2021 MA678 Midterm Project

The analysis of Google play store App

Yanbing Chen BUID:U32747296

Abstract

This report aims to find the relationship between App price and App review, rating, size and installs. I use multilevel model to analyze whether installs, review, rating and size would influence Google play store App price, and I plotted a varying intercept figure. The results are that rating, installs are all have negative influence on price, and size presents a positive impact on price only in Medical category. The result demonstrates that maybe there are other more important factors affect price. Therefore, I consider add more addiction factors to model in future analysis.

Introduction

Google play store is a digital application platform operated and developed by Google, and it is also a digital media store that runs Android devices. Android users can download a variety of software, such as books, games, movies, music and other apps on the Google play store to achieve a variety of needs. When users search for Apps on the Google play store, the platform displays a variety of content related to these Apps, including software prices, software ratings, downloads, reviews, etc., which will provide an effective reference for users to select software. Due to several factors that can make influence on customers choice, I explore some relationships among them, and find out which factors can exert effect on Apps' price and how can these make influence on App price based on the data I gain. In addition, I use multilevel model to do analysis about App price and it's factors in this report because there are so many kinds of App in the Google play store.

Method

Data Cleaning and Processing

The data is published on Kaggle: Google play store Apps, and it's publisher scraped these data of 10k Play Store apps in order to analyze the Android market. After downloading the data, I did the following steps to clean up and process the data.

There are 13 columns in the data, therefore, it is important to find out the meaning of each column in the first step. After understanding the meaning of each column, I deleted some column that would not be used. Secondly, I deleted some NA values and outliers to avoid negative results in future analysis. Thirdly, I handled some special symbols in the columns , consider doing log transaction and standardized processing to these factors.

Finally, I gained clean data containing 8 columns and 9367 observations. The explanation of 8 columns is shown in the following table 1.

Table 1: Explanation of columns

Column_names	**Explanation**
Category	The kind to which each app belongs
Rating	The grade given by someone who have downloaded and used the App
Reviews	The number of times a user has access to each app
Size	The amount of stored data required to download the app
Installs	The number of installations per app
Type	The payment way of Apps: Free or Paid
Type	Money spent by customers to download an App

Exploratory Data Analysis

In this data, there are many variables, such as 'Installs','Size', which range are so large that need us to make some transaction before doing exploratory analysis and model fitting. Therefore, in order to get some figures that are easier to understand and make the model fit better, I take log of these variables and do scatter plots to explore the relationship between these variables and price.

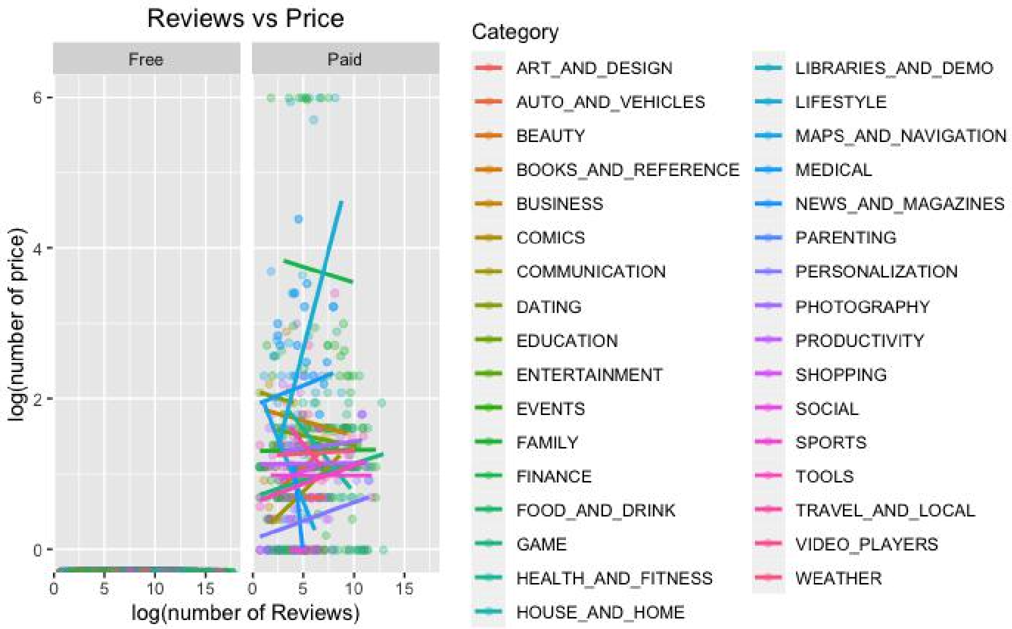


Figure 1 Review and Price

Figure 1 still shows the correlation between price and reviews. Prices and rating also differ in different categories. As can be seen from the graph, different categories have various intercepts and slopes. In one part of the category, price and rating present a positive correlation, while in the other part, a negative correlation is shown.

