2}[\log(\sum_{i=1}^{n} r_i'V_i^{-1}r_i/(n-p)) + 1] - \frac{1}{2}\sum_{i=1}^{n}\log[\det(V_i)] \\ -\frac{1}{2}\log[\det(\sum_{i=1}^{n} X_i'V_i^{-1}X_i)] \end{aligned} \tag{21}$$

Estimator of θ can be obtained from maximization from (21), which can be applied to get estimators of β and σ^2, respectively.

3.4 Confidence Interval estimation

Confidence Interval(CI) gives a range for a random variable based on a certain confidence level, that is, how much people believe in it. Therefore, people can regard the values in this confidence interval having the same or similar influence. Constructing confidence interval has significant meanings either in theory or empirical problem. On the one hand, when estimating parameters or predicting responses, we can get certain values of parameters or responses whereas considering the preciseness of science and mathematics, these values undoubtedly lacks certainty or accuracy. On the contrary, if we can locate estimation and prediction in ranges, the conclusion must sound more convincing. On the other hand, in reality, when salesmen recommend table for customers who are willing to purchase for their room, they should provide tables having several different sizes which can match rooms of customers rather than say that there is only one best table for you. Here, the size of table can be considered as a random variable and offering the lower bound and upper bound of this variable is the work the wise salesmen should do.

Nonetheless, confidence interval can be calculated from different method based on which three confidence intervals include profile likelihood confidence interval, Wald confidence interval and bootstrap confidence interval. Each one possesses disparate ideas and assumptions. Next we will discuss the underlying concepts of these CIs and their application in LMMs.

3.4.1 Profile likelihood confidence Interval

The assumption of profile likelihood confidence interval(PLCI) is that the estimator does not have to follow normal distribution[15]. The concept of PLCI is very similar to the profile likelihood technique previously mentioned. In a model, we assume β is our interest parameter and $\mathbf{b}$ is the vector of all nuisance parameters. Thus, $L(\beta, \mathbf{b})$ is the maximum likelihood function based on two random variables β and $\mathbf{b}$ and the profile likelihood function of β is defined as

$$L_1(\beta) = \underset{\mathbf{b}}{max} L(\beta, \mathbf{b}) \tag{22}$$

which means the maximum likelihood function of β with MLE value of $\mathbf{b}$. With this concept, we can consider the confidence interval next. In the hypothesis test, the null hypothesis is constructed like this: $H_0 : \beta = \beta_0$. In this circumstance, building confidence interval is equivalent to the question that finding all β_0 which can make the H_0 not be rejected under the $100(1-\alpha)\%$ confidence level. Then we use the **likelihood ratio test**[15]:

$$2[\log L(\hat{\beta}, \hat{\mathbf{b}}) - \log L_1(\beta_0)] < \chi^2_{1-\alpha}(1) \tag{23}$$

where $L(\hat{\beta})$ is the maximum likelihood with MLE of all parameters and $L_1(\beta_0)$ handle one less parameters so the left hand of this formula follows Chi-square distribution. Therefore, all β_0 satisfying above formula can form a confidence interval for β. Since $\log L(\hat{\beta})$ and $\chi^2_{1-\alpha}(1)$ are constant, we can rearrange the expression:

$$\log L_1(\beta_0) > \log L(\hat{\beta}, \hat{\mathbf{b}}) - \chi^2_{1-\alpha}(1)/2 \tag{24}$$

It is likely to get a graph like this:

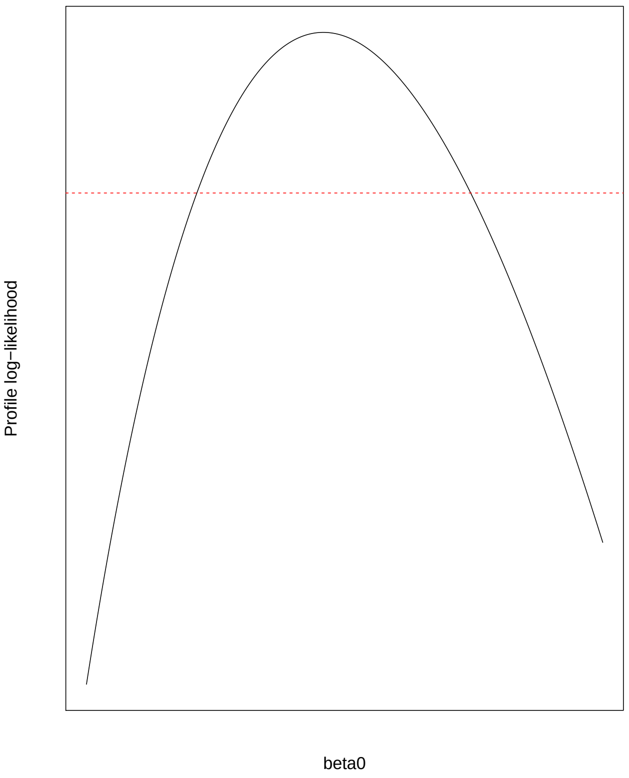


Figure 1: Profile likelihood function for β_0

Therefore, the part of the curve above the red line form the confidence interval we desire for.

3.4.2 Wald confidence Interval

Wald confidence Interval takes Wald Test into account that:

$$\frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} \sim N(0,1) \tag{25}$$

with the assumption that the difference between the two will be approximately normally distributed. According to this test, when considering the confidence Interval in LMM, we only need to know the estimates and variance of the parameters included. For the fixed effects β, we get the estimate above and its variance-covariance that:

$$\hat{\beta}_{ML} = (\sum_{i=1}^{n} X'\hat{V}^{-1}X)^{-1} \sum_{i=1}^{n} X'\hat{V}^{-1}y \tag{26}$$

$$\text{varcov}(\hat{\beta}) = \hat{\sigma^2}(\sum_{i=1}^{n} X'\hat{V}^{-1}X)^{-1} \tag{27}$$

Thus, we extract the diagonal of variance-covariance matrix as variance of $\hat{\beta}$, the wald confidence interval for $\hat{\beta}$ can be expressed that:

$$(\sum_{i=1}^{n} X'\hat{V}^{-1}X)^{-1} \sum_{i=1}^{n} X'\hat{V}^{-1}y \pm \sqrt{\text{diag}(\hat{\sigma^2}(\sum_{i=1}^{n} X'\hat{V}^{-1}X)^{-1})} \tag{28}$$

3.4.3 Bootstrap confidence Interval

Bootstrap confidence Interval comes from the idea that "pulling itself up by its own bootstrap"[6]. On simple understanding, it means doing large number of bootstraps from the original data. Assume we have original data $\{y_1, y_2, y_3, ..., y_n\}$, and we build LMM for this data which contains parameter $\Theta = (\beta, \theta, \sigma^2)$. After maximum likelihood estimation or restricted maximum likelihood estimation, we are able to obtain estimates $\hat{\Theta} = (\hat{\beta}, \hat{\theta}, \hat{\sigma^2})$. In order to find confidence interval, we should calculate the variation of $\hat{\Theta}$ around Θ, that is, $\delta = \hat{\Theta} - \Theta$. Hence, confidence interval based on $\alpha\%$ confidence level can be showed as:

$$\Theta \in [\hat{\Theta} - \delta_{\alpha/2}, \hat{\Theta} + \delta_{1-\alpha/2}] \tag{29}$$

For the purpose of finding δ, we process bootstrap operation. Firstly, we take resamples from the original data $\{y_1, y_2, y_3, ..., y_n\}$ and receive n new observations notated as $\left\{y_1^{(1)}, y_2^{(1)}, ..., y_n^{(1)}\right\}$ which has the same distribution as the original data. After the LMM estimation for the new data, we obtain new parameter estimates $\hat{\Theta}^{(1)}$ and the first variation is calculated as $\delta^{(1)} = \hat{\Theta}^{(1)} - \hat{\Theta}$. Then this kind of resampling operation should run repeatedly for m times, normally more 1000 times and a matrix of resample can form:

$$\begin{bmatrix} y_1^{(1)} & y_1^{(2)} & \dots & \dots & y_1^{(m)} \\ y_2^{(1)} & y_2^{(2)} & \dots & \dots & y_2^{(m)} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ y_n^{(1)} & y_n^{(2)} & \dots & \dots & y_n^{(m)} \end{bmatrix} \tag{30}$$

Each column represents one resampling and products one δ, thereby one sequence of δ is generated finally $\{\delta^{(1)}, \delta^{(2)}, \delta^{(3)}, ..., \delta^{(m)}\}$ and we sort them from smallest to biggest. Thus, $\delta_{\alpha/2}$ is at the $\alpha/2$ percentile and $\delta_{1-\alpha/2}$ is at the $1-\alpha/2$ percentile. In this way, the confidence interval of parameters can be given. Bootstrap introduces us a simple and straightforward angle to observe the variation of estimates. With the law of large number, the resample distribution can be a good approximation to the true distribution.

3.4.4 Comparison

Since these three methods follow different concepts, we should consider carefully about which method should be applied in different circumstances.

Profile confidence interval has very wide applications because of its low restriction that it is still available for the estimators not normally distributed[5]. Due to this reason, the default demand $confint()$ in R software adopts this profile likelihood method and it is useful when analyzing LMMs. Although Wald confidence interval is very common, it has difficult use in LMM because there is not clear method to account for the parameters contained in random effects, that is, σ^2 and θ. The Wald confidence interval cannot be calculated for these parameters in LMM. As for bootstrap confidence interval, the CI from this method is relatively valid. This kind of sampling cannot improve pointer estimates. It is obvious that every bootstrap is chosen from the same data pool and follows the same steps, there are no new information reflected even after all bootstraps[6]. This is also the reason why confidence interval calculated from bootstrap is more wider than profile likelihood method and Wald test so bootstrap confidence interval is hardly used due to its less preciseness.

To summarize, profile confidence interval is the most useful and convincing one during LMM analysis.

3.5 Structure of linear mixed-effects model in R

Optimization is always an important problem in model selection. In order to know the best linear mixed-effects model fitting the data, first we need to know how to construct linear mixed-effects model. To be clear, it is the same as the problem that how to choose variables as fixed effects terms and random effects terms. Normally, fixed effects terms can be variable of interest or the ones having clear influence pattern to response. Random effects can be the variables showing highly unclear pattern to response, such as subjects because subjects contains too many levels, it is hard to distinguish pattern.

