Report

Abstract

With the development of science and society, mobile phone has become an indispensable part in our daily life. This project is about Chinese phone users. Particularly, I pay more attention to Xiaomi, a phone brand in China, since this company is an industry leader. I use data from Kaggle to do Exploratory Data Analysis and fit 2 regression models. Through results, male is more likely to choose Xiaomi than female and young people like Xiaomi more than older group.

Introduction

The whole project is around these questions and the yellow one is a main question.

1. Which brand has the largest market share?
2. What brands do people in different ages like?
3. Which app is most popular?
4. Distribution of Xiaomi users in China.
5. The influence of age scale and gender on the choosing Xiaomi.

Data contains 6 datasets including users' information and applications' information.
User information: gender, age, brand, device name.
Application information: app_id, labels and tags of each app.

Method

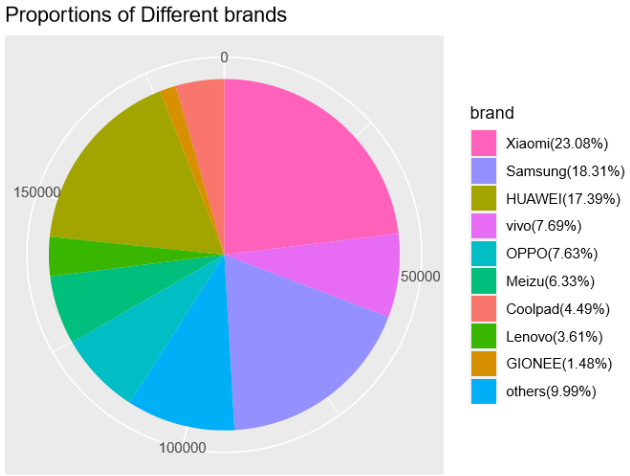
I mentioned five questions at the beginning. The first 4 questions can be answered by EDA. I state details in appendix. In this part, I will focus on the fifth question: The influence of age scale and gender on the choosing Xiaomi.

Data cleaning

I clean the data and join some datasets as preparation.

EDA

I use ggplot() to draw a brand pie to show proportion of Xiaomi.



I answer the first question according to this brand pie:

Xiaomi has the largest market share.

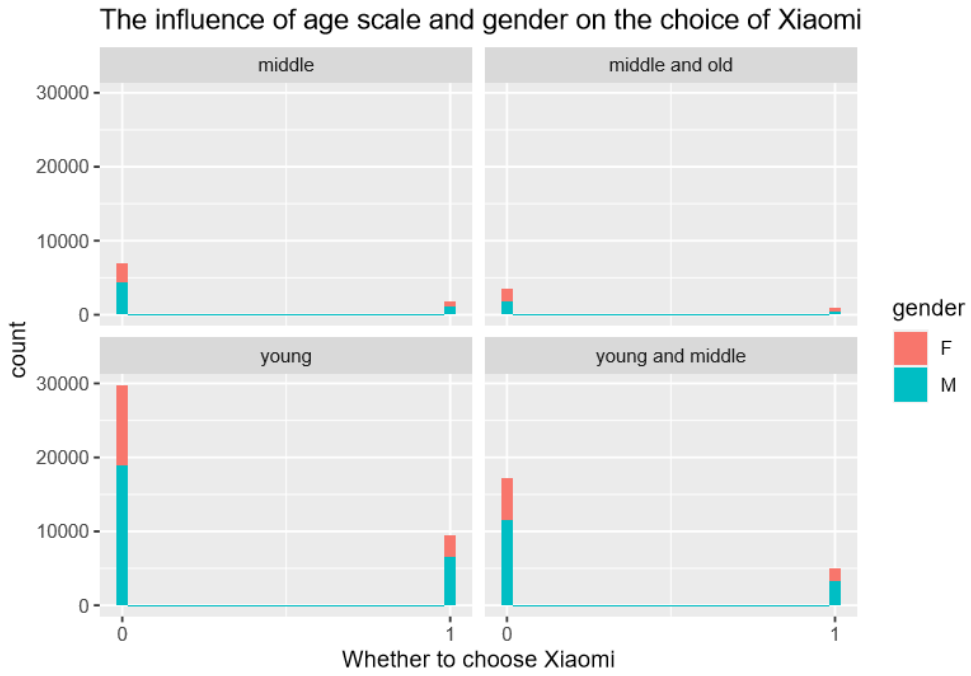
Then, according to age scale of Chinese people, I add a column to show each user' age scale. There are 4 age scales: 1-29 years old is young; 30-39 is young and middle; 40-49 is middle; over 50 is middle and old.