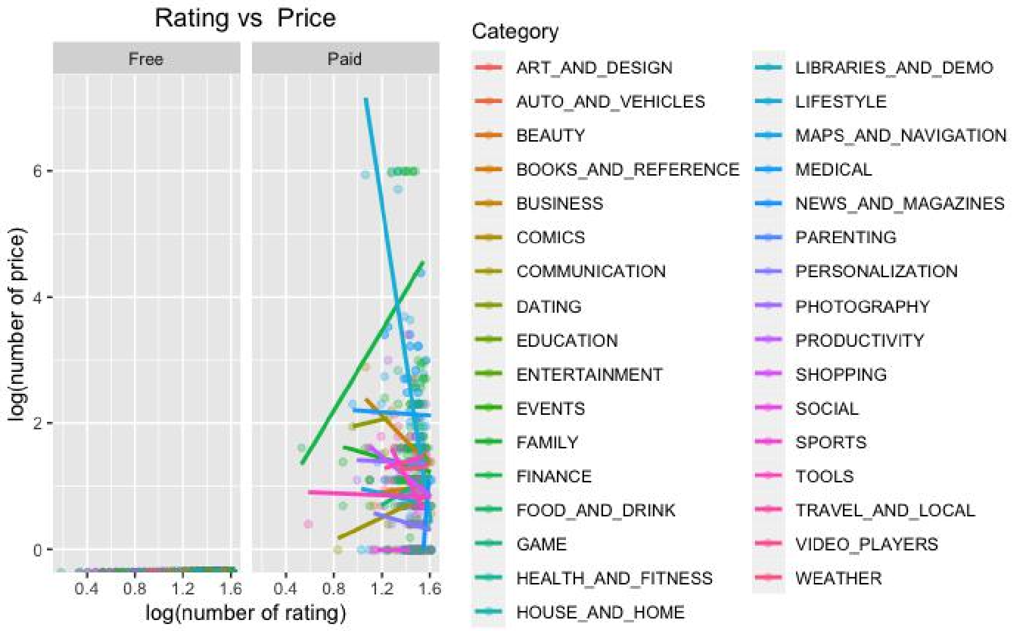


Figure 2 Rating and Price

Figure 2 illustrates the correlation between rating and price. It is clear that price and rating show a negative relationship in most category. In addition, it can be read from the result that in `Free` type that no matter how the rating grows, the price of App are always under zero and does not change significantly.

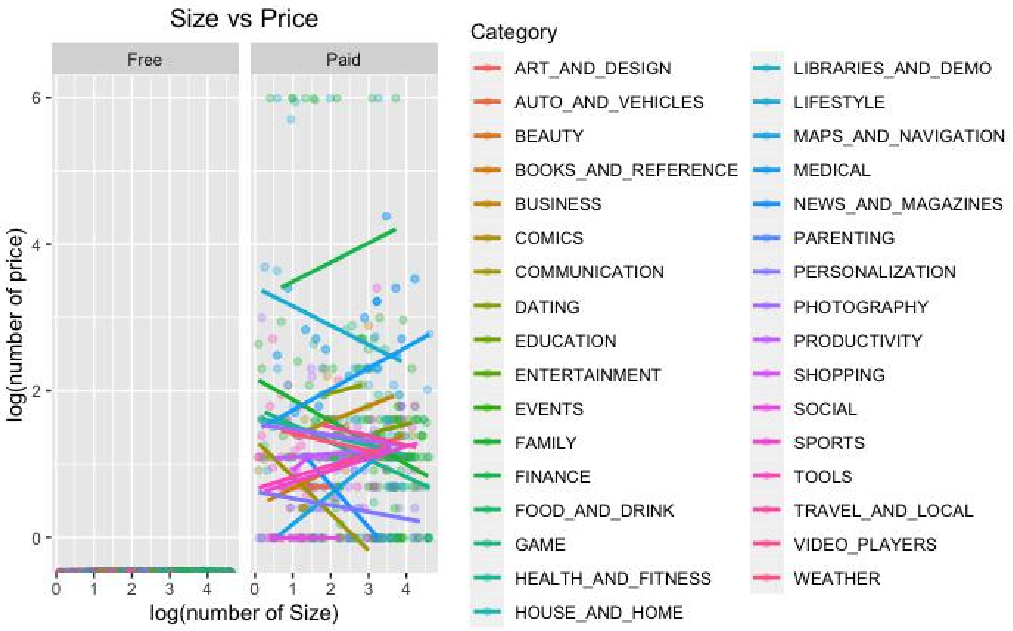


Figure 3 Size and Price

Figure 3 conveys an information which is similar with Figure 1 and 2 about 'Free' type.

Besides, in 'Paid' type, different categories also have different intercepts and slopes.