In R, we can build several different structures of linear mixed-effects models[18]. The purpose of constructing different models is firstly to investigate the effects of certain variables to responses and secondly to provide candidate models for choosing better models. Here I use y to represent response, x to represent fixed effects, b to represent random effects and $Data$ is the name of data set the model reads[12][8].

```
model1 <- lmer(y ~ 1+(1|b),Data)
```

The formula of model 1 is:

$$Y = Z_1 b_1 + \varepsilon \tag{31}$$

It shows no fixed effects here and there are random effects of one factor b_1 which have influence to response. This model has not much meaning since there are no interest effects included so it makes no sense with total random data assumed.

```
model2 <- lmer(y ~ x+(1|b),Data)
```

The formula of model 2 is:

$$Y = X\beta + Z_1 b_1 + \varepsilon$$

It is the classical linear mixed-effects model mentioned above and is also one of the common models we use that is composed of one interest fixed effect and one factor of random effects.

```
model3 <- lmer(y ~ -1 + x + (1|b),Data)
```

Model3 is very similar to Model2, "-1" means the result directly shows the effects of x to responses rather than shows the common effects, i.e. intercept and coefficient effects separately.

```
model4 <- lmer(y ~ x+(x|b),Data)
```

Model4 shows random effect b is correlated with fixed effect x, b has effects to x but has no effects to response. In this model, the slope of x at different level of b will not be the same because of the effects delivered from b[13].

```
model5 <- lmer(y ~ x+(1+x|b),Data)
```

Model5 shows random effects b not only has effects to fixed effects x, but also has effects to intercept of response[2]. The calculation of model 4 and model 5 is very complicated especially when the volume of data is large and there are lots of levels in a factor which may cause R runs slowly even crashes. Therefore, during the process we choose models, it is important to think what's our object of study and the possibly obvious random effects whose number in consideration should not be too large. Then construct the candidate models from simple one to more complex one and the next step is to compare models by some criterion.

3.6 Criterion for model selection

As for a data set, there might be several models applied to analyze it and how to measure which model is better is the main problem. There are some important and useful criterion to measure better model such as log-likelihood, Akaike information criterion (AIC), Bayesian information criterion (BIC), p-value. Obviously, different models possess different focus and purposes which result in different values. I will mainly discuss how these criterion derive and their underlying logic applied in linear mixed-effects model.

3.6.1 Log-likelihood

It is the most simplest criterion which has the expression that

$$l(\Theta) = \log L(\Theta) = \log \left(f(Y|\Theta)\right) \tag{32}$$

where $Y = (y_1, y_2, \cdots, y_n)'$ is the vector of observations; $\Theta = (\beta, \sigma^2, \theta)$ represents the vector of all parameters contained in linear mixed-effects model where $\beta = (\beta_1, \beta_2, \cdots, \beta_k)$ is the vector of fixed effects and $\theta = (\theta_G, \theta_R)$ is the vector of random effects; $f(Y|\Theta)$ is the likelihood function of observations.

According to this expression, we can see clearly the design of this criterion: Θ is the key component of our model so that it can represent our model, Y is the data set observed. Therefore, the meaning of f is that how much the data fits our candidate models and adding log is to avoid zero value in the likelihood[17]. The higher value of the log-likelihood achieves, the better the data fits our models.

However, log-likelihood criterion has a huge problem that it doesn't consider the number of parameters because we prefer a model with high log-likelihood and small number of parameters which can not only guarantee high level of fitting but also require less calculations and operations. Normally, it is widely considered that more number of parameters would increase the value of log-likelihood and the effects are significant when the number of parameters is few. However, when the number of parameters is large enough, every more parameters added in model would have highly little influence to log-likelihood value, which is showed below:

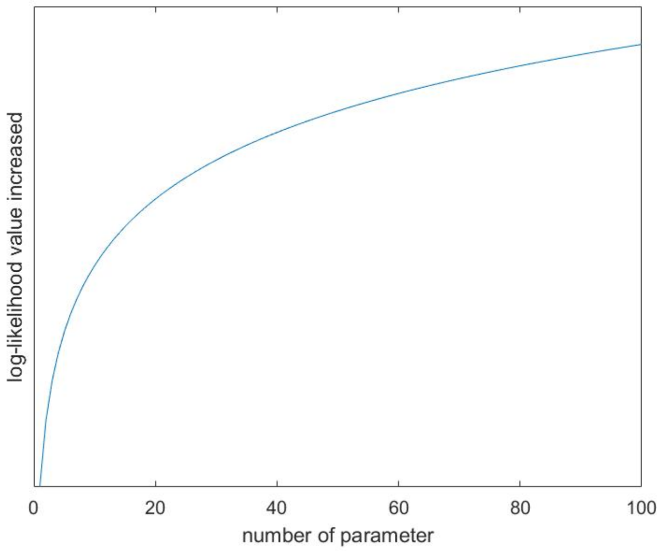


Figure 2: Relationship between log-likelihood value and paramater

That is, it makes little sense to having too many parameters. On the contrary, we prefer the model with less parameters when its log-likelihood value are just little less than the value of the model with more parameters. In this circumstance, akaike information criterion (AIC) and Bayesian information criterion (BIC) both consider the number of parameters and overcome log-likelihood's limitation.

3.6.2 Akaike Information Criterion (AIC)

Akaike information criterion is a standard to measure the goodness of statistical model fitting. It was founded and developed by Japanese statistician Akaike. This criterion suffices to weigh the complexity of the estimated model and the goodness of the fitting data of the model. When we use maximum likelihood estimation (MLE) in linear-mixed effect model, log-likelihood $l(\hat{\Theta})$ can be achieved where $\hat{\Theta} = (\hat{\beta}, \hat{\sigma^2}, \hat{\theta})$ contributes to maximum log-likelihood.

According to the significance of parameters mentioned above, $l(\hat{\Theta})$ is somehow an overestimate. AIC assumes it is overestimated by the difference between the expectation of maximum log-likelihood based on observed data and the expectation of log-likelihood using the parameters derived from former maximum log-likelihood process based on future data which can be collected in the future investigation, that is,

$$E_{\Theta^t}[l_Y(\hat{\Theta}_Y)] - E_{\Theta^t}[l_{Y^F}(\hat{\Theta}_Y)] \tag{33}$$

where Y-observed data, Y^F-future data, Θ^t-true model parameters, $\hat{\Theta}_Y$-maximum likelihood estimate of observed data[3]. We can understand it as the difference of the fitness of parameter from MLE in different data set, that is, observed data and future data. Since $E_{\Theta^t}[l_{Y^F}(\hat{\Theta}_Y)] - E_{\Theta^t}[l_Y(\hat{\Theta}_{Y^F})]$, then

$$E_{\Theta^t}[l_Y(\hat{\Theta}_Y)] - E_{\Theta^t}[l_{Y^F}(\hat{\Theta}_Y)] = E_{\Theta^t}[l_Y(\hat{\Theta}_Y) - l_Y(\hat{\Theta}_{Y^F})] \tag{34}$$

The reason is that observed data and future data are both from the same population so that the sample data as well as the log-likelihood value have the same or similar distribution which can be showed in graph below:

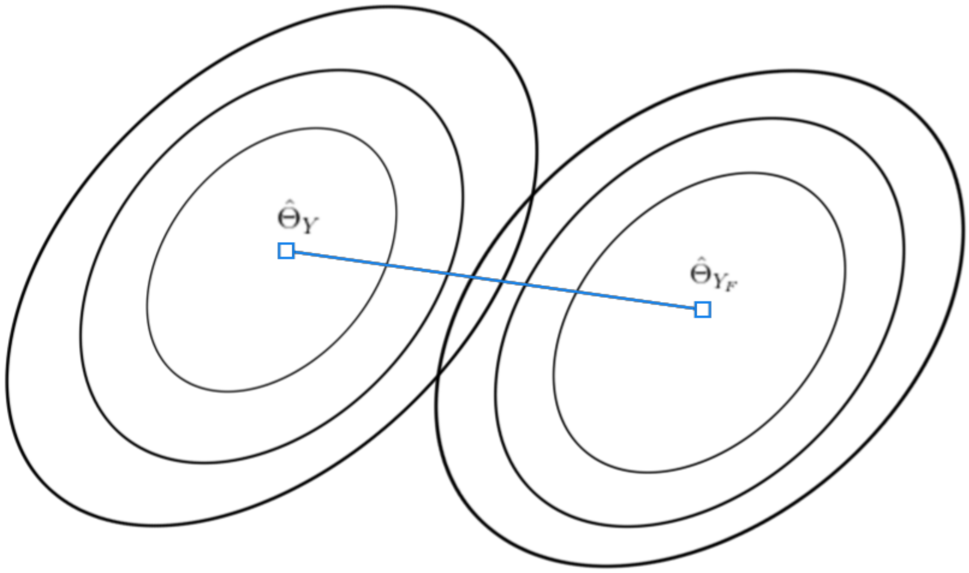


Figure 3: Distribution of log-likelihood

Referring to this graph, obviously the same or similar distribution shape and different position cause their symmetry. Therefore,

$$l_{Y^F}(\hat{\Theta}_Y) = l_Y(\hat{\Theta}_{Y^F}) \tag{35}$$

which concludes the equation (25). Then we need to find the value of $l_Y(\hat{\Theta}_Y) - l_Y(\hat{\Theta}_{Y^F})$. By using the Taylor expansion series, we do expand $l(\Theta)$ with second order at the point $\Theta = \hat{\Theta}$:

$$\begin{aligned} l(\Theta) &\approx l(\hat{\Theta}) + (\Theta - \hat{\Theta})\dot{l}(\hat{\Theta}) - \frac{1}{2}(\Theta - \hat{\Theta})'(-\ddot{l}(\hat{\Theta}))(\Theta - \hat{\Theta}) \\ &= l(\hat{\Theta}) - \frac{1}{2}(\Theta - \hat{\Theta})'(-\ddot{l}(\hat{\Theta}))(\Theta - \hat{\Theta}) \end{aligned} \tag{36}$$

Because $l(\hat{\Theta})$ is the maximum likelihood, $\dot{l}(\hat{\Theta}) = 0$. Next we still apply Taylor expansion on $\dot{l}(\hat{\Theta})$ with first order at point $\hat{\Theta} = \Theta^t$,

$$0 = \dot{l}(\hat{\Theta}) \approx \dot{l}(\Theta^t) + \ddot{l}(\Theta^t)(\hat{\Theta} - \Theta^t) \tag{37}$$

After rearranging, we get

$$\hat{\Theta} - \Theta^t \approx \frac{\dot{l}(\Theta^t)}{-\ddot{l}(\Theta^t)} \tag{38}$$

Multiply $\sqrt{n}$ at both sides, we get

$$\sqrt{n}(\hat{\Theta} - \Theta^t) \approx \frac{\frac{\dot{l}(\Theta^t)}{\sqrt{n}}}{\frac{-\ddot{l}(\Theta^t)}{n}} \tag{39}$$

According to central limit Theorem i.e.

