Report of MA678 Midterm Project

Zhe Yu

2020/12/1

Abstract

TED is an American online platform providing different topics of talks videos with slogan 'Ideas worth spreading'. But some videos have millions of hits while some only have a few views. Here rises a problem: Why do some talks are so popular on TED? To address this problem, I use some factors related with talks and build multilevel model. The model shows that the variables all have positive impact on number of comments and is slightly different between categories. This report are consisted 5 main parts: Introduction, Method, Result and Discussion.

Introduction

Since TED is famous all over the word, speakers with different native language will give speech on TED and some popular talks are translated into different languages. So each talk has its own features and some similarities with other talks as well, such as duration, number of views, the native language of this talks. And some features may lead the talks to be more out-standing. For example, talks in English may draw more attention for it is an American website, majority of viewers have English as their native languages. Besides, Viewers would give comments on the talks they are interested in. Talks with more comments always indicate that they are worth discussing or recommending.

Therefore I use multilevel models to see what and how factors may influence the number of comments of talks. Before that, I clean the data and combine some information collected from YouTube for same talk video.

Method

Data Cleaning and Processing

The main data set is published on Kaggle: TED-Ultimate Dataset. And I also found a data set on Kaggle: TEDTalks-transcript have some additional transcript of these videos.

Firstly, for some columns are 'dictionary' type so I need to clean them up, removing symbols that are useless. Secondly, I count the number of available languages, mutate it as a new column; Thirdly, I merged two data set into one and select the columns that I need; Finally, I transfered dates into POSIXct type by using Lubridate package. Here are some explanations of columns:

column names	explanation
title	The title of talk
num_lang	The number of available languages translations
comments	The number of comments
duration_ted	How long is the video in seconds
view_ted	The number of views
categories	The video was under which category on YouTube
english	Indicate the native language is English or not

Then, I got the cleaned data with 2165 observations and 32 variables. But I will choose only several variables to use.

Exploratory Data Analysis

For `views`, `comments` and some variables all have a large range and also if we see the density plots of them, there will be a long tale. Therefore, in order to make the plot more easy to read, I take log of these variables and draw some scatter plots to see if there is correlation between some variables with comments, since my question is how some factors effect the number of comments.

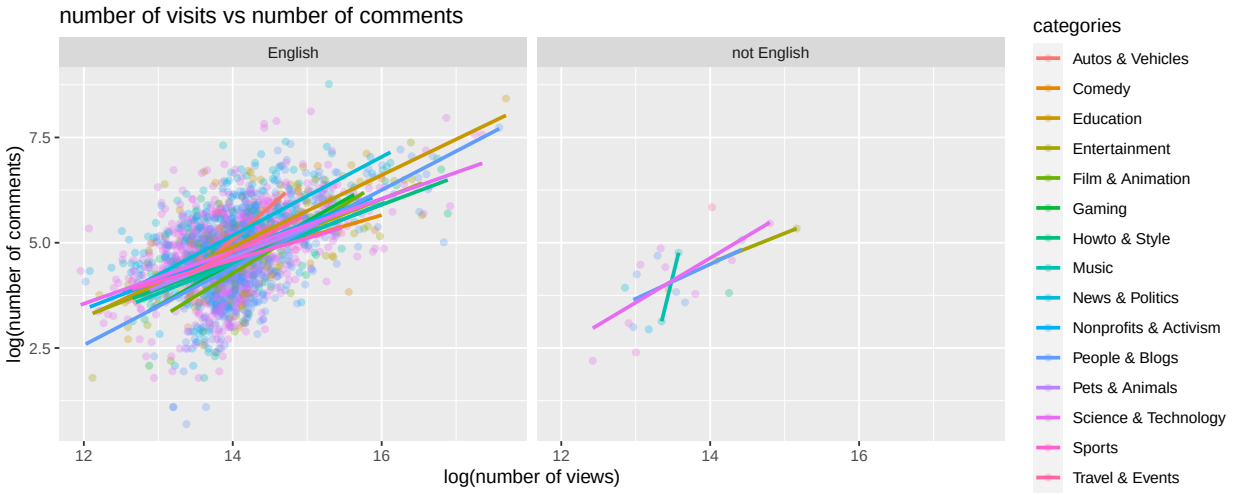

Figure 1: Data was separate into English talks and not English talks. Different colors represent different categories.talke log of both comments and views to make it more easy to read.

