

# STAT 605 Final Report

Fangyang Chen, Guanqi Lu, Chong Wei, Youhui Ye,Runfeng Yong  
12/5/2019

## 1.Introduction

### 1.1 Data resource

Our data is about movies' reviews on Amazon, the raw txt file can be found from kaggle by the link below:  
<https://www.kaggle.com/dm4006/amazon-movie-reviews> (<https://www.kaggle.com/dm4006/amazon-movie-reviews>)

### 1.2 Description of the data

In this dataset we have 8 variables : the movie's Product ID, the review's User ID, the Profile name(username),helpfulness(number of clicks on 'helpful'), the score({1,2,3,4,5}) user rated the movie, time(timestamp),summary(wrote by user) and review text. Below is part of a record in the txt file.

```
product/productId: B003AI2VGA
review/userId: A328S9RN3U5M68
review/profileName: Grady Harp
review/helpfulness: 4/4
review/score: 3.0
review/time: 1181952000
review/summary: Worthwhile and Important Story Hampered by Poor Script
review/text: THE VIRGIN OF JUAREZ is based on true events surrounding
```

### 1.3 Question pursued

- a. The relationship of scores and the length of review text.
  - b. The relationship of scores and the helpfulness.
  - c. The words frequency of high and low helpfulness rate.
  - d. The way that user's behavior changed over time.
- We'll use the bootstrap, regression and other statistics methods to achieve our goal.

## 2. Process

### 2.1 Data extraction and cleaning

We extracted the useful data from the 9GB txt file in HPC(parallel), and removed the duplications. In the "helpness" part, we removed the 0/0 items. For the timestamp, we converted this to real date. After the data cleaning, we have about 1000000 reccords.This step was done with shell script.

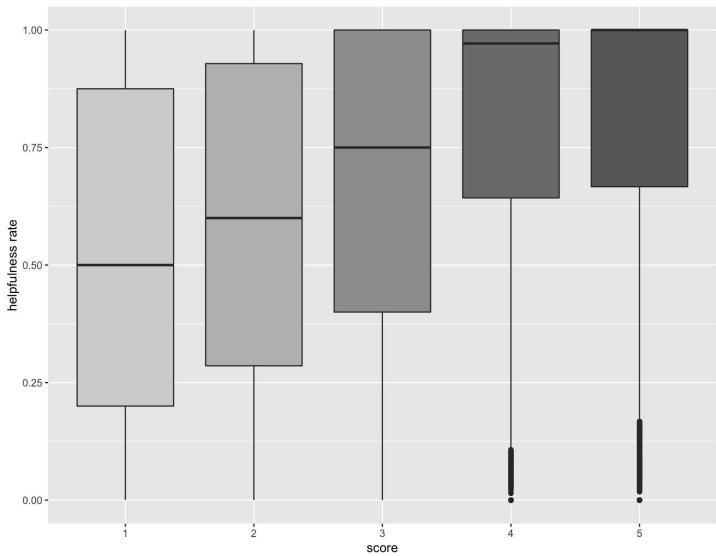
### 2.2 Analysis

#### 2.2.1 Score & Helpfulness

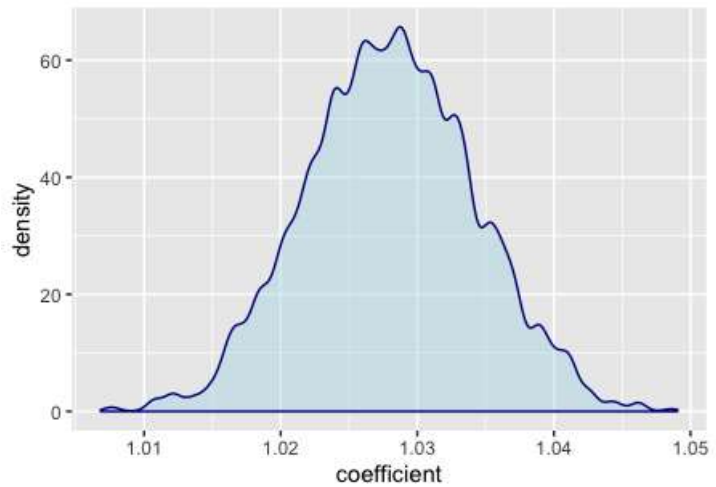
The average score is 4.1, the lowest is 1. From the summary we can see many users gave full score——5.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.0	4.0	5.0	4.1	5.0	5.0

Here is a boxplot of helpfulness rate of each score category.