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$$

$$\frac{\dot{l}(\Theta^t)}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial l(\Theta|y_i)|_{\Theta=\Theta^t}}{\partial \Theta} \sim N(0, var(\frac{\partial l(\Theta|y_i)|_{\Theta=\Theta^t}}{\partial \Theta})) \tag{40}$$

Then by the law of large number that when n comes large, $\bar{X_n} \to \mu = E[X]$

$$\frac{-\ddot{l}(\Theta^t)}{n} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l(\Theta|y_i)|_{\Theta=\Theta^t}}{\partial \Theta^2} = E[-\frac{\partial^2 l(\Theta|y_i)|_{\Theta=\Theta^t}}{\partial \Theta^2}] \tag{41}$$

Fisher information defines that

$$E[-\frac{\partial^2 l(\Theta|y_i)|_{\Theta=\Theta^t}}{\partial \Theta^2}] = var(\frac{\partial l(\Theta|y_i)|_{\Theta=\Theta^t}}{\partial \Theta}) = J \tag{42}$$

Thus,

$$\sqrt{n}(\hat{\Theta} - \Theta^t) \approx \frac{\frac{\dot{l}(\Theta^t)}{\sqrt{n}}}{\frac{-\ddot{l}(\Theta^t)}{n}} \sim N(0, \frac{1}{J})$$

Move out $\sqrt{n}$, we get the distribution that

$$(\hat{\Theta} - \Theta^t) \sim N(0, (nJ)^{-1}) \tag{43}$$

Since $\hat{\Theta}_Y$ and $\hat{\Theta}_{Y^F}$ are members in family $\hat{\Theta}$, then

$$(\hat{\Theta}_Y - \Theta^t) \sim N(0, (nJ)^{-1})$$

$$(\hat{\Theta}_{Y^F} - \Theta^t) \sim N(0, (nJ)^{-1})$$

After simple calculation of $(\hat{\Theta}_Y - \Theta^t) - (\hat{\Theta}_{Y^F} - \Theta^t)$, we get

$$(\hat{\Theta}_Y - \hat{\Theta}_{Y^F}) \sim N(0, 2V) \tag{44}$$

where $V = (nJ)^{-1}$

With conclusion achieved above, we come back to equation (25) and expand $l(\hat{\Theta}_Y)$ with second order at point $\hat{\Theta}_Y = \hat{\Theta}_{Y^F}$:

$$\begin{aligned} l(\hat{\Theta}_Y) &\approx l(\hat{\Theta}_{Y^F}) + (\hat{\Theta}_Y - \hat{\Theta}_{Y^F})\dot{l}(\hat{\Theta}_{Y^F}) - \frac{1}{2}(\hat{\Theta}_Y - \hat{\Theta}_{Y^F})'(-\ddot{l}(\hat{\Theta}_{Y^F}))(\hat{\Theta}_Y - \hat{\Theta}_{Y^F}) \\ &= l(\hat{\Theta}_{Y^F}) - \frac{1}{2}(\hat{\Theta}_Y - \hat{\Theta}_{Y^F})'(-\ddot{l}(\hat{\Theta}_{Y^F}))(\hat{\Theta}_Y - \hat{\Theta}_{Y^F}) \end{aligned}$$

According to (35),

$$l(\hat{\Theta}_Y) - l(\hat{\Theta}_{Y^F}) = \frac{1}{2}\Delta' V^{-1} \Delta \tag{45}$$

where $\Delta = \sqrt{2V} \cdot Z$, $Z \sim N(0,1)$. Therefore, we can calculate the main difference

$$E_{\Theta^t}[l(\hat{\Theta}_Y) - l(\hat{\Theta}_{Y^F})] = \frac{1}{2}E[\Delta' V^{-1} \Delta] = \frac{1}{2}E[Z'\sqrt{2V}V^{-1}\sqrt{2V}Z] = E[Z'Z] \tag{46}$$

where $Z'Z$ follows the Chi-square distribution i.e. $Z'Z \sim \chi^2(p)$ where p is the degree of freedom and here p can also represent the number of parameters in linear mixed-effects model. Obviously, from Chi-square distribution's property, the expectation of $Z'Z$ is p. Finally we gain the part overestimated:

$$E_{\Theta^t}[l(\hat{\Theta}_Y) - l(\hat{\Theta}_{Y^F})] = p \tag{47}$$

Thus, AIC can be expressed as $l(\hat{\Theta}) - p$. However, normally, in R software, the AIC formula is defined as

$$AIC = 2p - 2l(\hat{\Theta})$$

Both expressions have the same structure with slight difference including sign and multiplies. To be clear, in this essay, we use formula

$$AIC = 2p - 2l(\hat{\Theta}) \tag{48}$$

where p is the number of parameters, $l(\hat{\Theta})$ is the maximum log-likelihood. The lower AIC value a model has, the better the model is.

3.6.3 Bayesian Information Criterion (BIC)

In statistics, there are two ways to optimize models. On the one hand, adding more parameters in models can increase model's complexity. On the other hand, collecting more observations or data suffices to facilitate model's ability to describe data sets. AIC considers the parameter problems whereas the number of observations is not included. However, BIC considers both of them and takes them as measurement for models.

BIC provides an algorithm to approximate the log marginal likelihood of candidate models and chooses the one having smaller value as the better model. The formula of BIC is showed below:

$$BIC = p \cdot \log n - 2l(\hat{\Theta}) \tag{49}$$

where p is the number of parameters n is the number of observations

The key logic and idea behind BIC is calculating the marginal likelihood given the likelihood function $L(\Theta)$ and prior function $g(\Theta)$, i.e. the distribution of parameters.

$$P(Y|\Theta) = \int L(\Theta)g(\Theta)d\Theta = \int exp\left\{\log L(\Theta)g(\Theta)\right\}d\Theta \tag{50}$$

Next we take Taylor expansion of $\log(L(\Theta)g(\Theta)$ with second order at point $\Theta = \hat{\Theta}$, we get

$$\begin{aligned}\log(L(\Theta)g(\Theta)) \approx{}& \log(L(\hat{\Theta})g(\hat{\Theta})) + (\Theta - \hat{\Theta})\dot{\log}(L(\hat{\Theta})g(\hat{\Theta})) \\ &+ \frac{1}{2}(\Theta - \hat{\Theta})'\ddot{\log}(L(\hat{\Theta})g(\hat{\Theta}))(\Theta - \hat{\Theta})\end{aligned}$$

Because $\hat{\Theta}$ makes $\dot{l}(\hat{\Theta}) = 0$, the second term can be expanded and it is equal to zero, i.e. $(\Theta - \hat{\Theta})(\log(L(\hat{\Theta})) + \log(g(\hat{\Theta})))' = 0$ where $g(\hat{\Theta})$ is constant, we express it as a. Then

$$\begin{aligned}\log(L(\Theta)g(\Theta)) &= l(\hat{\Theta}) + \log(a) + \frac{1}{2}(\Theta - \hat{\Theta})'\ddot{\log}(L(\hat{\Theta}))(\Theta - \hat{\Theta}) \\ &= l(\hat{\Theta}) + \log(a) - \frac{n}{2}(\Theta - \hat{\Theta})'J(\Theta - \hat{\Theta})\end{aligned} \tag{51}$$

where $J = -\frac{l\ddot{o}g(L(\hat{\Theta}))}{n}$ which we have mentioned in AIC part. Moreover, from the AIC derivation process, we know $(\Theta - \hat{\Theta}) \sim N(0, (nJ)^{-1})$, so $\Theta \sim N(\hat{\Theta}, (nJ)^{-1})$ and $g(\Theta)$ is able to be expressed that:

$$g(\Theta) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(nJ)^{-1}}}\exp(-\frac{1}{2}\frac{(\Theta - \hat{\Theta})'(\Theta - \hat{\Theta})}{(nJ)^{-1}}) \tag{52}$$

In this circumstance,

$$\begin{aligned}L(\Theta)g(\Theta) &= a \cdot L(\hat{\Theta}) \cdot \exp(-\frac{n}{2}(\Theta - \hat{\Theta})'J(\Theta - \hat{\Theta})) \\ &= a \cdot L(\hat{\Theta})g(\Theta) \cdot (2\pi)^{p/2} \cdot \sqrt{\det(nJ)^{-1}}\end{aligned} \tag{53}$$

Come back to equation that

$$\begin{aligned}P(Y|\Theta) &= \int L(\Theta)g(\Theta)d\Theta \\ &= a \cdot L(\hat{\Theta}) \cdot (2\pi)^{p/2} \cdot \sqrt{\det(nJ)^{-1}} \int g(\Theta)d\Theta \\ &= a \cdot L(\hat{\Theta}) \cdot (2\pi)^{p/2} \cdot n^{-k/2} \cdot \sqrt{\det(J)^{-1}}\end{aligned}$$

Then take log at both sides,

$$\log P(Y|\Theta) = l(\hat{\Theta}) - \frac{k}{2}\log n + \log a + \log((2\pi)^{p/2} \cdot \sqrt{\det(J)^{-1}})$$

