Report of MA678 Midterm Project

Li Yuyang

December 6, 2021

Abstract

As the developing of mobile landscape, proportion of mobile device usage is increasing, so it's necessary for each tech company to improve the quality of their products in app market. However, there are a mass of different apps in one genre, then how to make your app seems attractive in app market and get more users is the key point for each company to consider. When an user wanna download an app from the market, the comments of the app is definitely one of the most important part for him to decide. This report are consisted of 4 parts: Introduction, Method, Result and Discussion.

Introduction

Comments of app are proportional to downloads, which means that an app has more comments represent it has more users in general. Meanwhile, the comments can be kind of feedback for companies to improve their product.

In this project I use the data from ios platform, which holds about 47% of the smartphone market to analyse the relationship between comments value of different genres and some factors may influence the comments by using multilevel model. For example, the price may be a common factor when a user viewing store to choose an app to download and comment.

Method

Data source

Link of the dataset from Kaggle:Mobile App Store (7200 apps)

Data clean

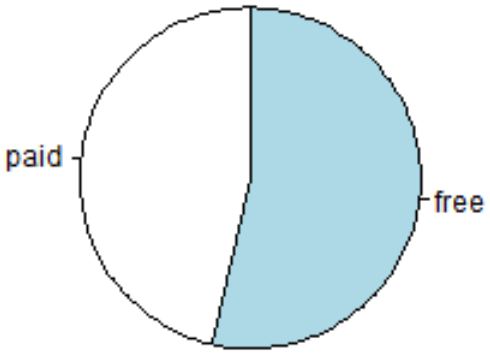
Firstly, I selected the columns that will be useful for my analysis, and then I cleaned the data by filtering by the genres, there are twenty-two primary genres in the data, and other rows are deleted because of containing some unreadable words. Also, I removed NA value and some meaningless symbols. For the price column, it looks strange for most of the value are end with 0.99, so I rounded up all the decimals in price.

variables interpretat

column names	explanation
id	App ID
track_name	App Name
price	Price amount
rating_count_tot	User Rating counts (for all version)
user_rating	Average User Rating value (for all version)
cont_rating	Content Rating
prime_genre	Primary Genre
lang.num	Number of supported languages
sup_devices.num	Number of supporting devices

EDA

Firstly, I separated the price into paid group and free group, and the difference of proportion of the two group is not remarkable, so I thought the price can be retained as one factor.



Also, most of the comments number in different genres is concentrated in 250, however, the comments number of game genre is larger a lot than others, so I took log of this variable to avoid long tail, the same treatment were taken for other similar variables.

Lastly, I tried different variables in data and plot the relation with comments number, and I found that some variables like number of supporting devices and content rating didn't have a consistent trend with contents number. If I use these variables to fit model may lose statistical significant. After filtering I decided to take user rating, number of supporting devices and price for fitting model.