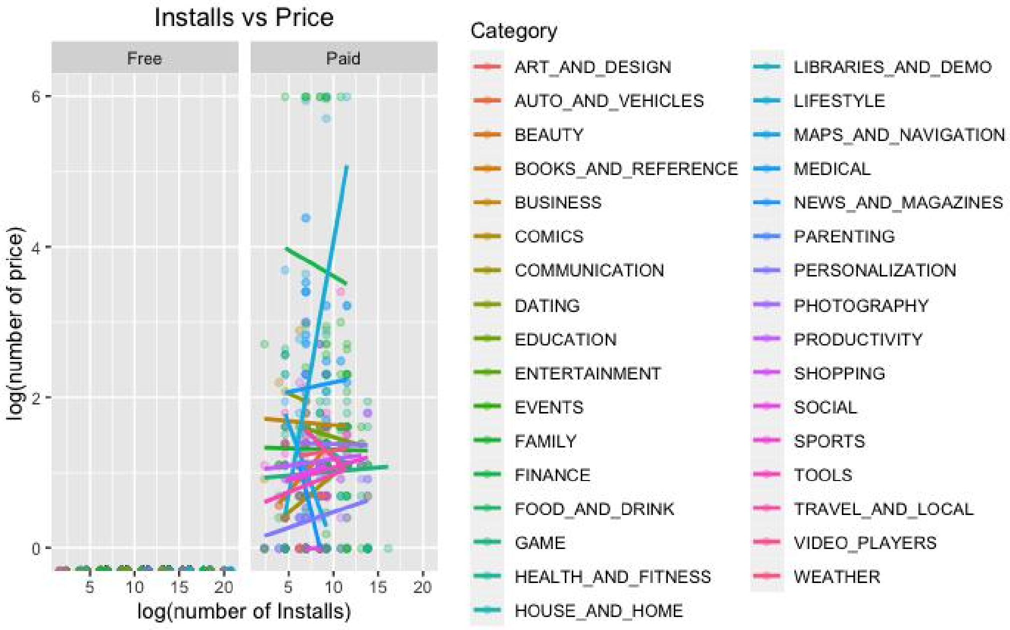


Figure 4   Size and Price

In Figure 4, dots in 'Paid' type look more concentrated, but it also shows that in different categories, the correlation between price and installs are different.

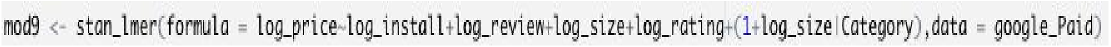
Above all, from Figure 1 to Figure 2, it is obvious that although the values of each variable changes, the App prices in each category change slightly or even never change. Therefore, I'll choose type to fit data in type 'Paid', for making the results more meaningful in later analysis.

Method

Model Fitting

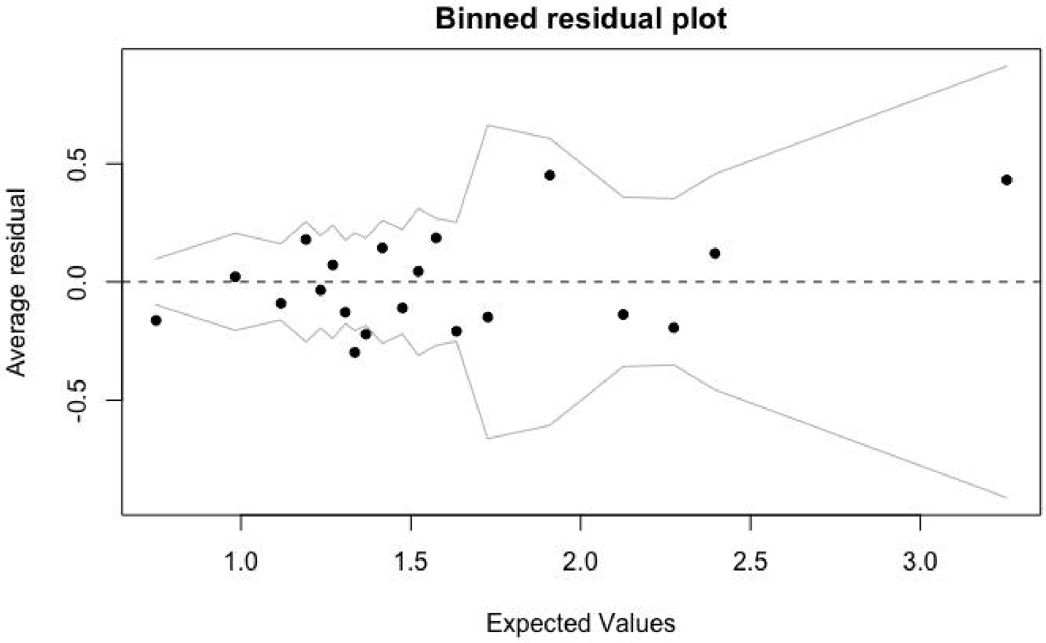
There are many methods can be chosen to fit model. When combing the real situation that this data contains more than 33 categories, I consider using multilevel model to fit the data. In the data cleaning and processing part, I have took log transaction to some predictors and the price. Besides, from the EDA results, it is clear that different categories have different intercepts and slopes, so it is reasonable to apply multilevel model.