When n comes to large enough, we can ignore the constant terms, here third term and fourth term is constant. Finally, we get

$$\log P(Y|\Theta) = l(\hat{\Theta}) - \frac{k}{2}\log n \tag{54}$$

Adding some slight rearrangement to this equation, we can get the BIC formula applied in R software:

$$BIC = k \cdot \log(n) - 2l(\hat{\Theta})$$

3.6.4 F-test

Sometimes AIC and BIC will give different choices between two models so under this circumstance, we need to refer to other criterion. F-test is able to calculate p-value which can be used to judge whether these two models are significantly different and make decision about model selection. The logic behind F-test is the comparison of models' deviance which is defined that:

$$D = 2[l(\hat{\Theta}_{max}; y) - l(\hat{\Theta}; y)] \tag{55}$$

where $\hat{\Theta}_{max}$ is the MLE for the parameter vector in the saturated model which has N parameters, $\hat{\Theta}$ is the MLE for the parameter vector in the candidate model. Assume there are two models $m0$ and $m1$ with degree of freedom, p and q respectively where $p < q$ and the parameters of $m1$ contains $m0$'s parameters. We can calculate their deviance denoted as D_0 and D_1 which are applied for F test:

$$F = \frac{(D_0 - D_1)/(p - q)}{D_1/(N - q)} \tag{56}$$

After we obtain the F value, p-value also can be achieved by referring to F table. Next the process of making decision depends on our own confidence degree to this test. Normally 95 percent confidence level and 99 percent confidence level are preferred choices, so based on 95 percent confidence level, if the p-value is less than 5 percent, there is significant difference between $m0$ and $m1$ because of the extra parameters in $m1$, we tend to choose $m1$ which has more parameters. On the contrary, when the p-value is more than 5 percent, there is not significant difference existing so that the model with less parameters $m0$ is our choice.

However, there is limitation in F-test using that the comparison between models with different types of parameters is difficult to give conclusion because we don't know which parameters play important role significantly fitting the data. Therefore, using F-test demands us to select appropriate models to compare.

To conclude, we can use Anova(model1, model2) which is an useful command in R software[7]. It can directly calculate log-likelihood, AIC, BIC of models respectively as well as p-value of F-test. What should be warned is that the comparison will only be valid if models are designed to fit the same data set. Based on the analysis above including the merits and weaknesses of different criterion, I want to emphasis the rule of model selection in linear mixed-effects model:

- Ignore log-likelihood value from Anova() because of its big disadvantage.
- For the purpose of studying certain parameters' significance, F-test is preferred,i.e. referring to p-value
- AIC and BIC are good ways to measure the goodness of fit to data and when two models' AIC and BIC value are quite similar, the model with smaller parameter' number should be considered.

4 Application: Knowledge discovery from high-throughput movie rating data

When people come to video website, one thing almost everyone can experience is that the website will automatically recommend videos or movies some of which are able to attract people. This is the work of the recommender system(RS) behind the websites. However, their idea to recommend mostly relies on the watching history. For example, if the system has observed $\ll$ *The Avengers* $\gg$ is in one's watching history, it will discovery some movies which have the similar traits and operate the recommendation. From the perspective of the name, RS is possible to provide $\ll The\,Avengers\,2 \gg$ and relative sequal movies. As for the same firm or actors/actresses, it is likely to offer the movies manufactured by Marvel. Even from the genre angle: science fiction, RS can give people some science fiction movies, such as $\ll Interstellar \gg$.

However, what if the people who visit the website are the new users which deliver the crucial information that they have no watching history. In this way, how to recommend movie for these people? Nowadays RS considers the similarities of movies whereas we can shift to discover the sameness of users' traits[11]. Thinking about the favorite movies for people with different age, we may have the feeling that people from different age group are likely to show distinct taste or preference to movie. The most obvious example which everyone may experience is that adventure or science fiction movie may be teenagers' favorites whereas seniors possibly reveal more love on comedy films. Then here comes a question that how about the attitude difference of people from different age levels toward the same movies or the same type of movies? Here we use figures and data to do knowledge discovery rather than depending on intuition or feelings. If we measure the "attitude" as the rating on movies, this question is equivalent to the problem that what is the hidden pattern of the relationship between age and the rating to movies. If we can find a particular pattern, it has high possibility to be referred by recommender system. RS may know which types of movies are the favorite ones of people from certain age level, and then recommend these types of movies to the new users within the same age level.

In order to process this topic, firstly we need database about detailed information of users which includes age and other traits, such as gender, occupation, nation, hobby, etc. Moreover, movie information is also the necessity which composes of movie name, movie genre and movie rating from users. For this purpose, The ***ml-1m*** dataset is used in this experiment. It provides us 6040 users, 3952 movies and movie rating made on a 5-star scale. Among these ratings, each user has at least 20 ratings. Furthermore, User information contains age, gender, occupation and zip-code which are all in the category type. Movie information including Movie ID, title and genres is listed and movie genre can be found in **Appendix**. Therefore, this is a real big data and the pattern excavated from it will be highly convincing. Although some false information exists in the dataset, the huge volume allows us to ignore its effect to our data analysis. The data structure is showed as below:

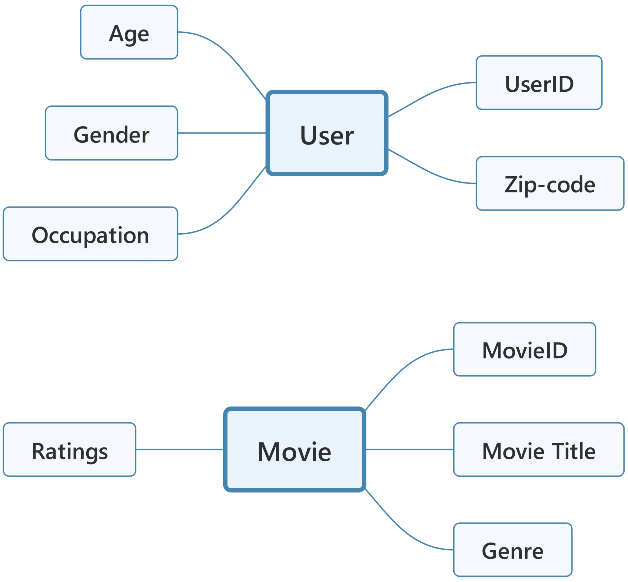


Figure 4: Data structure of user and movie

Our task is to discover the age influence on movie rating, thereby we need to rearrange the data provided above. Since UserID from User dataset is linked with the ratings in Movie dataset, we can ignore zip-code which makes little sense in this experiment and construct a data chain aiming for one movie or types of movies that:

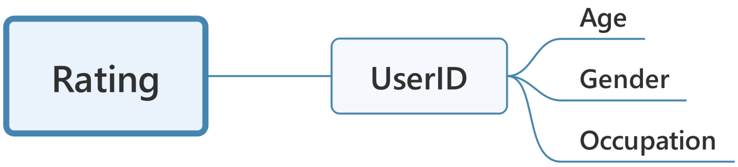


Figure 5: Data structure

According to this data structure, we can construct linear mixed-effects models where age is our interest variable treated as fixed effect and gender is candidate random effects as well as occupation. Three different models can be designed that:

```
Model1 <- lmer(Rating ~-1 + Age + (1|Gender), data)
```

```
Model2 <- lmer(Rating ~-1 + Age + (1|Occupation), data)
```

```
Model3 <- lmer(Rating ~-1 + Age + (1|Gender) + (1|Occupation
    ), data)
```

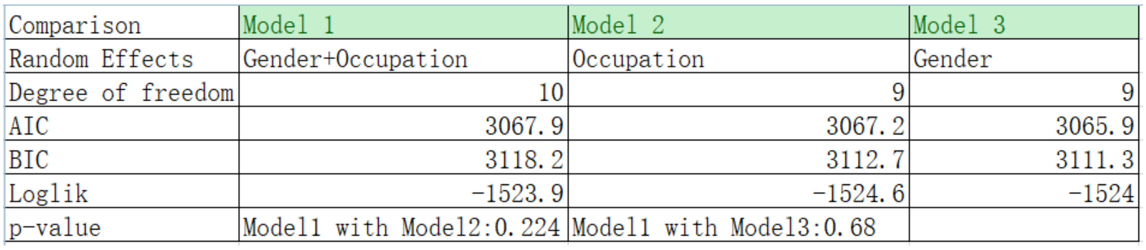
As for how to choose from these models, we should focus on the real data analysis as well as our model selection criteria.

Given the data as well as candidate models above, our study about age discovery starts from one certain movie to movies of different genres. During this process, we are willing to find general pattern underlying.

4.1 One certain movie: $\ll Life\,Is\,Beautiful \gg$

Firstly, we are willing to see the relationship between users' age and rating for one specific movie. Here we choose the film $\ll Life\,is\,beautiful \gg$ labeled as comedy and romantic movie which won the best foreign language film at the 71st Oscar Awards. Since it is a known work, it received lots of rating records from users that there are 1152 ratings given.