Figure 1 show the relationship between number of views and comments. But this plot indicates that the speaker of majority of videos is speaking English, so I will consider only to fit model on English talks. And in the English group, it is obviously that with different contents the slope and intercept are slightly different, but the trend is almost the same, pointing to top-right.

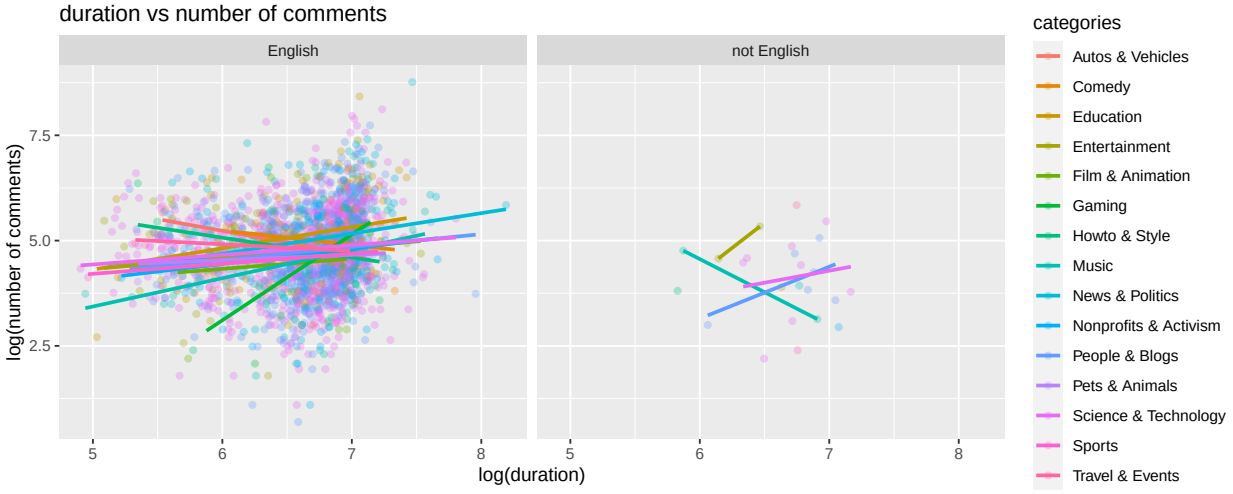

Figure 2: correlation between duration and number of visits.

Figure 2 is the same situation for English and not English group as above shows. So I will focus on the English plot. It also shows that in different categories the correlation between duration and number of comments are not same. And it is more clear than Figure 1 because some categories seems to have negative correlations but others have positive correlations.

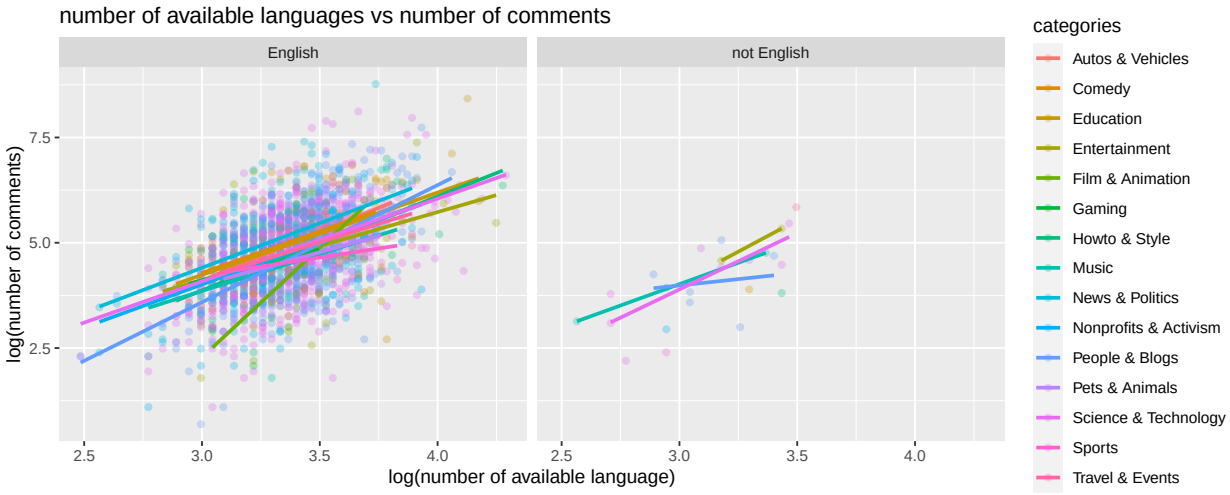

Figure 3: correlation between number of available languages and number of comments.