In order to gain the best result, I tried many ways to fit the model, such as 'lmer', 'stan_lmer' and so on. Besides, trying to delete or add different variables in the model was also an important part during model fitting. After many attempts, the optimal model fitting method I selected is `stan_lmer`. I put the fitting results of some other models that I tried in the appendix. In five `stan_lmer` models that I tried to fit, the most efficient model is mod9, for its looic value is smaller. Here is the model function:



```
mod9 <- stan_lmer(formula = log_price~log_install+log_review+log_size+log_rating+(1+log_size|Category),data = google_Paid)
```

I also plotted a binned residual plot to check model validation. In figure 5, it shows that most points fall inside the confidence bands and there is not a distinctive pattern to the residuals. Therefore, the model can be considered a relatively good fit based on the message from binned residual plot.



Result

Model coefficient

Due to there are many categories in this data, here just an example for Education category, and we can conclude this formula:

$$log(price) = -2.37 - 0.05 \cdot log(install) + 0.07 \cdot log(review) - 0.06 \cdot log(size) - 0.41 \cdot log(rating)$$

From the coefficients result, it is clear that the parameters of three predictors are not totally large than 0, meaning that the impact of different predictors on price in each categories is not entirely positive. However,the coefficient in log_rating are all equal to -0.41.For example, in category `Education`, each 1% difference in review, the predicted difference in price is 7% when other variables do not change.

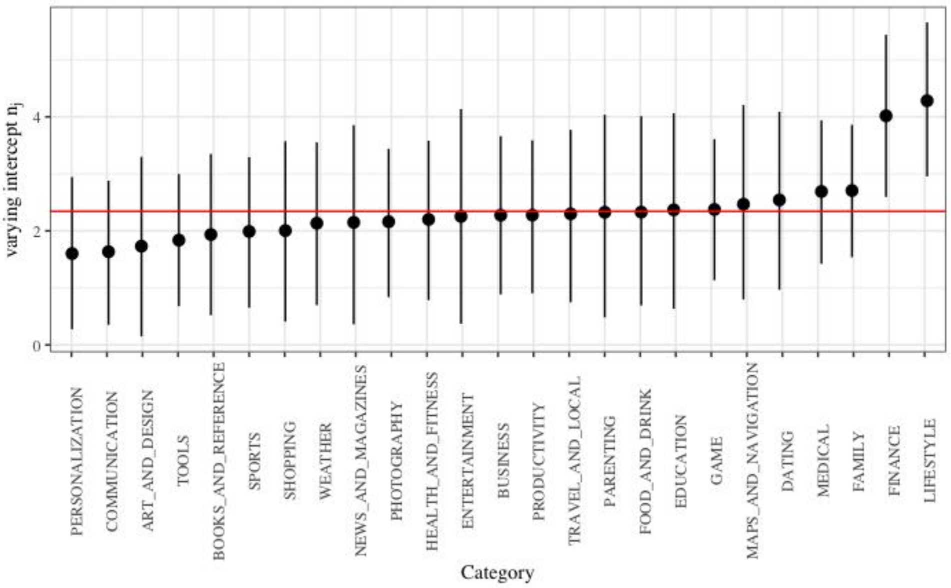


Figure6 Varying intercept

Figure 6 shows different intercept in different categories. The red line is the average intercept value in the data. In 33 categories, there are 11 categories' intercept under average line, and `Life Style` has the largest intercept. Besides, in these four predictors, only App review have positive influence on App price, and the rest three have negative impact on it except in `Medical`.

Discussion

This report analyzes the variables in the Google play store that may have various effects on the price of Apps and their relationship to Apps price. According to the fitting results of the model, almost every predictor has negative effect on price in each category. Only in `Medical`, review shows a positive impact on App price.

However, there are some limitations in this reports. Firstly, this report only taking some external factors that may affect the App itself into account, but ignoring some of the App internal characters, which are also affecting its price. For example, the cost of the app, functional requirements, the target group of App and so on. Therefore, in further analysis, I plan to add more predictors into the model so that I can gain better result. In addition, there are 33 categories in the data, and some categories have common characters, so the other important step that I will do is combine them into the same group to analyze. Secondly, in the EDA part, I plotted 4 pictures to find the relationship of each predictor and price. However, in model fitting part, I only consider one predictor, which is size, with price. It is no doubt that I need to fit more models to make correspondent to my EDA part.