For our model, the users' age is regarded as our interest, i.e. the fixed effect and we need to consider how to choose random effects based on the models provided above. After using **Anova()** test, the values of model selection criteria are provided that:



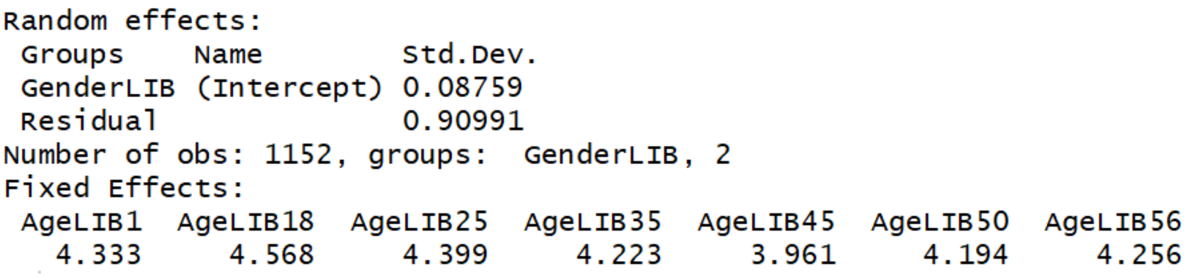
Comparison	Model 1	Model 2	Model 3
Random Effects	Gender+Occupation	Occupation	Gender
Degree of freedom	10	9	9
AIC	3067.9	3067.2	3065.9
BIC	3118.2	3112.7	3111.3
Loglik	-1523.9	-1524.6	-1524
p-value	Model1 with Model2:0.224	Model1 with Model3:0.68	

Figure 6: Model selection

According to this diagram, Model 3 has the lowest AIC and BIC value so Model 3 is the best model based on AIC and BIC criteria, but Model 1 obtains the largest loglikelihood value. Literally, Model 1 is the best one referring to loglikelihood and it has more parameters. However, after the comparison on p-value, we can make our decision. Under 95% confidence interval, the p-value of Model 1 and Model 2 is larger than 0.05 which means there is little difference between Model 1 and Model 2. Since Model 2 has the fewer parameters, it is better than Model 1. Due to the same reason for Model 1 and Model 3, Model 3 is better than Model 1. In the meantime, observing from AIC, BIC and loglikelihood criteria, we can see the superiority of Model 3 compared with Model 2. Therefore, in conclusion, Model 3 is the best model for the data *«Life is beautiful»*, which can be constructed that:

```
LIB_Model <- lmer(Rating ~-1 + Age + (1|Gender), Life is
    beautiful)
```

lmer command is capable of estimating the parameters contained in models and we can have the result that:



```
Random effects:
 Groups    Name        Std.Dev.
 GenderLIB (Intercept) 0.08759
 Residual              0.90991
Number of obs: 1152, groups:  GenderLIB, 2
Fixed Effects:
 AgeLIB1  AgeLIB18  AgeLIB25  AgeLIB35  AgeLIB45  AgeLIB50  AgeLIB56
   4.333     4.568     4.399     4.223     3.961     4.194     4.256
```

Figure 7: Model result

It shows the standard deviation of random effect, gender and residual as well as what we desire: fixed effects of all age levels. To be clear, the radar chart is able to show the fixed effects and their differences intuitively.

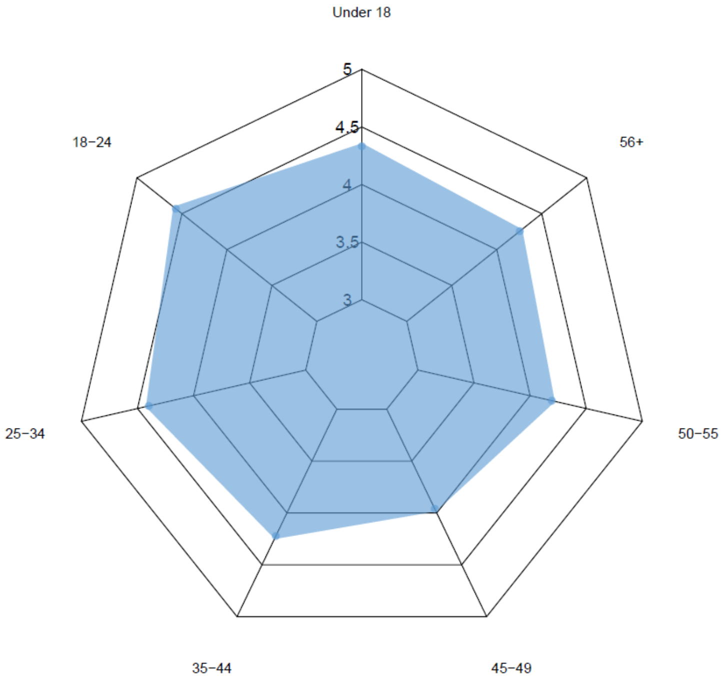
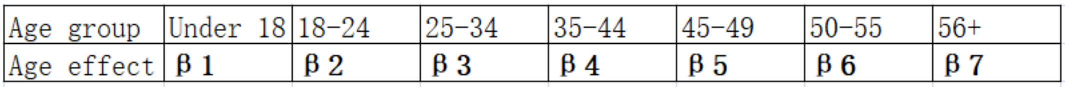


Figure 8: Age effect distribution for $\ll life\, is\, beautiful \gg$

In this graph, visually we can see the rating from users between 18-24 is much higher than the one from other age groups. However, this deduction is unreasonable since we are not sure about whether the rating of users from 18-24 is significantly higher than others. To judge the significance of differences, we need to operate hypothesis test. The effects of every age group, that is, the parameters of age levels can be represented that:



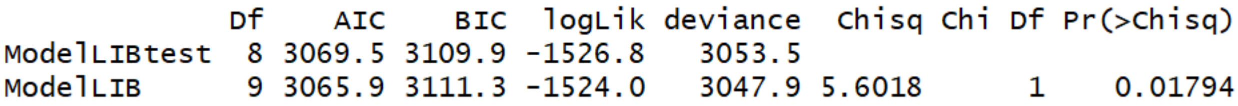
Age group	Under 18	18-24	25-34	35-44	45-49	50-55	56+
Age effect	β 1	β 2	β 3	β 4	β 5	β 6	β 7

Figure 9: Age effects expression

Therefore, our task is to prove whether β_2 is significantly the largest effect. From the result of $LIBmodel$, β_3 is the closest one to β_2 which leads to our hypothesis test that:

- $H_0 : \beta_2 = \beta_3$
- $H_1 : \beta_2 \neq \beta_3$

H_1 represents the original model $LIBmodel$ and we can replace all data belonging to 25-34 age level to 18-24 age level, that is, combining β_2 and β_3 together which brings new model labeled as $LIBmodeltest$. We compare these two models and obtain result that:

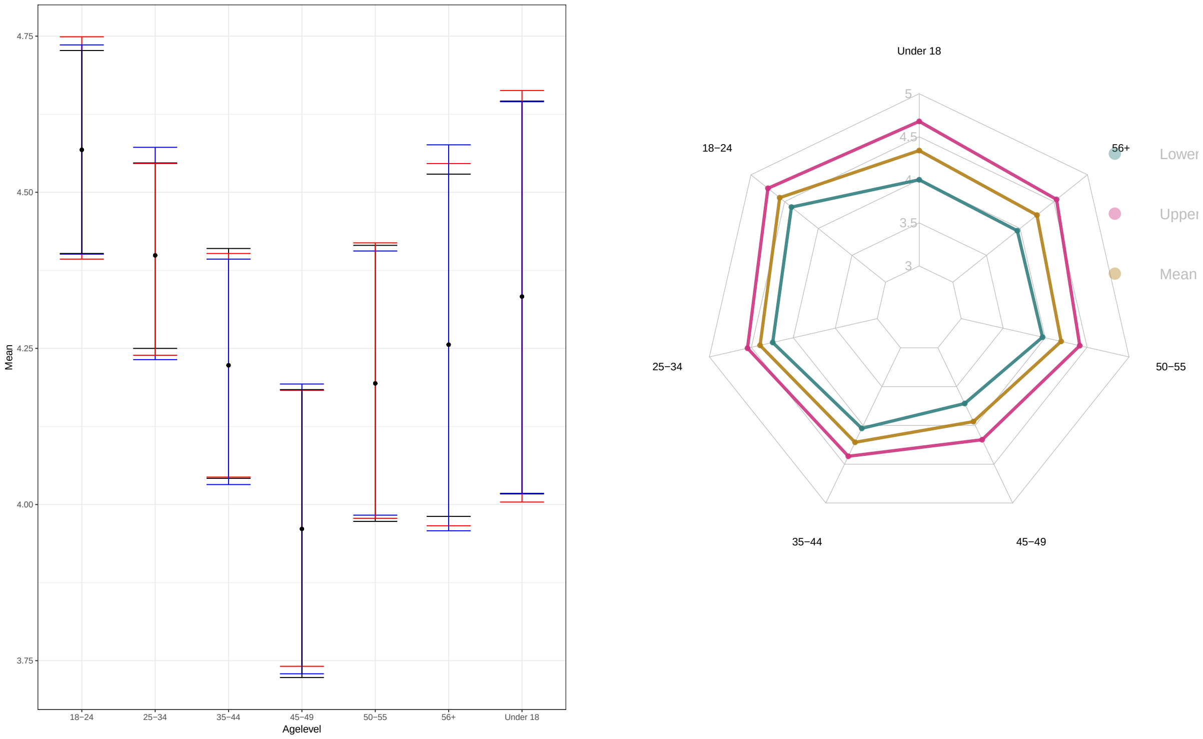


```
             Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
ModelLIBtest  8 3069.5 3109.9 -1526.8   3053.5
ModelLIB      9 3065.9 3111.3 -1524.0   3047.9 5.6018      1    0.01794
```

Figure 10: Anova analysis of β_2 and β_3

The low $p-value$ indicates that it is significant to reject H_0 under 95% confidence interval and there is big difference between β_2 and β_3, thereby the conclusion that 18-24 users give the most high rating for *«Life is beautiful»* than other age groups can be drawn.

Besides observing the estimates of age effects, we are able to use package *lm4* to calculate the confidence interval of age effects. Based on three types of confidence interval(profile CI, wald CI, bootstrap CI), their ranges can expressed respectively in one graph that:



(a) Three types of CI for LIB

(b) radarchart for CI

Figure 11: Confidence Intervals of $\ll Life\, is\, beautiful \gg$

where the black range line represents profile confidence interval, the red range line explains wald confidence interval and the blue line shows the bootstrap confidence interval. From the graph, we can see that there are just little differences among these three confidence intervals and they are almost overlapping with each other so that it is easy for us treat all confidence intervals as the same one here. Therefore, taking the wald confidence interval, we are capable of building a radar chart for CI of age effects where the area between red line and green line is the CI place showed above.