Just focus on the English group and see the regression lines on Figure 3, if the number of available language increases, the number of comments may increase. It makes sense for the more the number of available language, the more people from different countries will see the talks and leave their comments. Also different categories have different slopes and intercepts.

Model Fitting

Considering different categories, I will use multilevel model to fit the data. And for three continues variables-duration, number of language and duration-in order to match the plots in EDA, I take log of them as predictors and also take log of number of comments as outcome. Since from EDA it is clear that talks in different categories have different correlation with variables, so I use varying slope and varying intercept in multilevel models. Besides, just as I mentioned before, English talks will only be taken into account. Below is the function:

```
model <- lmer(log_comments~log_duration+log_views+log_numlang+(1+log_duration|categories)
              +(1+log_views|categories)+(1+log_numlang|categories),talks)
```

And to see the fixed effects below, all variables are significant at alpha = 0.05 level.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-12.93	0.67	16.75	-19.43	6.34e-13 ***
log_duration	0.79	0.04	24.53	17.57	2.10e-15 ***
log_views	0.33	0.04	12.21	8.09	2.99e-06 ***
log_numlang	2.33	0.13	10.98	18.56	1.21e-09 ***

Result

Model Coefficients

Just take some example here, for Entertainment category, we can conclude this formula:

$$log(comments) = -12.70 + 0.77 \cdot log(duration) + 0.32 \cdot log(views) + 2.15 \cdot log(numlang)$$

Because duration is counted in seconds, number of views always can be higher than 1000, and there are always more than 20 available languages, but number of comments always much less than number of views, so it makes sense that intercept in the formula is negative. But for log(duration), log(views) and log(numlang) can not be 0, it is hard to interpret. And all the parameters of three predictors are all bigger than 0, which means they all have positive impact on number of comments. For each 1% difference in duration, the predicted difference in comments is 0.77%. And the same for number of views and number of available languages.

For different categories, the influence of each predictor is always not the same. For News & Politics, the Intercept is smaller than others, I think it may because when things related to politics, it is serious so many people will withhold their opinion. For Science & Technology, number of language have bigger influence than other categories, it may because we always need to learn the advanced technology from other countries and language is the biggest problem. So the number of languages is much more important in talks about Science and Technology.

	(Intercept)	log_duration	log_views	log_numlang
Entertainment	-12.70	0.77	0.32	2.15
News & Politics	-14.37	0.90	0.38	2.21
People & Blogs	-12.28	0.73	0.37	2.47
Science & Technology	-13.37	0.82	0.21	2.49

Model Validation

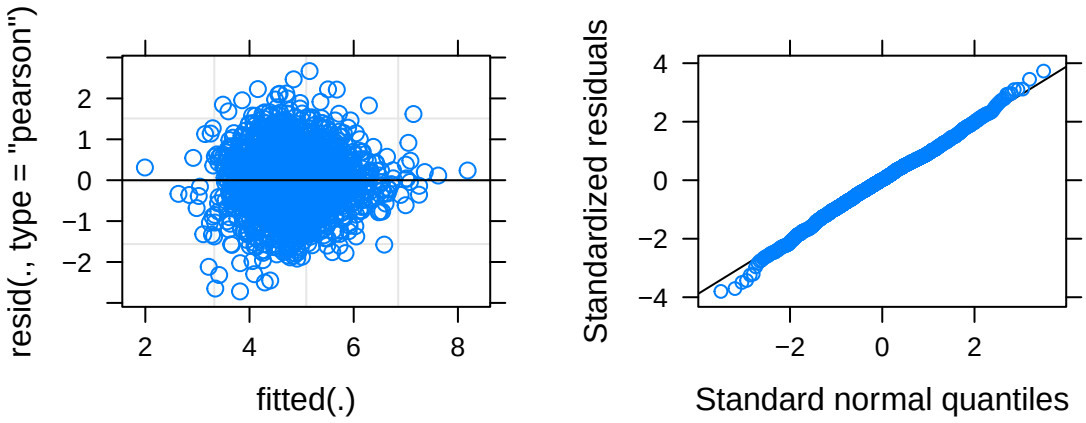

Figure 4: Residual plot and Q-Q plot.