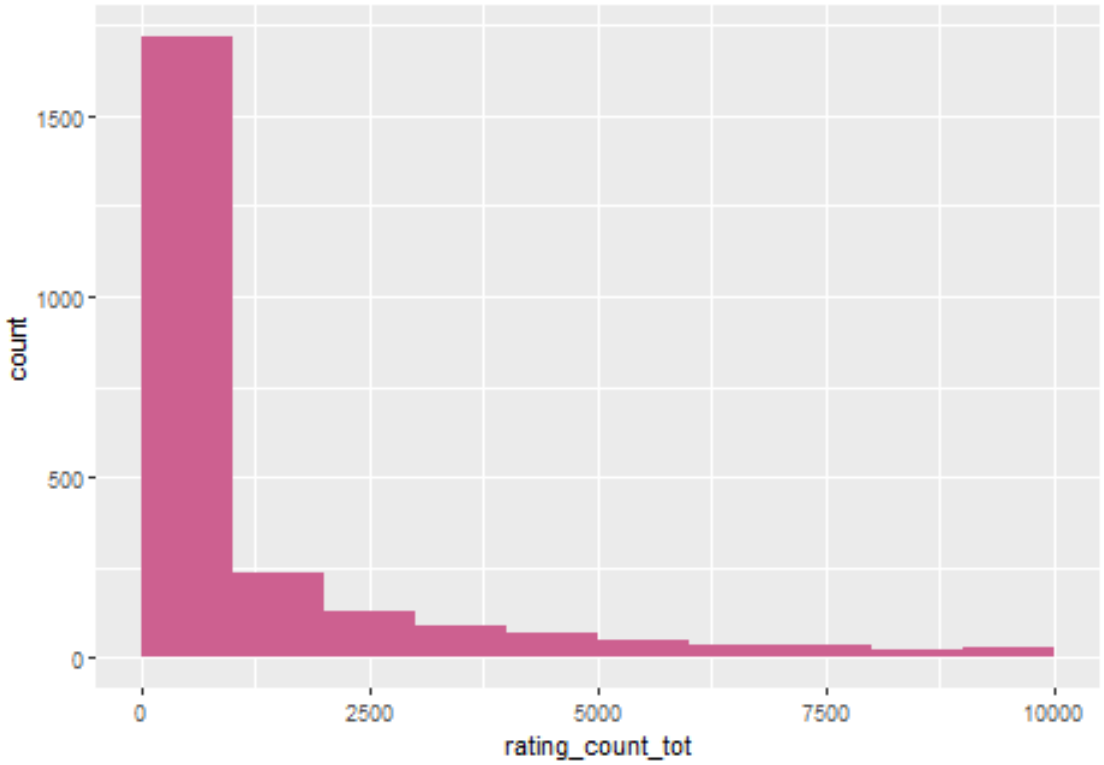


Figure 1: histogram of number of comments

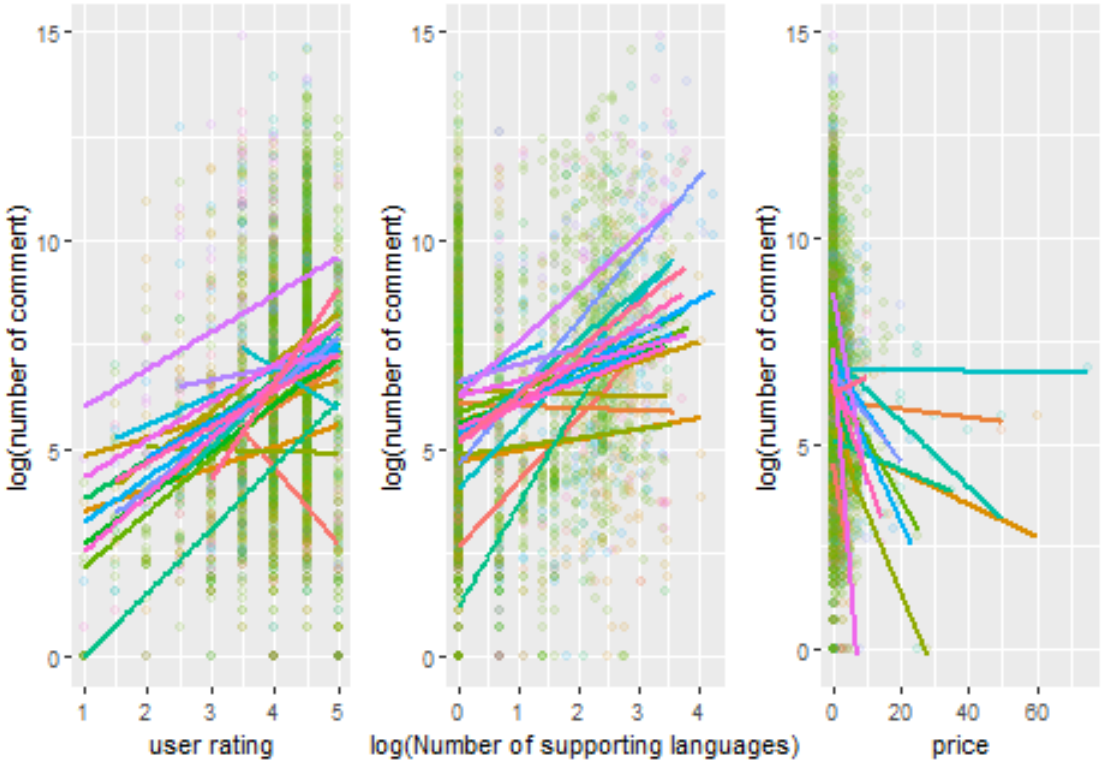


Figure 2: Three variables vs log(number of comment)

Model fitting

As I see in the EDA part, though the trend of variables is similar, the difference between genres is significant, so I chose multilevel model for fitting model, and adding random effects in both intercept and slope to make them different among genres. Besides, I took log of the rating counts and number of supporting devices because these two numbers have a relatively large scale.