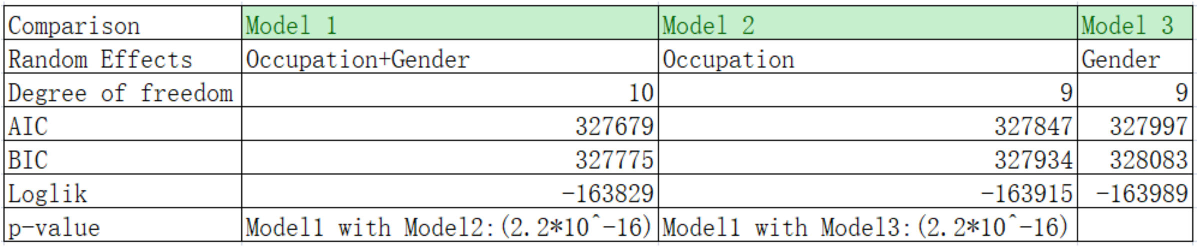
Back to the graph of *three types of CI for LIB*, it is not difficult to discover that the CI ranges of 18-24, 25-34, 35-44 are narrower than the ones of 45-49, 50-55, 56+ and under 18. According to formula of wald confidence interval, it can be deduced that the standard deviations of age effects of 18-24, 25-34, 35-44 are smaller than the ones of 45-49, 50-55, 56+, under 18. More narrow a confidence interval is, more new information is reflected. Therefore, the age effects of 18-44 for $\ll Life\,is\,beautiful \gg$ are more meaningful and it can better display the range where the true age effects lie.

Nonetheless, the analysis above only comes from one movie which hardly possesses much meaning used for application and reference. What we want to search for is the general pattern so that although there is some interesting discoveries of $\ll Life\,is\,beautiful \gg$, it still has poor influence on our main target. On the other hand, if one specific movie genre's certain pattern can be excavated, then this knowledge would be more plausibly used for recommending movies for people.

4.2 One genre: COMEDY

Comedy is a very big genre in ***ml-1m*** because it has the largest number of rating data (107009 ratings) among all movie genres. Therefore, the analysis of this big dataset will be more meaningful than one certain movie. The target is still to find the relationship between age and rating whereas the research range expands to comedy, one certain genre.

Firstly, we still need to choose the proper model to analyze the data and the results are showed:



Comparison	Model 1	Model 2	Model 3
Random Effects	Occupation+Gender	Occupation	Gender
Degree of freedom	10	9	9
AIC	327679	327847	327997
BIC	327775	327934	328083
Loglik	-163829	-163915	-163989
p-value	Model1 with Model2:(2.2*10^-16)	Model1 with Model3:(2.2*10^-16)	

Figure 12: Model selection

The comparison is very clear that Model 1 has the smallest AIC and BIC values and the largest loglikelihood value. Moreover, the p-values of Model 1 with Model 2 and Model 1 with Model 3 are very small which gives the information that although Model 1 contain one more parameter than Model 2 and Model 3, the deviance of it is significantly lower than the other two. Therefore, we should choose Model 1 as our design model:

```
Comedy_Model1 <- lmer(Rating ~-1 + Age + (1|Occupation) +
    (1|Gender), comedy)
```

According to the result from ***lmer*** operation, we can obtain the radar chart of age effects in comedy movies showed below.

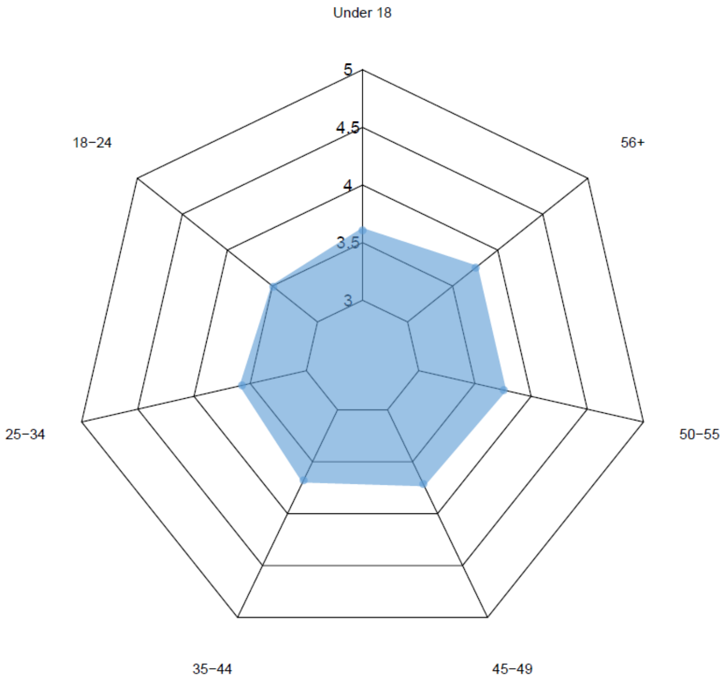
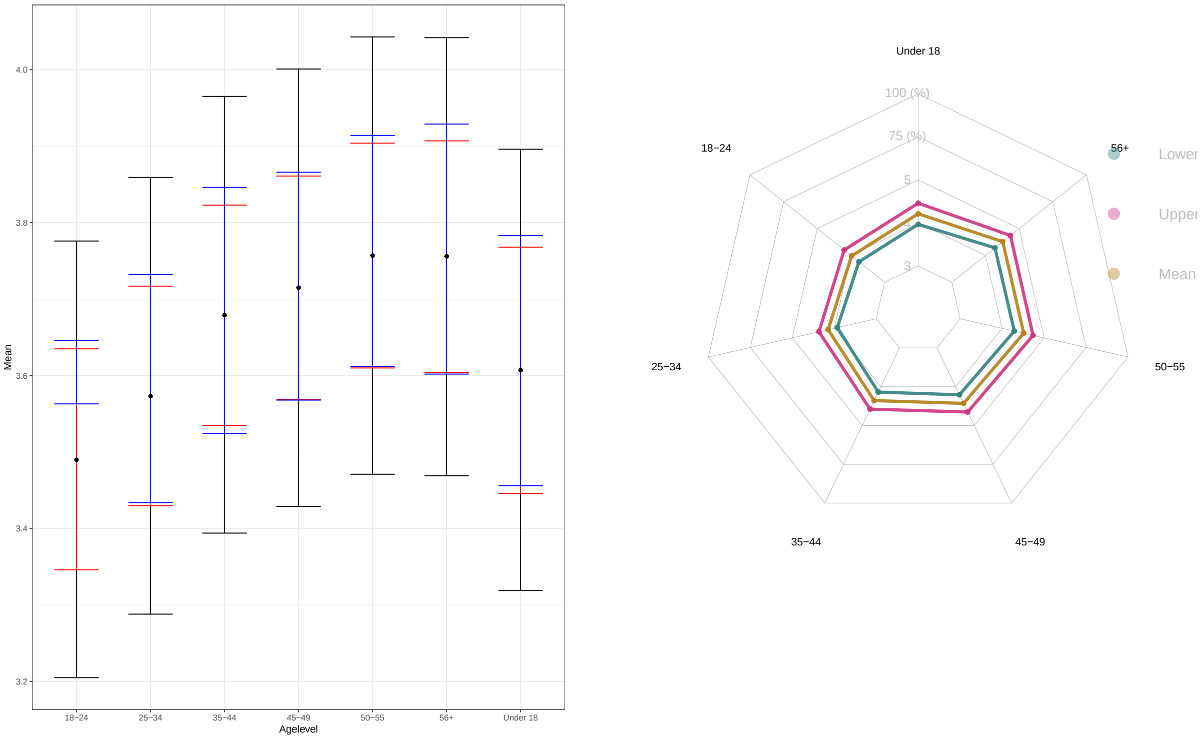


Figure 13: Age Effects for comedy

Since age effect mentioned here is just the mean value of corresponding parameter, it is impossible for us to guess 100% accurately that the age effect to one's rating for comedy is the number displayed in the graph above. The convincing way is to provide confidence interval for each age effect, here we choose 99% confidence interval because the comedy data is much larger than the data of *«Life is beautiful»* which is about tenfold higher, we can put higher confidence level on the comedy analysis, that is, 99%. Using this level, three types of confidence interval for each age effect possess images below (Black for profile CI, red for wald CI, blue for bootstrap CI). Surprisingly, these confidence intervals here perform great differences and we should make choice among them. Firstly, there is a clear observation that profile confidence interval has much wider ranges for every age effects contrasted to the other two so that it contain less new information for the location of age effects' estimates and we will not use profile confidence interval here. Moreover, although wald confidence interval and bootstrap one look very similar, one problem on bootstrap CI is remarkable on the age level 18-24. At this age level, bootstrap CI is highly narrow whereas it does not contain the estimate which is a serious issue indeed. Considering backing the definition of bootstrap CI, every bootstrap is random and its merit is just based on the large number of repetitions to simulate as far as possible whereas no improvement is produced. Hence, it is reasonable for the existence of the estimate exclusion. Then looking at the wald CI, for every age level, the range performs high proximity and it demonstrates the significance of every age effect CI is similar. Thereby, we can trust the CIs of different age levels evenly. In this circumstance, wald confidence interval should be taken account for the range of age effects and its radar chart is showed.



(a) CIs for comedy

(b) Wald CI of comedy

Figure 14: Confidence Intervals of comedy

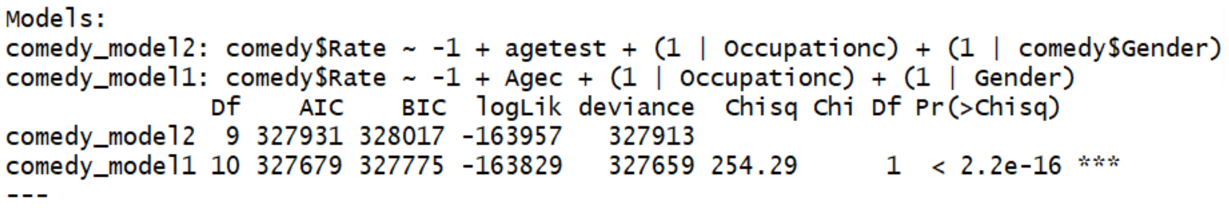
From age effect of comedy graph, on the first cursory glimpse, age above 45 seems to have higher effects and people with age 18-24 are likely to have relatively lower effect, whereas the whole effects look like a round pie. To investigate the difference level of effects, we are trying to process hypothesis test for all differences. Since it would be a tedious task, we use simple method to operate it for the purpose of understanding easily.