We use bootstrap(parallel) to caculate coefficient of the regression between each score and each rate of helpfulness.Here is the distribution of the coefficients.

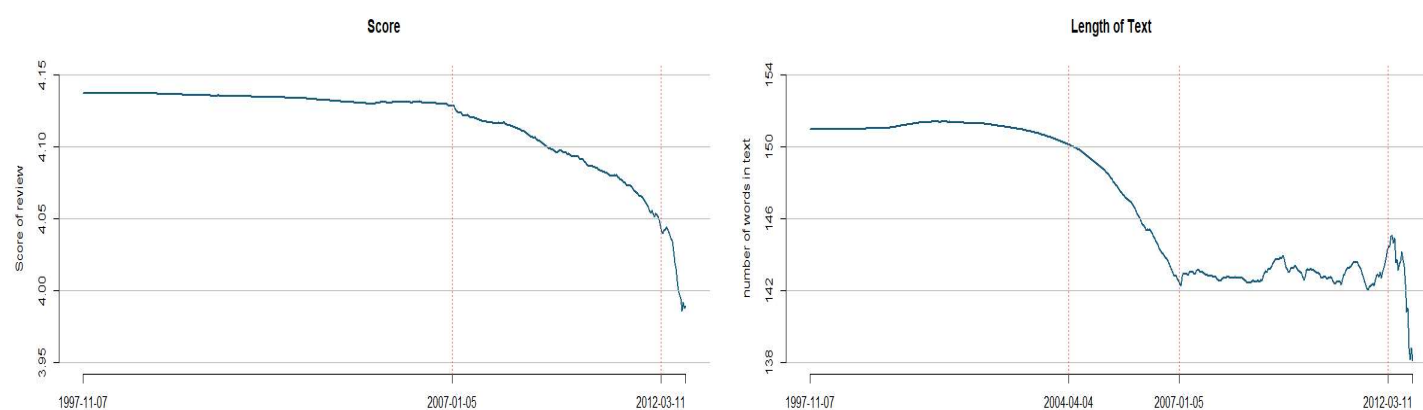


### 2.2.2 Score & Length of review text

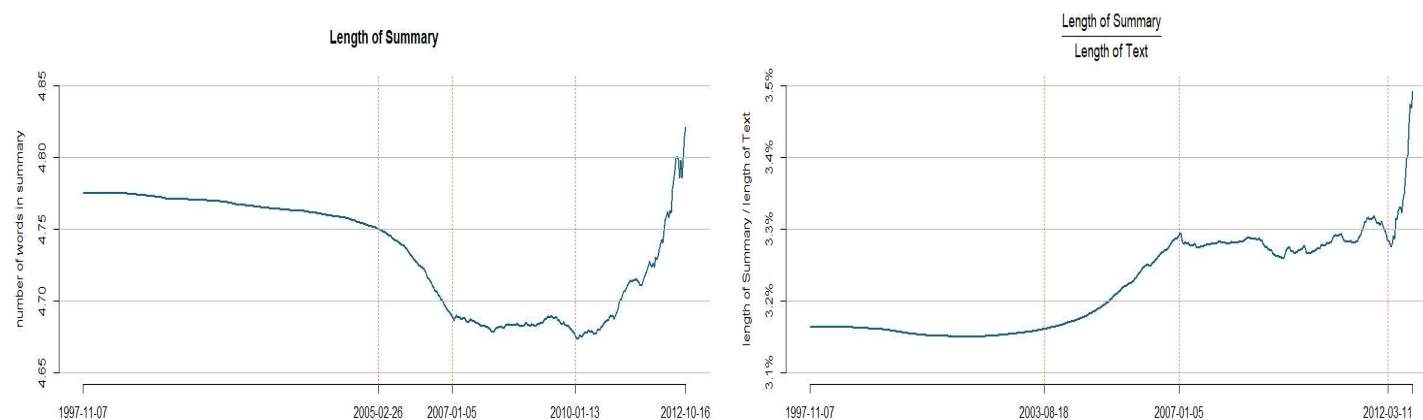
### 2.2.3 The Contents of reviews & Helpfulness rate

[illegible]

The length of review text also dropped a lot between 2004 and 2007, then it became steady till March 2012, after 2012, the length of review started to drop again.



Then we checked the ratio of length of summary and length of review text, we can conclude that when users give comments to the movies, they are more likely to write long reviews in the past. Nowadays, people tend to write shorter reviews, and focus on much shorter version of comments —summary.



Our weaknesses are: The regression we did is only linear regression, there may be other better models. The analysis of time is only done by plots, we didn't get into the quantity relationships.