I tried two similar multilevel models, the difference between these is making the effects of price different. In the first model I made both of intercepts and slopes of price different in genres, and in the second model I removed the random effects for slope of price. Comparing for two models, I found that the second one fitted better and had a smaller p-value, so I decided to take the second model as my final outcomes.

```
#Final model
finalmodel <- lmer(log(rating_count_tot)~user_rating+log(lang.num)+
                  price+(1+user_rating|prime_genre)+
                    (1+log(lang.num)|prime_genre),app1)
```

Result

Coefficients

The basic fixed effects of the model are showed as follow:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.88719	0.42137	14.24388	6.852	7.22e-06
user_rating	0.73732	0.12068	13.09960	6.110	3.60e-05
log(lang.num)	0.66253	0.07223	13.97388	9.172	2.74e-07
price	-0.10116	0.01343	2642.36114	-7.530	6.93e-14

Coefficients of random effects of some standard genres as follow:

prime_genre	(Intercept)	user_rating	log(lang.num)	price
Entertainment	4.3307592	0.49918081	0.6524462	-0.101162
Games	0.4978502	1.13147382	0.4412793	-0.101162
Book	7.6949329	-0.05578763	0.4416768	-0.101162
Food & Drink	4.3396180	0.49771943	0.4795743	-0.101162

Model Specific

Taking the game group as an example, the formula of fitting model is as follow:

$$log(ratingcount) = 0.4978502 + 1.13147382 \cdot userrating + 0.4412793 \cdot log(lang.num) + -0.101162 \cdot price$$

The coefficient of price in negative, it's common because when people viewing the app store, if the price of the app is too high, they may abandon the willing to download this app, and another possible reason is the number of deliberate praise comments will decrease when the app charges. And the remaining parameters is positive, for the rating, which means that a higher rating of the app, more comments it will receive, because users are willing to praise after good experience, the number of supporting languages has the same positive impact, in the model of game group,if one app supports kind of language increase by 1%, the comments value can increase by 0.44%.

Model Validation

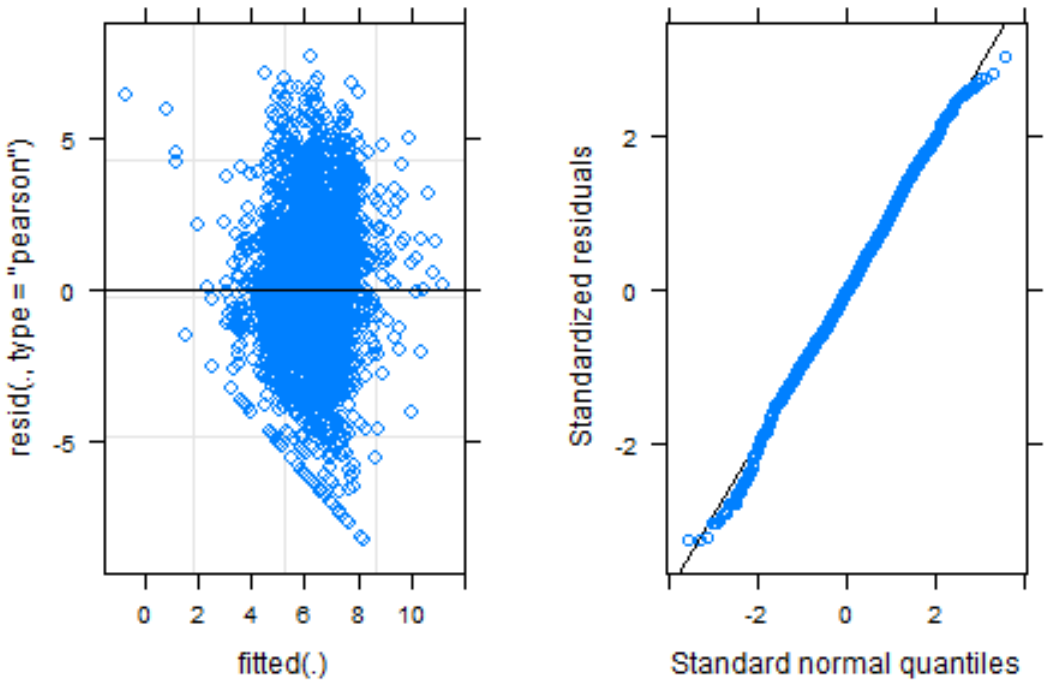


Figure 3: Residual plot and Q-Q plot of final model

In the residual plot we can see the residual points are evenly distributed on both sides of zero, it' a reasonable sign, also in the Q-Q plot most of the points located in the line, which means that the normality of model is good.