At first, we choose two effects having the largest difference i.e. in the model, they will share the same parameter and the model contains one less parameter. Then, we use **Anova()** to compare previous model and the model after change with 99% confidence level. If the result shows they are not significant, then all differences between age effects are of no significance. On the contrary, when the result reflects significance, we need to contrast the value of the second high difference and do hypothesis test again until the result shows the difference of models are not significant. In this way, we are able to sort the age effects according to significance.

In this experiment, we know the effects $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ are 3.607, 3.490, 3.573, 3.679, 3.715, 3.757, 3.756 respectively. Therefore, we choose β_6, β_2 and do the first hypothesis test:

- $H_0 : \beta_6 = \beta_2$
- $H_1 : \beta_6 \neq \beta_2$

where H_0 shows comedyModel2, H_1 is the comedyModel1 defined previously. we can get the **Anova()** result:



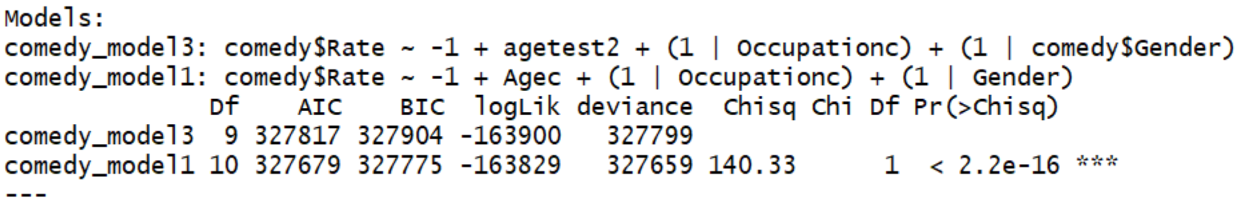
```
Models:
comedy_model2: comedy$Rate ~ -1 + agetest + (1 | Occupationc) + (1 | comedy$Gender)
comedy_model1: comedy$Rate ~ -1 + Agec + (1 | Occupationc) + (1 | Gender)
              Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
comedy_model2  9 327931 328017 -163957   327913
comedy_model1 10 327679 327775 -163829   327659 254.29      1  < 2.2e-16 ***
---
```

Figure 15: Comedy: Model 1 vs Model 2

The p-value is quite small and less than 0.01, so we should do the second hypothesis test with β_7 and β_2 :

- $H_0 : \beta_7 = \beta_2$
- $H_1 : \beta_7 \neq \beta_2$

where H_1 indicates comedyModel3 and we obtain the **Anova()** result:



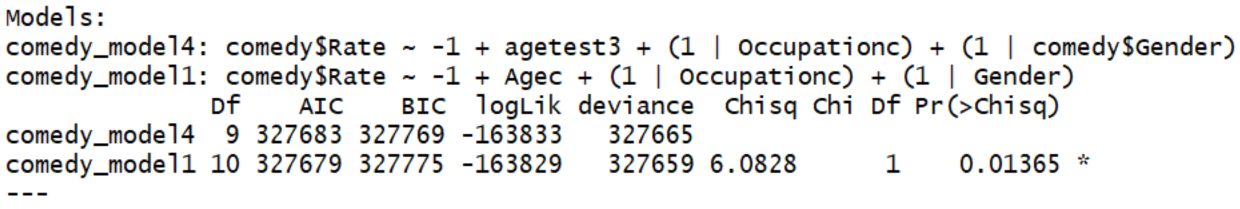
```
Models:
comedy_model3: comedy$Rate ~ -1 + agetest2 + (1 | Occupationc) + (1 | comedy$Gender)
comedy_model1: comedy$Rate ~ -1 + Agec + (1 | Occupationc) + (1 | Gender)
              Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
comedy_model3  9 327817 327904 -163900   327799
comedy_model1 10 327679 327775 -163829   327659 140.33      1  < 2.2e-16 ***
---
```

Figure 16: Comedy: Model 1 vs Model 3

The p-value still shows the non-significance. Then, according to this rule and after several runs of tests, we come to the difference between β_4 and β_5:

- $H_0 : \beta_4 = \beta_5$
- $H_1 : \beta_4 \neq \beta_5$

where H_0 indicates comedyModel4 so the result is:



```
Models:
comedy_model4: comedy$Rate ~ -1 + agetest3 + (1 | Occupationc) + (1 | comedy$Gender)
comedy_model1: comedy$Rate ~ -1 + Agec + (1 | Occupationc) + (1 | Gender)
              Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
comedy_model4  9 327683 327769 -163833   327665
comedy_model1 10 327679 327775 -163829   327659 6.0828      1    0.01365 *
---
```

Figure 17: Comedy: Model 1 vs Model 4

The p-value is larger than 0.01, so the difference is significant and the loop hypothesis tests are finished. According to the tests, there are some surprising discoveries which can be mentioned:

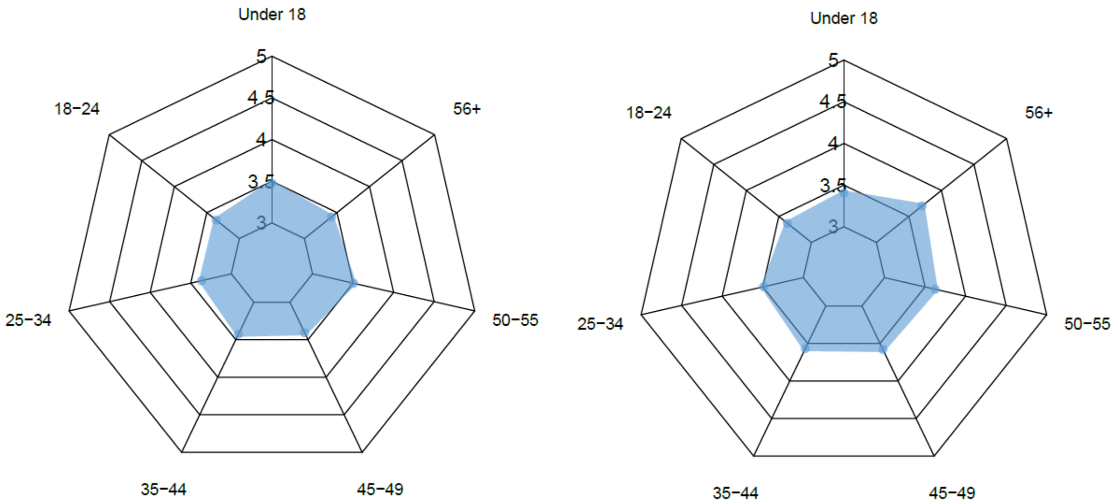
- The differences of effects of rating from people with age more than 35 are insignificant so that they can be regard as a unit with high rating on comedy which we can notate as "The high rating group". In the meantime, the differences of effects of rating from people with age less than 35 are also insignificant which we notate as "The low rating group".

- The difference between every member from "The high rating group" and every member from "The low rating group" is significant.

In this circumstance, we can define young people as the ones with age 1-34 and senior people as the ones with age more than 35. Therefore, the hidden pattern from comedy analysis is that **YOUNG PEOPLE ARE MORE PARTICULAR THAN SENIOR ONES FOR COMEDY**. Thus, it may deliver a message for some video websites that they can recommend more comedy movies to the senior and less for young people and in return receive more positive feedback. However, how about movies of other genres, we have not done researched on the movies of other types and the whole movies recorded. If they give the similar patterns in the radar chart, then it can deliver the information that people from the same age group have the same attitudes to all movies of different genre. Thus, now we are continuing to analyse movies of other types and observe the finding.

4.3 Others movie genres

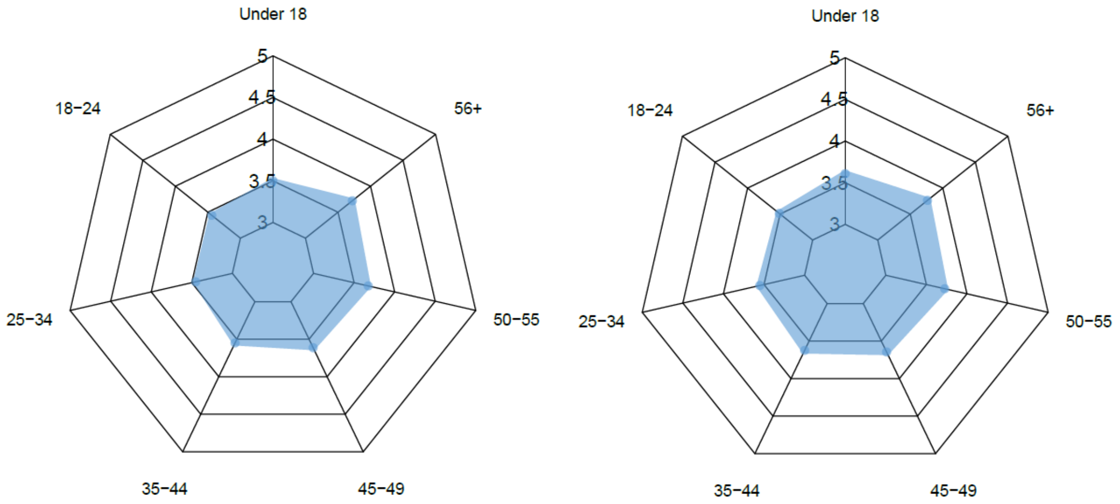
Furthermore, We use linear mixed-effects models to analyse data of other genres in which age is regarded as fixed effect, occupation and gender are considered as random effects chosen from the model selection criteria mentioned before. Here we choose science fiction movie, children movie and adventure movie to show in radar chart. The age effects for these models are showed below:



(a) Age Effect of sci-fi

(b) Age Effect of children

Figure 18: Age effects for sci-fi and children movie



(a) Age Effect of adventure

(b) Age Effect of the whole movie

Figure 19: Age effects for adventure and whole movies

After hypothesis tests respectively, these three graphs all give patterns that young people which is defined as ones in age groups "Under 18", "18-24", "25-34" marks the movies lower than the old people which is defined as people in age groups "45-49", "50-55", "56+". It contributes to our thinking that as for ranking movie, young people are picky than old ones which is also supported by comedy analysis which is displayed above. Moreover, in order to provide more evidence, rather than analysing movies of other types, we directly investigate age effects distribution in all movies recorded, i.e 3,000,000 pieces of users' rating for movies, from which we get chart above.