From the Residual plots in Figure 4 we can see that the mean of residuals is almost 0, but dots reduce on the right and left of the plot since the number of samples are small there, so it makes sense. And for the Q-Q plot in Figure 4, majority dots are on the lines so the normality is good. Figure 5 shows that there are not obvious leverage point.

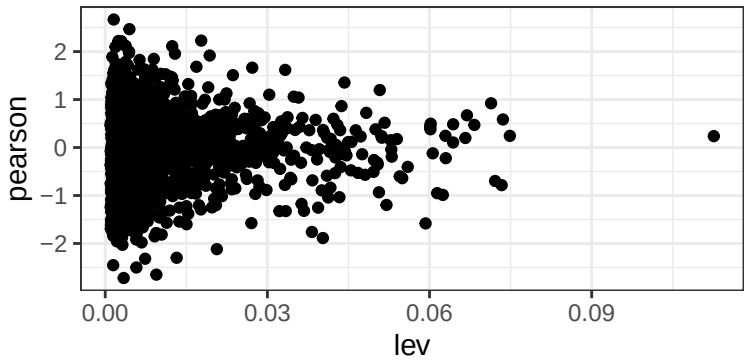

Figure 5: Residuals vs Leverage.

Discussion

The estimates are all reasonable in some extents. The longer the duration, the more information the speaker conveys, so people are more likely to have some perspectives; The larger the number of views, the more people will debate and have some comments; The more the number of available languages, the more people will have access to these talks, so people all over the world will talk together. Due to features of different categories, the different estimates of three predictors are also convincing.

The result of model almost matches well for different categories in EDA, but there are still some problems. In EDA part we see that some lines have negative slopes which means that the increasing in duration will make the number of comments decrease. But because for majority of categories, duration have a positive impact on number of comments, so although I use multilevel model with varying slope to fit the data, all the estimates of duration indicate that it has positive influence on comments for all categories.

I only use three variables to predict the outcome, but there are still some variables that may have big effects on comments, such as who is the speaker, the occupation of the speaker and the contents of talks. Actually there is a column contains the information of occupation, but it seems that the occupation in that column is too specific, so for further improvement, I will clean them up, and separate them into several categories, then add them into the model.

Citation

University of Wisconsin. Mixed Models: Diagnostics and Inference. https://www.ssc.wisc.edu/sscc/pubs/MM/MM_DiagInfer.html

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Rune Haubo Bojesen Christensen. lmerTest: Tests in Linear Mixed Effects Models. R package version 3.1.3. https://CRAN.R-project.org/package=lmerTest