Citation

1.https://www.kaggle.com/lava18/google-play-store-apps.
2.http://ritsokiguess.site/docs/2019/06/25/going-to-the-loo-using-stan-for-model-comparison/
3.https://cran.r-project.org/web/packages/loo/vignettes/loo2-example.html
4.https://stats.stackexchange.com/questions/99274/interpreting-a-binned-residual-plot-in-logistic-regression
5.https://cran.r-project.org/web/packages/loo/vignettes/loo2-example.html
6.https://mc-stan.org/users/documentation/case-studies/tutorial_rstanarm.html

Appendix

More EDA

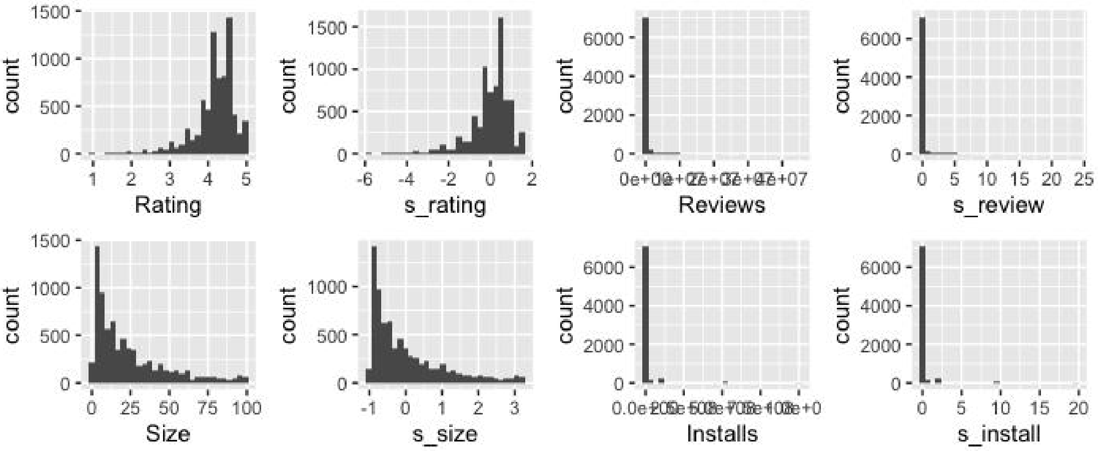


Figure7 Distribution of each predictor

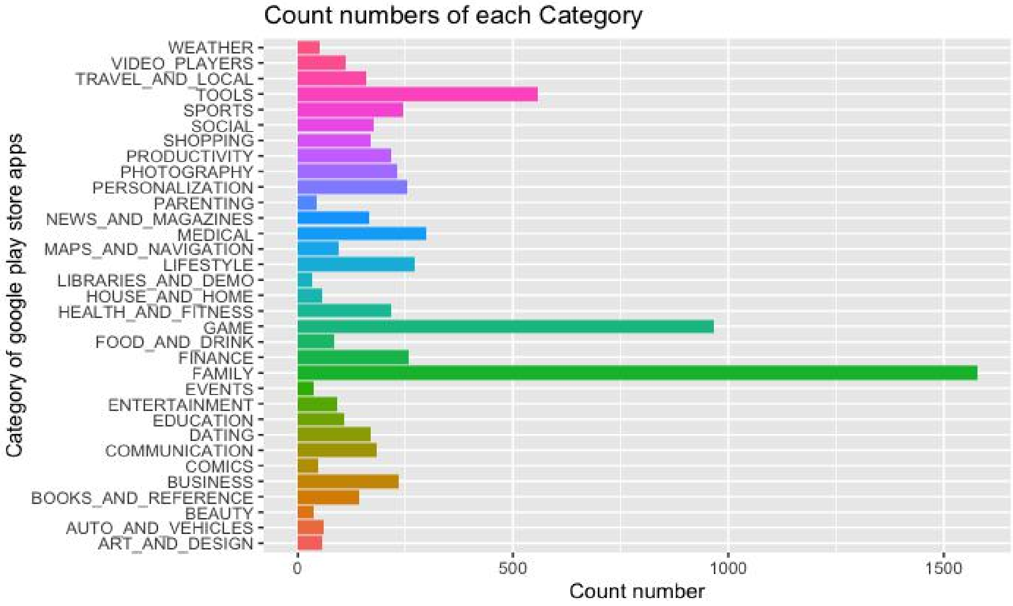


Figure8 Count numbers of each category

Model selection