Clearly, whatever by visually observing or under hypothesis test, age effect differences between young people (Under 34) and old people (Beyond 45) are considerable. Therefore, to summarize, from analysing the data from ***ml-1m***, we are able to deduce the information that young people are picky than old people to mark movies.

5 Summary

Facing today's society filling with millions of, even billions of data, knowledge discovery becomes a important channel for transforming data to information used for multiple applications. In microcosm, everyone's rational decision should be based on knowledge discovery. For example, people should consider and decide which job they are willing to take by observing the trend of this job by collecting a large number of data may including degree of satisfaction, salary, etc. For study, it is the main method for scholars to do research. For profit, business is able to earn more money by applying discovery few people know. Due to its high significance, how to discover knowledge is a hot topic.

In this paper, we are interested in the recommender system which undoubtedly is also one of the knowledge discovery method. Recommender system is widely used in lots of fields such as video websites, shopping applications. It mainly focus on the similarity of movies for recommendation with the assumption that users have history on websites or applications, but it cannot solve the problem of providing proper movies for new users without history. Therefore, instead of thinking of movies, we are trying to consider the sameness of users. If one pattern related to traits of users, then it can be regarded as a reference for recommender system.

In this research, we use the movie data for study and are willing to find the relationship between age and the rating for movie. After ensuring target, we should find a appropriate method to analyse it. Linear mixed-effects model(LMM) is our choice because it breaks thought the restrictions on linear model. The data under LMM do not need to be normally distributed and they can be dependent with different variance. Besides these advantages, LMM allows us to add random effects in models. Because in our analysis, we have many variables some of which are our interest and some of which are not our object. Nonetheless, we are somewhat sure they will have some influence on our response but their patterns are not our need. Therefore, placing them into position labeled as random effects is a wise and valid way. The data of movie also have this characteristic, age is our interest and other traits of users will be random effects.

In the data package ***ml-1m***, gender and occupation of user is the random effects so that we are capable of building three different models only with random effects' difference. By using model selection criteria including log-likelihood, Akaike Information criterion(AIC), Bayesian Information criterion(BIC) and F-test, the best model can be selected among three candidate models.

In order to find the pattern of age on movie rating, we divide our analysis into three parts according to different dataset. Firstly, we analyse the movie rating of one movie $\ll Life\, is\, beautiful \gg$ within which we are familiar with some operation of linear mixed-effects model, such as how to select best model, express the age effects as well as corresponding confidence intervals and hypothesis test on age effects differences. However, since one movie's discovery is meaningless and lacks universality, we next do analysis on dataset from one genre: comedy. From this result, we find young people defined as ones with age less than 35 have significant lower rating on comedy than senior people with age more than 35. After getting this interesting discovery, we have a further question that what about movies of other genres. Surprising, most of them show the very similar pattern even the all movie rating data gives the same conclusion. Thus, we obtain the discovery that young people are more particular in marking movie than senior ones.

This discovery has several meanings. The first one is for recommender system. It can recommend more movies especially from different genres for senior new users in one time. Since they will give relative high rating, this behavior can leave a good expression for the senior new users that this movie website can provide movies with high quality, thereby more amount of play and more profit can be earn for the website. The another meaning is that it is our first time using data to analysis a common phenomenon in real life. In our daily life, many people may have the feeling that old people are easily satisfied and young people's demand is difficult being met. For example, when a family is having a meal in a restaurant, it is very likely that the word to evaluate the meal from grandparents is good whereas young people's assessment may be just so-so or not good. It is hard for them to speak highly positive word. In this paper, we verify the phenomenon in movie rating and what about other fields, such as book rating, music rating which is still worth being discussed in the future.

References

[1] Al Borchers John Riedl Badrul Sarwar, Joseph Konstan. Applying knowledge from kdd to recommender systems. *University of Minnesota*, 1999.

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Statistics Computing*, arXiv:1406(1):133–199, 2014.

[3] J. Deleeuw. *Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle.* 1992.

[4] Andrzej Gałecki and Tomasz Burzykowski. *Linear Mixed-Effects Models Using R.* 2013.

[5] Hulya Bayrak Hatice Tul Kubra Akdur, Heniz Ozonur. A comparison of confidence interval methods of fixed effect in nested error regression model. *Journal of Natural and Applied Sciences*, 20(2), 2016.

[6] Jonathan Bloom Jeremy Orloff. Bootstrap confidence interval. 2014.

[7] Alexandra Kuznetsova, Per B. Brockhoff, and Rune Haubo Bojesen Christensen. lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 2017. This work is licensed under the licenses Paper: Creative Commons Attribution 3.0 Unported License.

[8] J.A.Bovaird L.Locker, L.Hoffman. On the use of multilevel modeling as an alternative to items analysis in psychlinguistic research. *Behavior Research Methods*, 39.

[9] C. J. Matheus, P. K. Chan, and G. Piatetsky-Shapiro. Systems for knowledge discovery in databases. *IEEE Transactions on Knowledge and Data Engineering*, 5(6):903–913, Dec 1993.

[10] Sangkil Moon, Paul K Bergey, and Dawn Iacobucci. Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *Journal of Marketing*, 74(1):108–121, 2010.

[11] 1 Abelardo Montesinos-López † José Crossa ‡ 1 José C. Montesinos-López § David Mota-Sanchez ** Fermín Estrada-González * Jussi Gillberg †† Ravi Singh ‡ Suchismita Mondal ‡ Osval A. Montesinos-López, * and Philomin Juliana‡. Prediction of multiple-trait and multiple-environment genomic data using rcommender systems. *Genes/Genomes/Genetics.*

[12] R.H.Baayen. Analyzing linguistic data: A practical introduction to statistics using r.

[13] D.M.Bates R.H.Baayen, D.J.Davidson. Mixed-effects modeling with crossed random effects for subjects and items. *journal of Memory and Language*, 59, 2008.

[14] J. Schafer. The application of data-mining to recommender systems. 2009.

[15] Christensen J Stryhn H. Confidence intervals by the profile likelihood method, with applications in veteinary epidemiology.

[16] Gregory Piatetsky-Shapiro Usama Fayyad and Padhraic Smyth. From data mining to knowledge discoveries in databases. *AI Magazine*, 17(3):37–54, 1996.

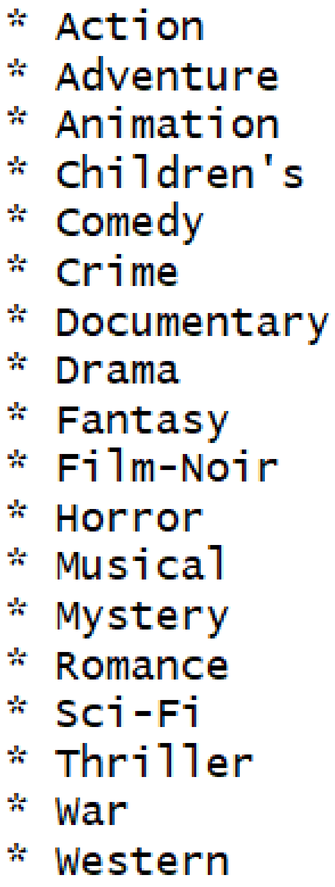
[17] Larry Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000.

[18] Bodo Winter. A very basic tutorial for performing linear mixed effects analyses. 2013.

[19] Dong Jin Yang and Xue Ying Zhong. The perception of film attractiveness and its effect on the audience satisfaction, intention and investment. *Journal of Service Science Management*, 09(1):21–27, 2016.

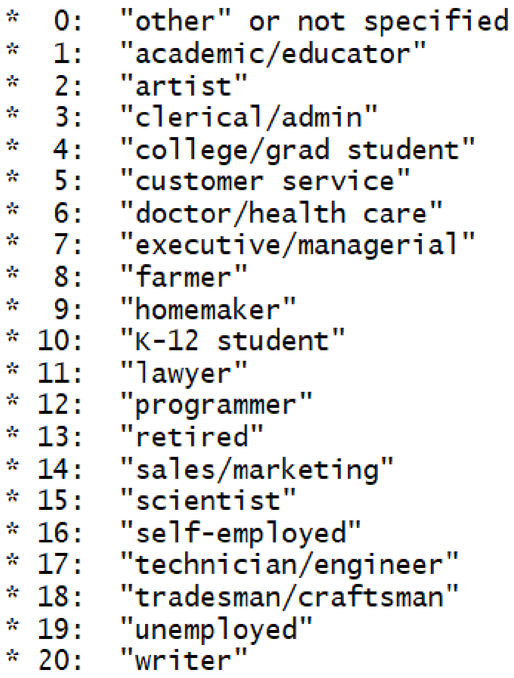
[20] XianXing Zhang, Yitong Zhou, Yiming Ma, Bee-Chung Chen, Liang Zhang, and Deepak Agarwal. Glmix: Generalized linear mixed models for large-scale response prediction. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 363–372, New York, NY, USA, 2016. ACM.

[21] Huan Zhao, Quanming Yao, Yangqiu Song, James Kwok, and Dik Lun Lee. Learning with heterogeneous side information fusion for recommender systems. 2018.



```
* Action
* Adventure
* Animation
* Children's
* Comedy
* Crime
* Documentary
* Drama
* Fantasy
* Film-Noir
* Horror
* Musical
* Mystery
* Romance
* Sci-Fi
* Thriller
* War
* Western
```

Figure 20: Movie Genre



```
*  0:  "other" or not specified
*  1:  "academic/educator"
*  2:  "artist"
*  3:  "clerical/admin"
*  4:  "college/grad student"
*  5:  "customer service"
*  6:  "doctor/health care"
*  7:  "executive/managerial"
*  8:  "farmer"
*  9:  "homemaker"
* 10:  "K-12 student"
* 11:  "lawyer"
* 12:  "programmer"
* 13:  "retired"
* 14:  "sales/marketing"
* 15:  "scientist"
* 16:  "self-employed"
* 17:  "technician/engineer"
* 18:  "tradesman/craftsman"
* 19:  "unemployed"
* 20:  "writer"
```

Figure 21: Occupation

A m1-1m