Discussion

The results showed before are reasonable, the mainly trend of influence of app rating and supporting language numbers on comments counts is positive, which represents if an app has a higher user rating or supports more languages, it will gain more user comments. However, the relation between price and comments number is negative, more expensive app has fewer comments. And all the trends are suitable for almost every genre in my data, which are consistent with the result in the EDA part.

There are still some weaknesses in my model. Firstly, I pretended the relation between comments number and downloads number is positive, more comments represent more downloads will make it easier to interpret outcomes, but this precondition doesn't have any data supporting. Also, I didn't distinguish the nature of comments are good or bad, so it will cover up something because the influence of predictors may be opposite with the result of model for negative reviews. Besides, I believe there are still other factors to effect comments, but I just chose three have remarkable and consistent influence. Further research should be considered beside my project.

In the future I can analyse the connection between number of comments and downloads, also I can add some other important variables into model to make it more accurate.

Appendix

EDA

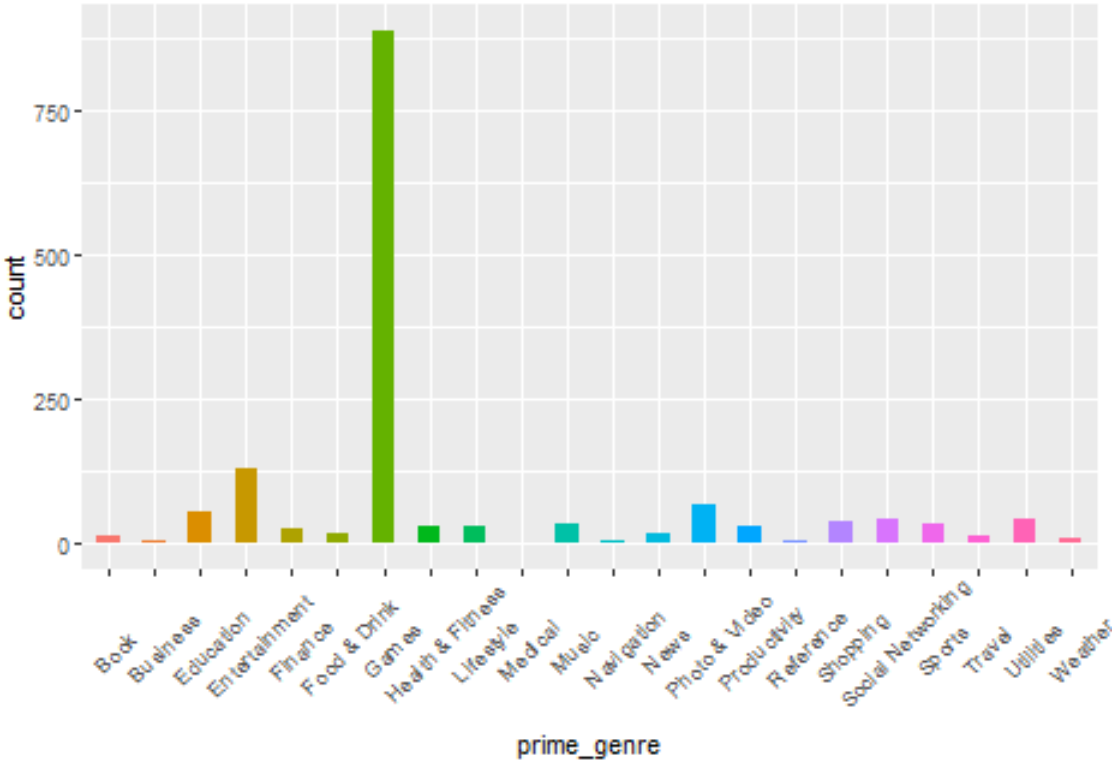


Figure 4: plot of the count of different genres in free subset

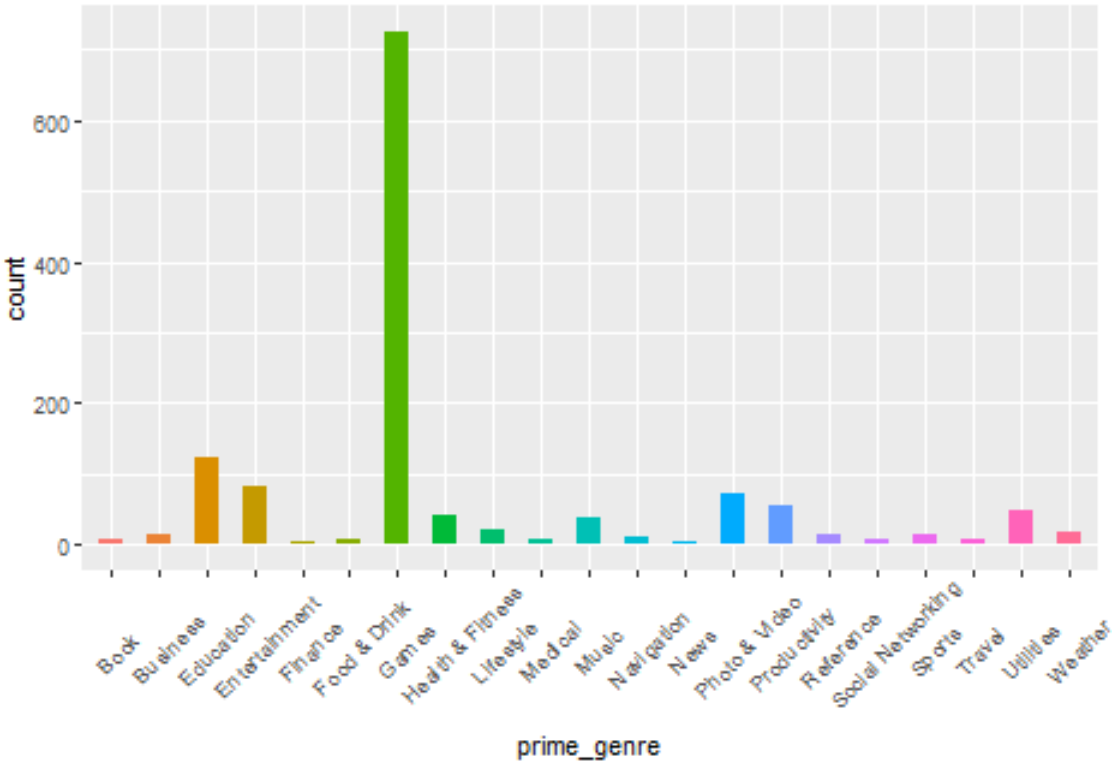


Figure 5: plot of the count of different genres in paid subset

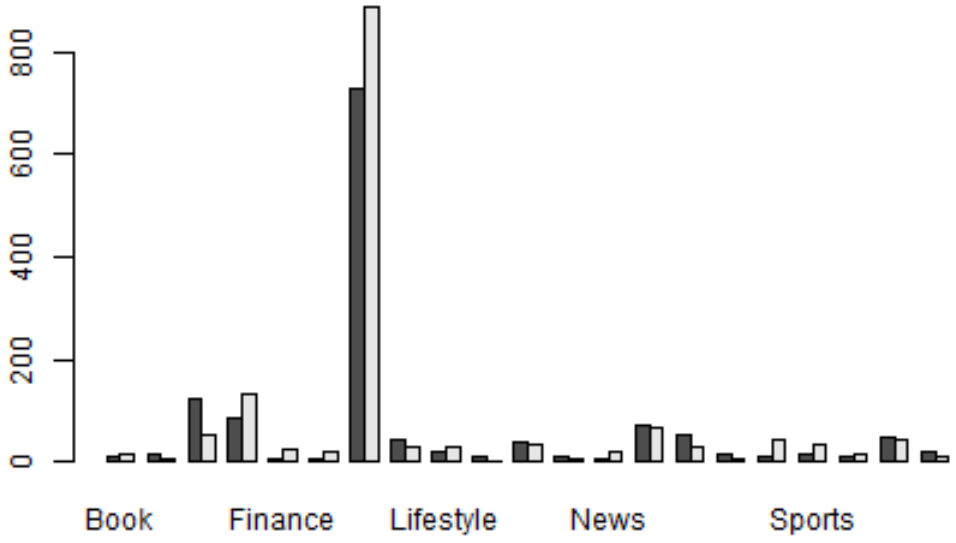


Figure 6: plot for comparing free and paid group

Full Results

Random effects of model

```