Appendix

More EDA

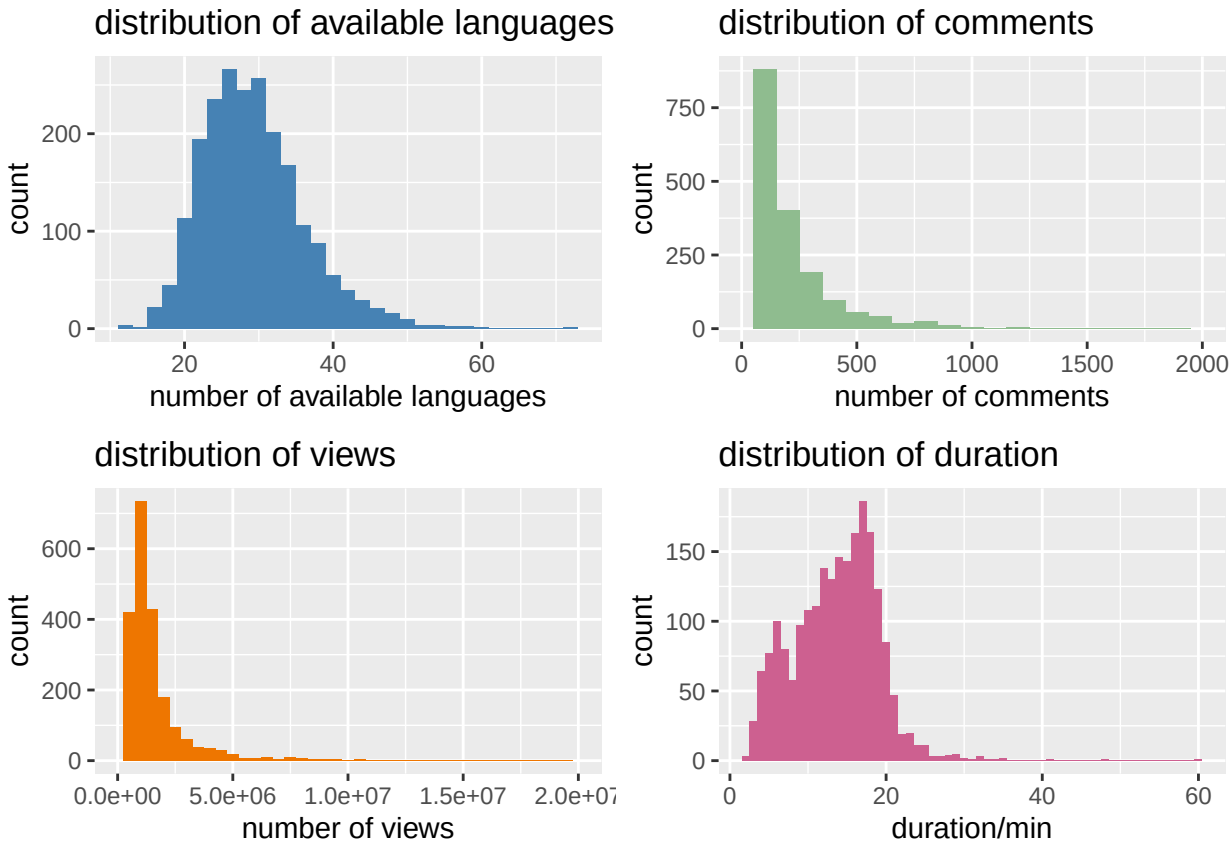

Figure 6: distribution plots for number of language, number of comments, number of views and duration