device_id	gender	age	phone_brand	device_mode l	age_scale
8842779817574750000	M	35	Xiaomi	MI 2	young and middle
9217427614282480000	F	27	Samsung	Galaxy S4	young

Since I want to explore the situation of choosing Xiaomi, I add a column to show whether the device is Xiaomi. 1 means choosing and 0 means not.

device_id	gender	age	phone_brand	device_model	age_scale	choose
8842779817574750000	M	35	Xiaomi	MI 2	young and middle	1
9217427614282480000	F	27	Samsung	Galaxy S4	young	0

Next, I use facet() to plot bars to show the situation of people in different ages choosing Xiaomi.



From these 4 pictures, I find in those people who choose Xiaomi, there are more males than females in each age scale. But there are more men in total than women.

Different to age scale, I calculate proportion of Xiaomi from 1 to 100 year-old. And I use it to draw a line chart and smooth each point. The proportion increases in young age scale, slightly decreases in the second age scale, holds steady in the third age scale and decreases in the forth age scale. But data from people who are very young or very old are way less than other age scale. The proportion will be extreme.

Model

I fit two models “M5” and “M6”. M5 is a multilevel logistic regression model and M6 is a logistic regression model.

M5=stan_glmer(choose~gender+(1|age_scale), data=Xiaomi, family=binomial(link = "logit"))

M6=stan_glm(choose~gender+age,data =Xiaomi,family = binomial(link = "logit"))

I set link = "logit" because in the data, outcome is binary and final output is probability.

```
M5$coefficients
```

```
##                                 (Intercept)
##                                -1.299257162
##                                     genderM
##                                 0.089602973
##           b[(Intercept) age_scale:middle]
##                                -0.117141416
##   b[(Intercept) age_scale:middle_and_old]
##                                 0.005931916
##            b[(Intercept) age_scale:young]
##                                 0.098536111
## b[(Intercept) age_scale:young_and_middle]
##                                 0.006795577
```

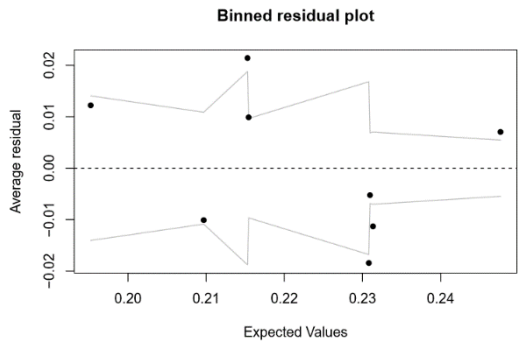
```
M6$coefficients
```

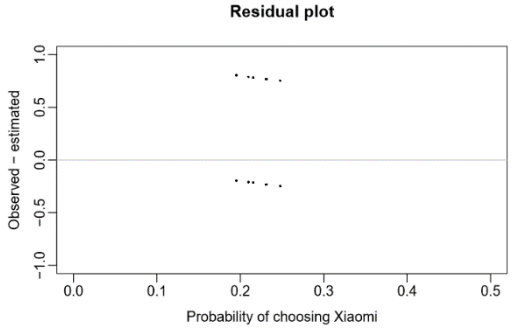
```
##  (Intercept)      genderM          age
## -1.080814060  0.086704253 -0.005554542
```

Validation (details in appendix)

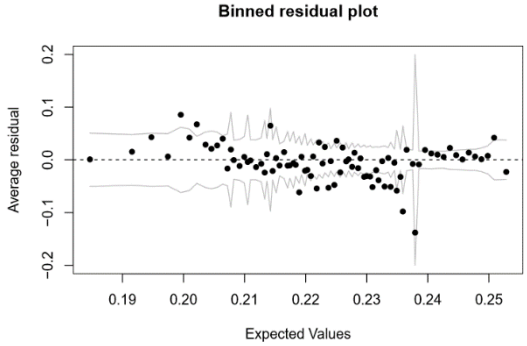
In this part, I use arm and package to draw binned residual plot and rstanarm package to check model. There 2 models' residual plot look similar. But binned residual plots are different. First, they have different ordinate order of magnitude. Second, in M5'binned residual plot, the points are at the edge of the confidence interval. In M6'binned residual plot, most points are in the confidence interval.