Try stan_lmer

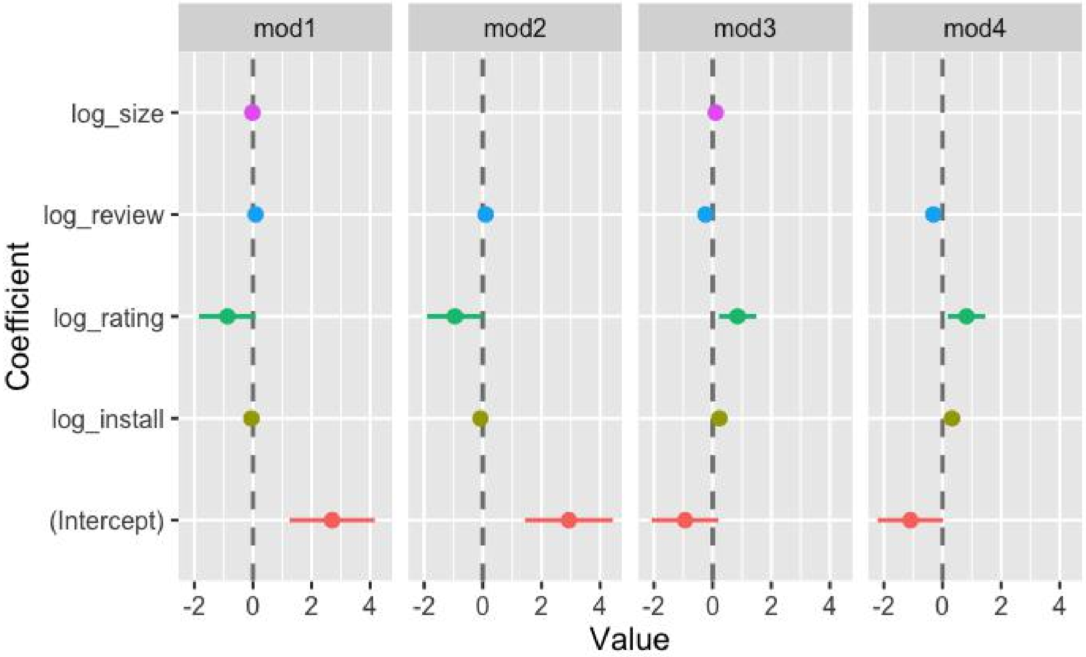


Figure8 Coefficient of each model

Model Validation

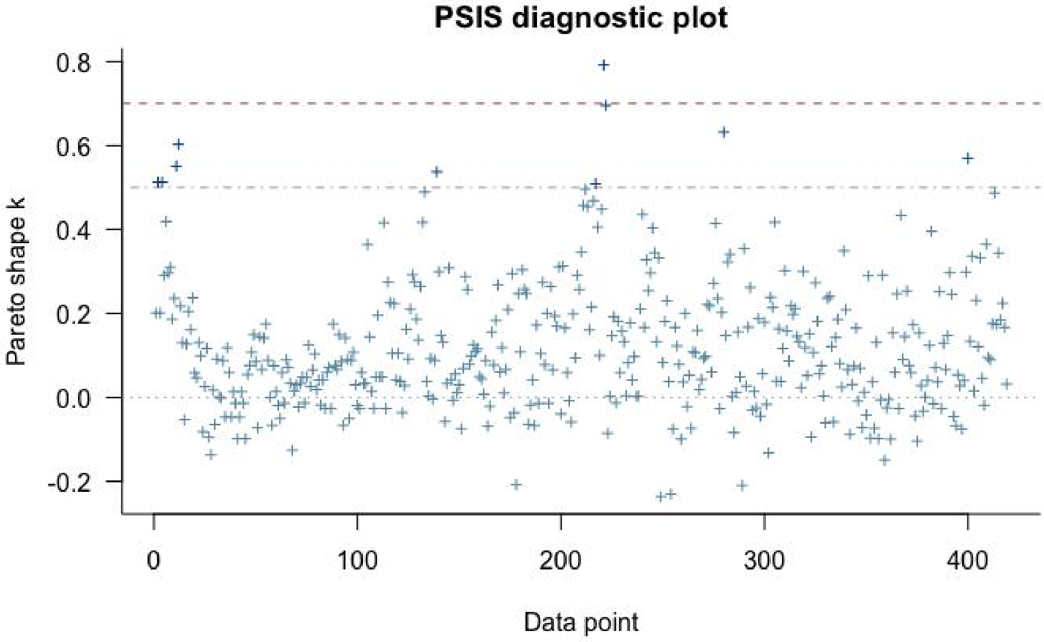


Figure9 PSIS diagnostic plot

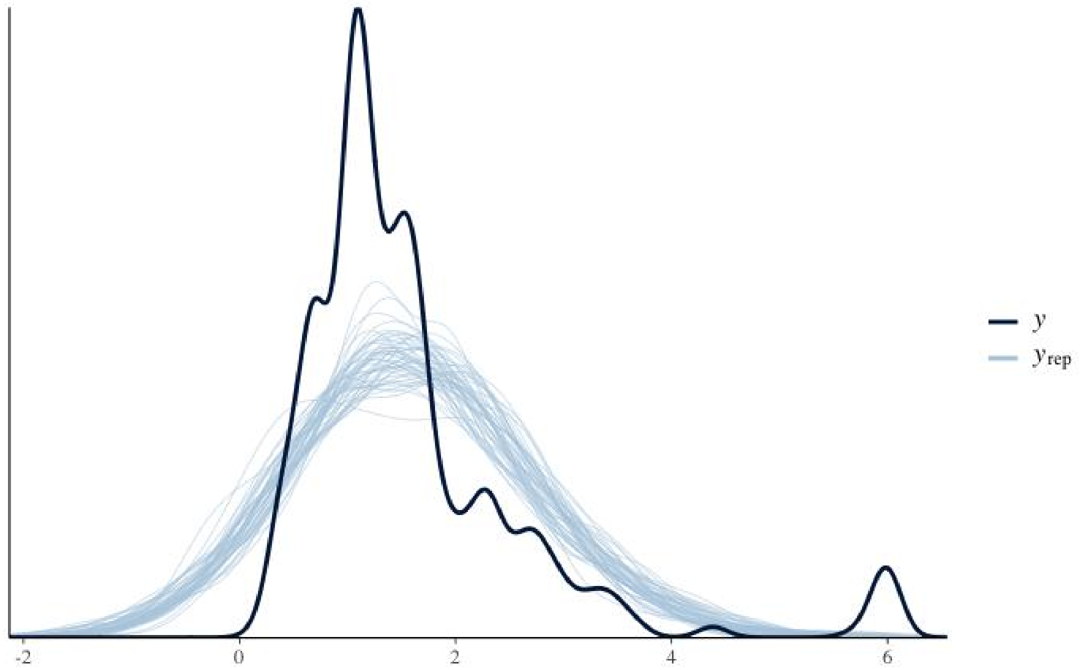


Figure10 PP-check model plot

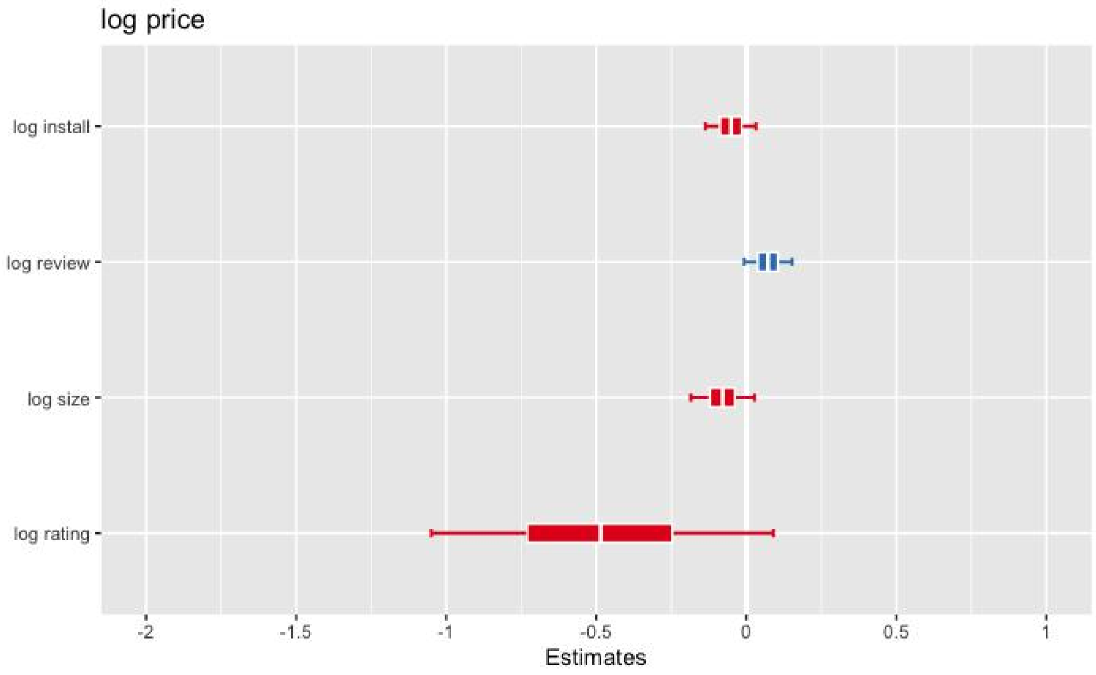
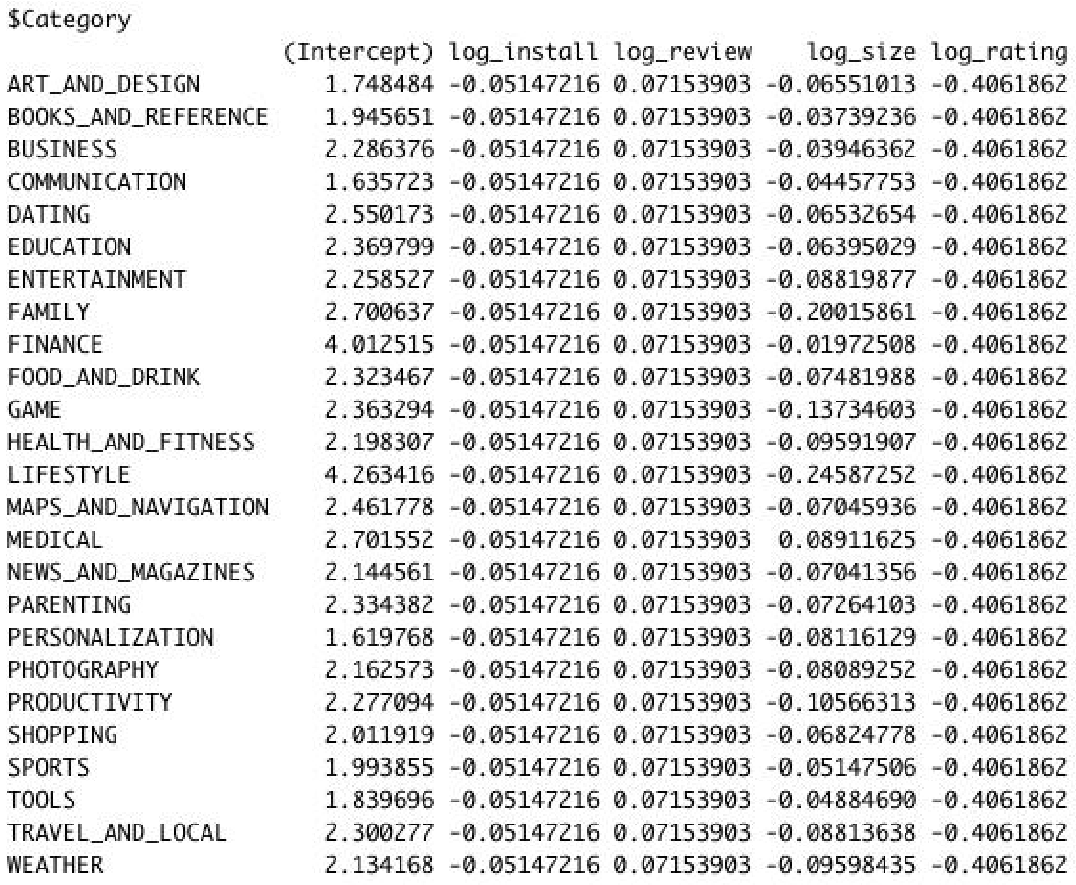