## $prime_genre
##                    (Intercept) user_rating  (Intercept) log(lang.num)
## Book                2.40387302 -0.79310628 -0.313840856   -0.22085747
## Business           -0.48121049  0.15876507 -0.244869668   -0.17232076
## Education           0.67372333 -0.22228054 -0.555226818   -0.39072666
## Entertainment       0.72178618 -0.23813785 -0.014335249   -0.01008806
## Finance            -0.51705017  0.17058960  0.078951139    0.05555984
## Food & Drink        0.72621556 -0.23959923 -0.259988040   -0.18295993
## Games              -1.19466833  0.39415516 -0.314405680   -0.22125495
## Health & Fitness   -0.42683870  0.14082626 -0.188123983   -0.13238744
## Lifestyle          -0.26870440  0.08865325  0.007842914    0.00551925
## Medical             0.12722607 -0.04197551 -0.075111472   -0.05285778
## Music              -0.55872036  0.18433779  0.183101028    0.12885266
## Navigation          0.70838955 -0.23371792  0.338854196    0.23845997
## News               -0.32067688  0.10580045  0.271470909    0.19104070
## Photo & Video      -0.24157117  0.07970122 -0.155164081   -0.10919275
## Productivity        0.05441153 -0.01795191  0.044145240    0.03106608
## Reference           0.06895845 -0.02275136  0.260963081    0.18364609
## Shopping           -0.07940060  0.02619652  0.079041864    0.05562369
## Social Networking -0.24183586  0.07978855  0.647155753    0.45541929
## Sports             -0.80887597  0.26687126  0.204327479    0.14379023
## Travel             -0.27470895  0.09063432 -0.111391198   -0.07838870
## Utilities           0.27721177 -0.09146007  0.064374680    0.04530203
## Weather            -0.34753357  0.11466124  0.052228762    0.03675465
##
## with conditional variances for "prime_genre"
```

Fixed effects of model

```