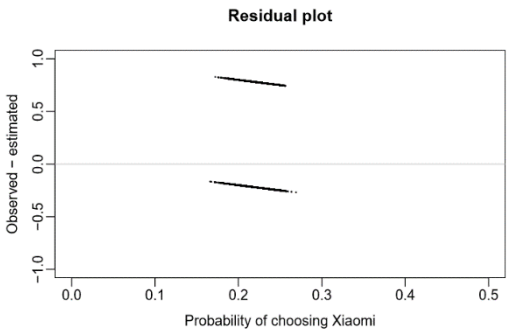
M5:





M6:





Inference

In this part, I interpret each model.

Interpretation of M5:

Intercept -1.299 means on average, female in young age scale have invlogit(-1.299+0.099)=23.15% probability of choosing Xiaomi. Sd is way less than mean, which means the sample size is enough to estimate.

Gender coefficient is 0.090 which means a difference of gender corresponds to no more than an 0.090/4=2.24% positive difference in the probability of choosing Xiaomi. Generally speaking, comparing to female, male are more likely to choose Xiaomi.

We have 4 groups and we can see σ(age_scale) is estimated at 0.18. Dividing by 4 tells us that the age scales differed by approximately ±4.5% on the probability scale. b[(Intercept) age_scale:young] is 0.099 which gives a positive effect on choosing Xiaomi. b[(Intercept) age_scale:young_and_middle] and b[(Intercept) age_scale:middle_and_old] also have positive effect. But effect degree is very light. b[(Intercept) age_scale:middle] is -0.117 which has a negative effect on choosing Xiaomi. These 4 groups' result match the age-proportion plot.

Interpretation of M6:

Intercept -1.081 means on average, female in 0-year-old have invlogit(-1.081)=25.33% probability of choosing Xiaomi. Gender coefficient is 0.087 which means a difference of gender corresponds to no more than an 0.087/4=2.18% positive difference in the probability of choosing Xiaomi. Generally speaking, comparing to

female, male are more likely to choose Xiaomi. Age coefficient is -0.006 which means a difference of 1 in age category corresponds to no more than an 0.15% negative difference in the probability of choosing Xiaomi.

Result

In this part, I answer the five questions mentioned above.

1.Which brand has the largest market share?

No.1: Xiaomi. No.2: Samsung. No.3: HUAWEI.

2.What brands do people in different ages like?

In young age scale(1-29), top 5 brands are Xiaomi, Samsung, HUAWEI, VIVO and OPPO; in young and middle age scale(30-39), top 5 brands are Xiaomi, Samsung, HUAWEI, OPPO and VIVO; in middle age scale(40-49), top 5 brands are HUAWEI Samsung, Xiaomi, OPPO and Coolpad; in middle and old age scale(>=50), top 5 brands are Xiaomi, Samsung, HUAWEI, Coolpad and OPPO.

3.Which app is most popular?

The most download and the most active app is a parkour game developed by Tencent.

4.Distribution of Xiaomi users in China.

Xiaomi users are mainly concentrated in eastern and southeastern China.

5.The influence of age scale and gender on the choice of Xiaomi.

Generally speaking, male are more likely to choose Xiaomi than female. With the growth of age, probability of choosing Xiaomi will slightly decrease. If we devide age into 4 groups, young age scale(1-29), young and middle age scale(30-39), middle and old age scale(>=50) have positive effect on choosing Xiaomi. But the effect degree is light. Middle age scale(40-49) has negative effect on choosing Xiaomi.

Discussion

In the model, I have some concerns:

1. The coefficients are very small. This may be solved by changing link.
2. Although I have enough data, number of predictors is too small.
3. Another model may also fit this situation: zero inflated model. Because according to brand pie, Xiaomi accounts for 23.08%, which means there are a number of “0” in the “choose” column.
4. These models took me a day to run and I saved them as RData files. When trying get linpred to draw my model, I failed because the file is too big. Sampling is a way.

Bibliography

[1] GH book

《Data Analysis Using Regression and Multilevel》

[2] Count() and tally()
https://dplyr.tidyverse.org/reference/count.html#arguments
[3] How to plot a percentage plot with ggplot2
https://sebastiansauer.github.io/percentage_plot_ggplot2_V2/

Appendix

In the main body part, I answer question 1 and question 5 in detail. This part will answer the remaining questions.