Figure11 Random effect plot

Model Coefficient



```
$Category
                    (Intercept) log_install log_review    log_size log_rating
ART_AND_DESIGN         1.748484 -0.05147216 0.07153903 -0.06551013 -0.4061862
BOOKS_AND_REFERENCE    1.945651 -0.05147216 0.07153903 -0.03739236 -0.4061862
BUSINESS               2.286376 -0.05147216 0.07153903 -0.03946362 -0.4061862
COMMUNICATION          1.635723 -0.05147216 0.07153903 -0.04457753 -0.4061862
DATING                 2.550173 -0.05147216 0.07153903 -0.06532654 -0.4061862
EDUCATION              2.369799 -0.05147216 0.07153903 -0.06395029 -0.4061862
ENTERTAINMENT          2.258527 -0.05147216 0.07153903 -0.08819877 -0.4061862
FAMILY                 2.700637 -0.05147216 0.07153903 -0.20015861 -0.4061862
FINANCE                4.012515 -0.05147216 0.07153903 -0.01972508 -0.4061862
FOOD_AND_DRINK         2.323467 -0.05147216 0.07153903 -0.07481988 -0.4061862
GAME                   2.363294 -0.05147216 0.07153903 -0.13734603 -0.4061862
HEALTH_AND_FITNESS     2.198307 -0.05147216 0.07153903 -0.09591907 -0.4061862
LIFESTYLE              4.263416 -0.05147216 0.07153903 -0.24587252 -0.4061862
MAPS_AND_NAVIGATION    2.461778 -0.05147216 0.07153903 -0.07045936 -0.4061862
MEDICAL                2.701552 -0.05147216 0.07153903  0.08911625 -0.4061862
NEWS_AND_MAGAZINES     2.144561 -0.05147216 0.07153903 -0.07041356 -0.4061862
PARENTING              2.334382 -0.05147216 0.07153903 -0.07264103 -0.4061862
PERSONALIZATION        1.619768 -0.05147216 0.07153903 -0.08116129 -0.4061862
PHOTOGRAPHY            2.162573 -0.05147216 0.07153903 -0.08089252 -0.4061862
PRODUCTIVITY           2.277094 -0.05147216 0.07153903 -0.10566313 -0.4061862
SHOPPING               2.011919 -0.05147216 0.07153903 -0.06824778 -0.4061862
SPORTS                 1.993855 -0.05147216 0.07153903 -0.05147506 -0.4061862
TOOLS                  1.839696 -0.05147216 0.07153903 -0.04884690 -0.4061862
TRAVEL_AND_LOCAL       2.300277 -0.05147216 0.07153903 -0.08813638 -0.4061862
WEATHER                2.134168 -0.05147216 0.07153903 -0.09598435 -0.4061862
```