Project 2

Coral Chen and Jessica Li

Apr 9, 2020

```
books <- read.csv("C:/Users/coral/OneDrive/Desktop/Wake Forest/2020Spring/STAT212/Project 2/books.csv")
```

Executive Summary/ Abstract

In this project, we choose a dataset that records ratings and the information about books on a website called "Goodreads," including the publication dates, author, number of people who wrote reviews, and so on. By exploring these different features of a book, we hope to draw a conclusion on what decides the popularity of a book on this website, and that leads to our research question - What contributes to the average rating of books on the Goodreads website?

To answer the question, we first change our response variable rating score into a binary term by categorizing them into which is higher than the median score 3.69, and this isn't so that we could construct an empirical logistic model to explain our question. We then proceed to the EDA process, where we dropped some insignificant levels for categorical predictors and transformed the numerical variables to ensure its linear relationship against the response variable by comparing the deviance. After checking the multicollinearity to build an interaction between predictors, we have a rough model ready at this point. The next step is to use the nested F test and the BSS techniques to select out the best model with the lower AIC and compare it to the rough model; we then check the condition for inference and draw out a final model:

$$log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$$
$$= -0.3247 + 0.09153languageCodeEng + 3.352languageCodeJpn$$
$$+ 0.000001764sqNum_pages + 0.01106SqrtRatings_count$$
$$- 0.05160sqrtText_reviews_count$$
$$- 0.00000009923SqrtRatings_count: sqrtText_reviews_count$$

With this model, we can now made a conclusion for the rating score of the books in regard to the research question: the language as English and Japanese, number of pages, the number of rating are positively contributes to the odds of gaining a high rating score for a book in the Goodreads website, while the number of text reviews count is making negative contribution to it.

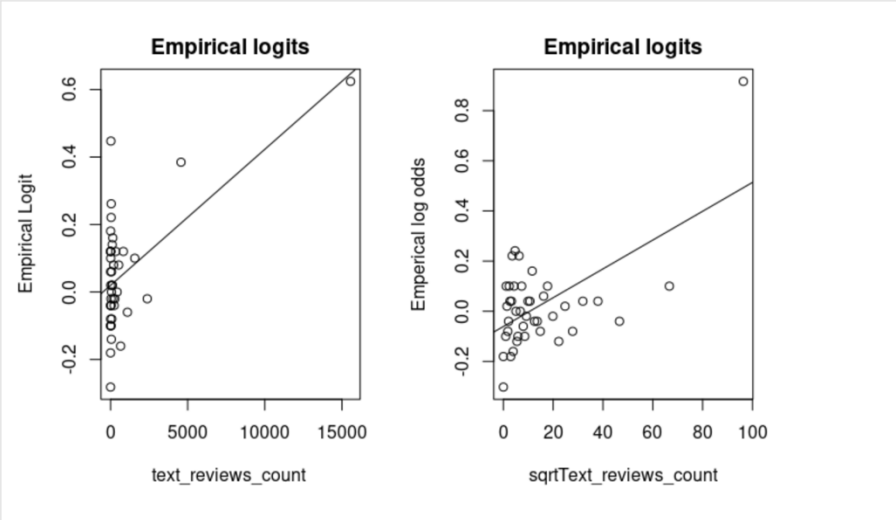
Data

This set of data is about the information on books that comes from the Goodreads website. Soumik, the creator, collected the data via Goodreads API because Goodreads allows its members to access the Goodreads website. In the data, it contains 11132 books and twelve relevant variables for each book. Each row of data corresponds to different variables for one particular book. In this project, we decided to choose an average rating as our response variable, which is represented by "average_rating" in the dataset. Since we are analyzing a logistic regression model, the response variable needs to be binary. Therefore, we calculated the mean of an average rating, which is 3.96. Anything higher or equal to 3.96 will be a high rating, and anything lower than 3.96 will be below rating. The possible predictors for our dataset are bookID, language code, number of pages, number of ratings, number of text reviews, and publication date. In this dataset, we have 8472 rows of data, and it is a very clean dataset, so there are missing data, but only as few as 4 data points. The citation for the data is:

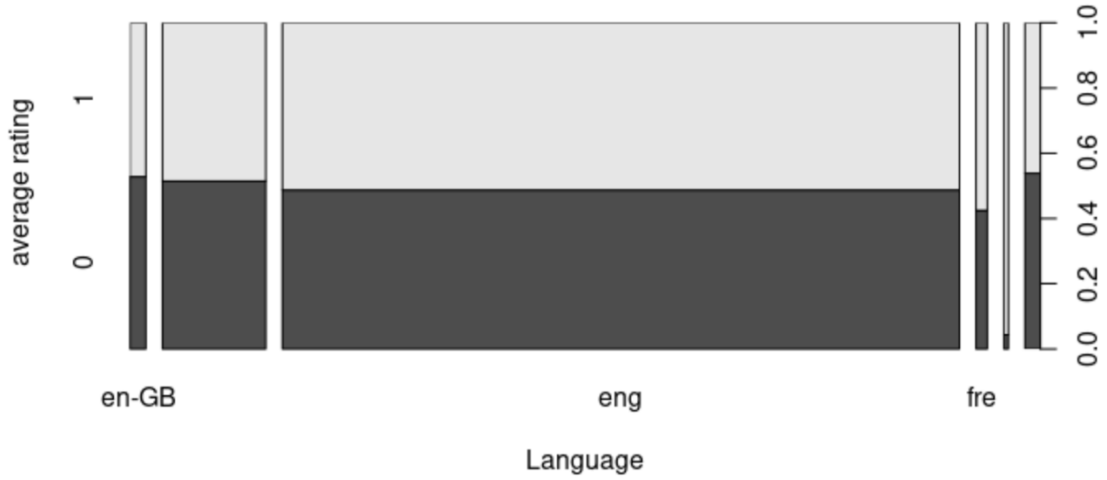
Soumik. (2019) Goodreads-books – comprehensive list of all books listed in goodreads [Data set]. Retrieved from https://www.kaggle.com/jealousleopard/goodreadsbooks.

Exploratory data analysis (EDA)

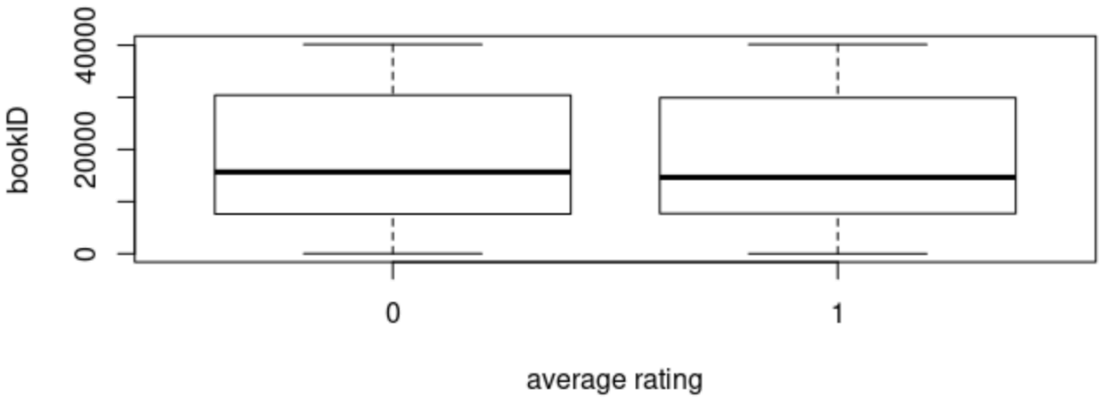
Before the exploratory data analysis, we decided to delete a few columns. After our careful analysis and consideration, title, authors, isbn, isbn13, and publisher were deleted. We choose to delete titles, authors, and publishers because those three variables are names, and there are too many variations that do not contribute to the analysis. We decided to delete isbn and isbn13 because they are randomly assigned code that does not have meaning behind them, and there is no need to include them. Therefore, we are left with four numerical variables (bookID, Num_pages, Ratings_count, Text_reviews_count) and two categorical variables (language code and Publication date). For each numerical variable, we decided to use an empirical logit plot to verify linearity. After analyzing the plots, we figured out several outliers that can be very influential for the regression line. Therefore, the thing we did to fix the plots and make them better is to remove influential outliers. After outliers are removed, we think several plots need transformations because those plots are far from the regression line. Therefore, we decided on the transformations based on the plots. The second plot is the number of pages, and since the plot is concave up, we decided to transform up and square the variable. The third plot is rating count, and we did a square root transformation to transform down and make the plot more linear. The fourth plot is text reviews count, and the plot is concave down. Therefore, we did a square root transformation on the variable.