Data

I find data on kaggle. Everyone can download the data. Here is the link: https://www.kaggle.com/chinapage/china-mobile-user-gemographics

My data are in six datasets:

1. App_events:

event_id	app_id	is_installed	is_active
2	5927333115845830000	1	1
2	-5720078949152200000	1	0
2	-1633887856876570000	1	0

2. App_labels:

app_id	label_id
7324884708820020000	251
-4494216993218550000	251
6058196446775230000	406

3. Label_categories:

label_id	category
155	psychology
156	science
157	Information

4. Events:

event_id	device_id	timestamp	longitude	latitude
1	29182687948017100	2016/5/1 0:55:25	121.38	31.24
2	-6401643145415150000	2016/5/1 0:54:12	103.65	30.97

5. Gender_age_train:

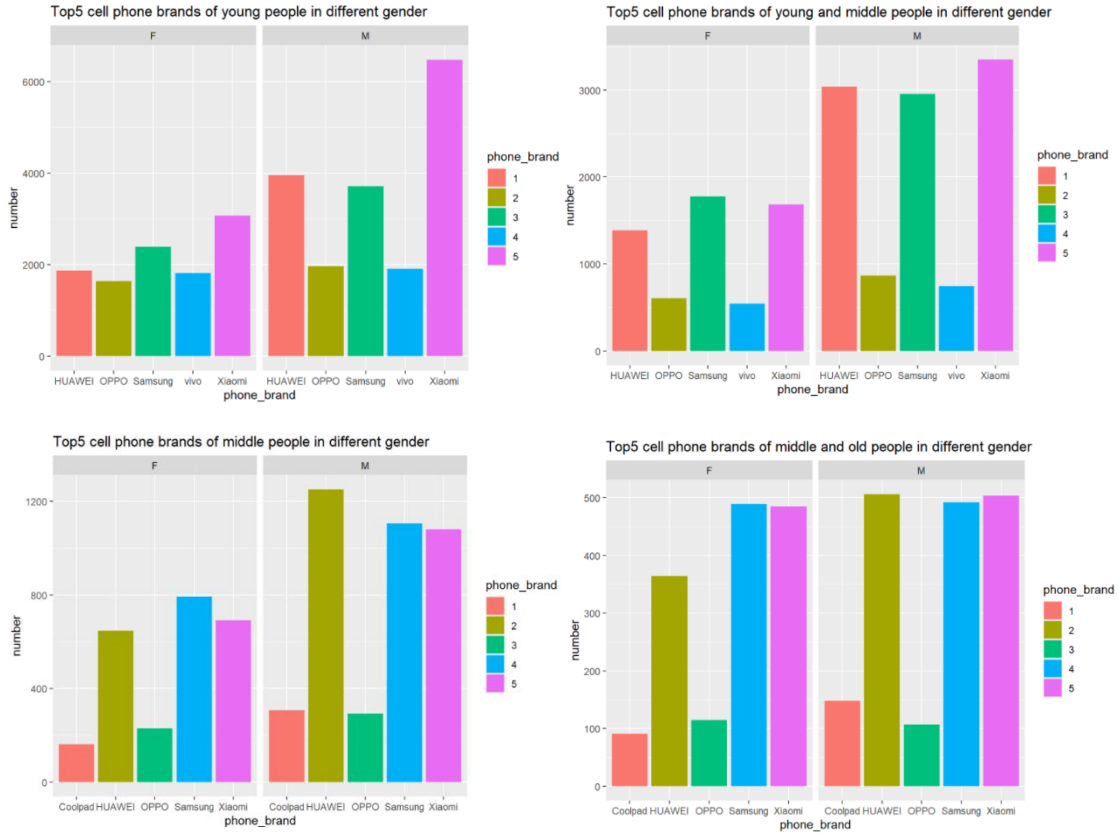
device_id	gender	age
-8076087639492060000	M	35
-2897161552818060000	M	35
-8260683887967670000	M	35

6. Phone_brand:

device_id	phone_brand	device_model
-8890648629457970000	小米	红米
1277779817574750000	小米	MI 2
5137427614288100000	三星	Galaxy S4