##   (Intercept)   user_rating log(lang.num)         price
##     2.8871868     0.7373187     0.6625342    -0.1011620
```

Coefficients of model

```
## $prime_genre
##                    (Intercept) user_rating log(lang.num)     price
## Book                 7.6949329 -0.05578763     0.4416768 -0.101162
## Business             1.9247659  0.89608372     0.4902135 -0.101162
## Education            4.2346335  0.51503811     0.2718076 -0.101162
## Entertainment        4.3307592  0.49918081     0.6524462 -0.101162
## Finance              1.8530865  0.90790826     0.7180941 -0.101162
## Food & Drink         4.3396180  0.49771943     0.4795743 -0.101162
## Games                0.4978502  1.13147382     0.4412793 -0.101162
## Health & Fitness     2.0335094  0.87814492     0.5301468 -0.101162
## Lifestyle            2.3497780  0.82597190     0.6680535 -0.101162
## Medical              3.1416390  0.69534315     0.6096765 -0.101162
## Music                1.7697461  0.92165644     0.7913869 -0.101162
## Navigation           4.3039659  0.50360073     0.9009942 -0.101162
## News                 2.2458331  0.84311911     0.8535749 -0.101162
```

```
## Photo & Video       2.4040445  0.81701987     0.5533415 -0.101162
## Productivity        2.9960099  0.71936674     0.6936003 -0.101162
## Reference           3.0251037  0.71456730     0.8461803 -0.101162
## Shopping            2.7283856  0.76351518     0.7181579 -0.101162
## Social Networking   2.4035151  0.81710720     1.1179535 -0.101162
## Sports              1.2694349  1.00418991     0.8063245 -0.101162
## Travel              2.3377689  0.82795297     0.5841455 -0.101162
## Utilities           3.4416104  0.64585858     0.7078363 -0.101162
## Weather             2.1921197  0.85197990     0.6992889 -0.101162
##
## attr(,"class")
## [1] "coef.mer"
```