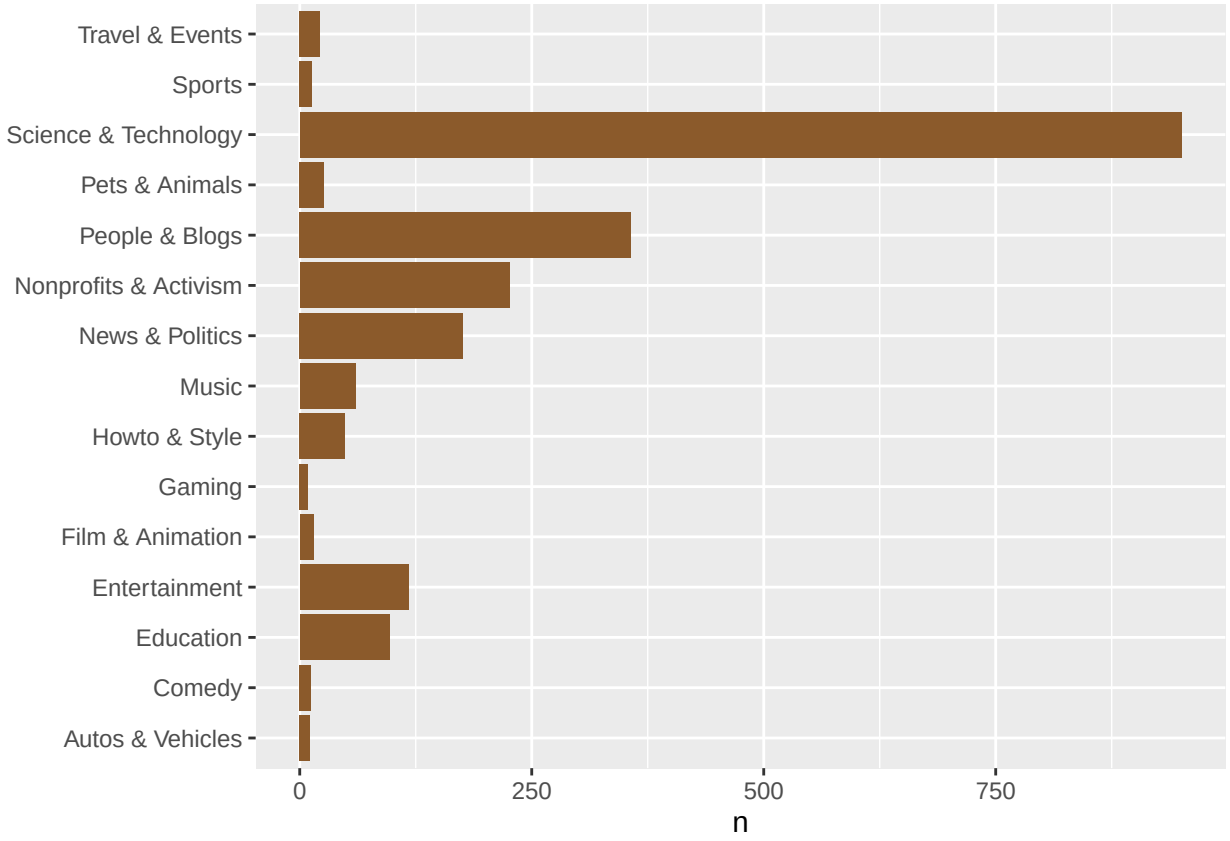

Figure 7: plots for count of category

Full Results

Random effects of model

```
## $categories
##                         (Intercept) log_duration (Intercept)    log_views
## Autos & Vehicles      -0.134017640  0.033987892  0.02889658 -0.002009319
## Comedy                 0.009442396 -0.002394663  0.12860642 -0.008942625
## Education             -0.114265010  0.028978475 -0.35702741  0.024825839
## Entertainment          0.078224255 -0.019838265  0.07890154 -0.005486405
## Film & Animation       0.180666704 -0.045818449 -0.48777959  0.033917670
## Gaming                 0.008540544 -0.002165947 -0.52931471  0.036805807
## Howto & Style          0.119506954 -0.030307871  0.43811323 -0.030464128
## Music                  0.174339364 -0.044213787  0.35775084 -0.024876143
## News & Politics       -0.480160733  0.121772410 -0.73278993  0.050954422
## Nonprofits & Activism -0.068411584  0.017349698  0.12465970 -0.008668191
## People & Blogs         0.216498568 -0.054905681 -0.54325022  0.037774810
## Pets & Animals         0.098584859 -0.025001869 -0.39815327  0.027685519
## Science & Technology  -0.147410406  0.037384399  1.72854775 -0.120194270
## Sports                 0.090144703 -0.022861382  0.14012937 -0.009743872
## Travel & Events       -0.031682973  0.008035043  0.02270971 -0.001579116
##                         (Intercept)  log_numlang
## Autos & Vehicles       0.044418307 -0.012179593
## Comedy                 0.051956166 -0.014246494
## Education              0.282574197 -0.077482458
## Entertainment          0.639997257 -0.175488639
## Film & Animation      -0.433107323  0.118758965
## Gaming                -0.012617377  0.003459711
## Howto & Style         -0.059230726  0.016241192
## Music                  0.225749540 -0.061901017
## News & Politics        0.426354922 -0.116907436
## Nonprofits & Activism  0.136106070 -0.037320579
## People & Blogs        -0.592739855  0.162530550
## Pets & Animals        -0.069985574  0.019190194
## Science & Technology  -0.600459261  0.164647238
## Sports                -0.037915494  0.010396509
## Travel & Events       -0.001100851  0.000301857
##
## with conditional variances for "categories"
```

Fixed effects of model

```
##  (Intercept) log_duration    log_views  log_numlang
##  -12.9308207    0.7865322    0.3294409    2.3281179
```

Coefficients of model

```
## $categories
##                       (Intercept) log_duration log_views log_numlang
## Autos & Vehicles        -13.33287    0.8205201 0.3274316    2.315938
## Comedy                  -12.90249    0.7841375 0.3204983    2.313871
## Education               -13.27362    0.8155107 0.3542668    2.250635
## Entertainment           -12.69615    0.7666939 0.3239545    2.152629
## Film & Animation        -12.38882    0.7407138 0.3633586    2.446877
## Gaming                  -12.90520    0.7843663 0.3662467    2.331578
## Howto & Style           -12.57230    0.7562243 0.2989768    2.344359
## Music                   -12.40780    0.7423184 0.3045648    2.266217
```

```
## News & Politics         -14.37130    0.9083046 0.3803954    2.211210
## Nonprofits & Activism   -13.13606    0.8038819 0.3207727    2.290797
## People & Blogs          -12.28132    0.7316265 0.3672157    2.490648
## Pets & Animals          -12.63507    0.7615303 0.3571264    2.347308
## Science & Technology    -13.37305    0.8239166 0.2092467    2.492765
## Sports                  -12.66039    0.7636708 0.3196971    2.338514
## Travel & Events         -13.02587    0.7945673 0.3278618    2.328420
##
## attr(,"class")
## [1] "coef.mer"
```