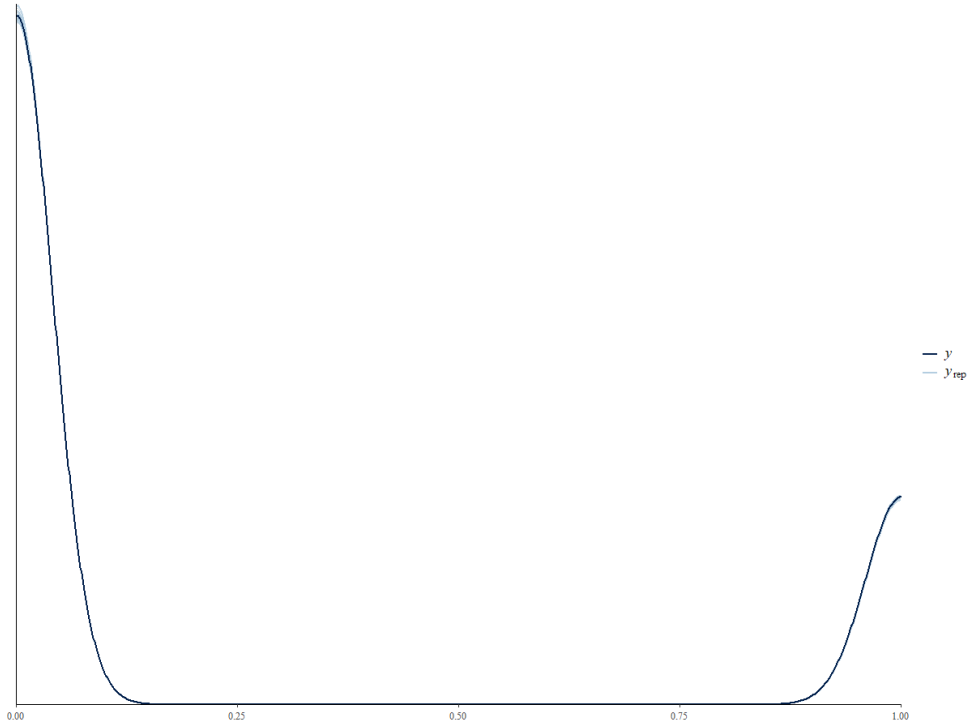
After cleaning, I still have around 760000 data about users.

Brand-number bar in different age scale

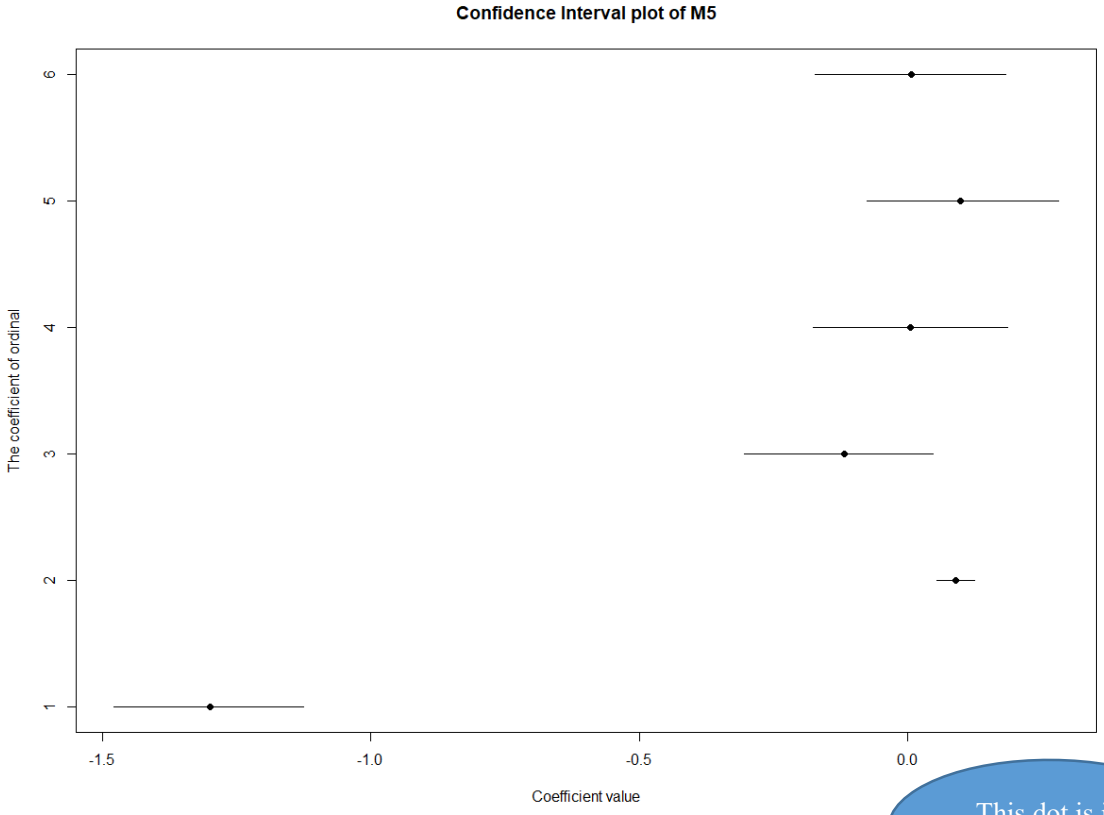


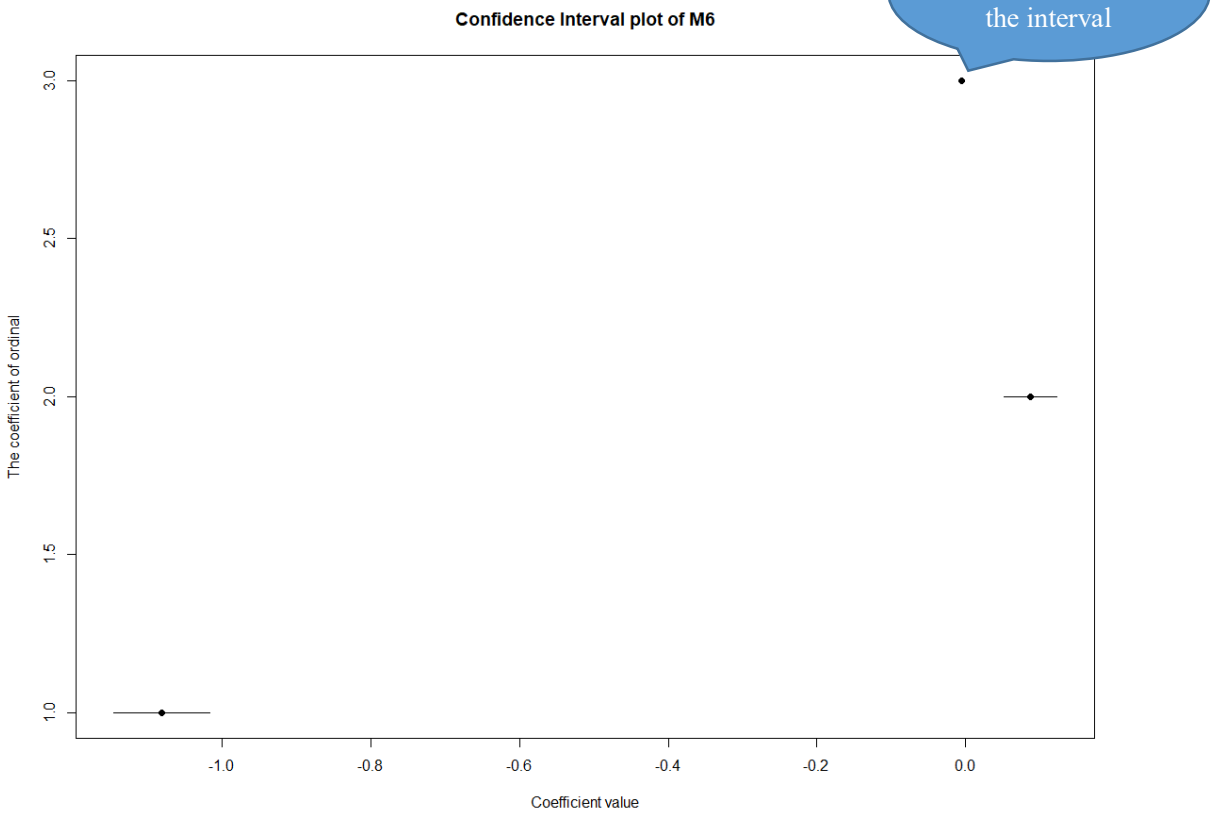
Validation (supplement)

I also use pp_check() function and results of these 2 models are almost exactly the same. These blue lines are very close to the black line, which is good.



Here are Confidence interval plots (95%) of 2 models.





This dot is in the interval

Favorite app

To answer this question, I define “most popular” by finding an app with most downloads or max activity. Every app has several tags to describe itself. Since I don’t have name of each app, I use tags to infer which app it is and order them in a table.

According to my result, a parkour game developed by Tencent has both most downloads or max activity.

Map

In this part, because the total Xiaomi users are around 170000, I sample 20000 Xiaomi users to draw a distribution plot.