After the transformation, we finished the analysis for numerical variables, and the next step is to analyze the categorical variables. For language code, we first plotted the data and showed a table that includes all the variables and counts of data points. Then we decided to only keep Japanese, French, Spanish, English, English-US, and English-GB, because other levels have not a lot of data points that have significant contributions to the overall analysis.



The second variable is publication data, the plot we got shows total black that data points are all over the plot, and it is very hard to change the form of data. Therefore, we decided to remove the publication date. The next step is to check for multicollinearity. We plot the table and figure out that the only multicollinearity is between SqrtRatings_count & sqrtText_reviews_count, which is 0.914. In the analysis, we figured out that BookID is not a great predictor, because the box plots are quite similar, and the mean is nearly the same.



Therefore, it does not provide much information for use. However, we decided to keep it for now and do more analysis at the model selection. In the end, our rough model includes the following predictors: language_code,bookID, sqNum_pages, SqrtRatings_count, sqrtText_reviews_count, and SqrtRatings_count:sqrtText_reviews_count.

Model Selection

We choose to do both nested tests and subset selection, because by combining both BSS and nested likelihood ratio tests, our result will be more precise and logical.

We first use BSS techniques to find out the model with the lowest AIC, where AIC is a measure that indicates how well the model explains the data. We started out the process by putting the rough model into BSS, to do this we need to build a matrix with all the predictors that are included in the rough mode by inputting the code model.matrix. The following step is to run the BSS for the matrix that we built and that gives out various possible models with different combinations of predictors. The BSS techniques would compute the AIC for each model and that enables us to find the one with the lowest AIC by the code bssOUT$Best Model. We name the model that was selected as " Refined model". We then use the nested F test again after the BSS, comparing the rough model and the refined model with the code "anova(RefinedModel, RoughModel, test ="Chisq")", and the result of p-value equals to 0.3919 indicates we should choose to use the refined model as it exceed the level 0.05.

To further refine our model, we took a look at the p value for each predictors in the refine model to see whether all of them are significant. Using α=0.05 as the significant level, we can see that the predictors language codeen-US, eng, fre, and spa exceed 0.05. We then remove those insignificant predictors but keep the predictor code-eng, considering there are 6815 data points fall into this category and we don't want to delete the majority of the data.

At the end of this process, we have our model as: $log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -3.247e-01 + 9.153e-02languageCodeEng + 3.352e+00languageCodeJpn + 1.764e-06sqNum_pages + 1.106e-02SqrtRatings_count - 5.160e-02sqrtText_reviews_count - 9.923e-08SqrtRatings_count:sqrtText_reviews_count$

Final Model and condition for inference

There are four conditions that we need to check: binary response variable, linearity, randomness, and independence. In the EDA section, our response variable was originally a numerical variable, but we changed it to binary that "1" represents the average rating is higher or equal to 3.96, and "0" means that the average rating is lower than 3.96. Therefore, the first condition is met. The second condition is linearity. In our EDA, we drew all the plots and checked for linearity in order to determine the transformation of numerical variables. The violations do exist for numerical variables. Therefore, we transformed each numerical variable and removed outliers in order for our plots to be more linear (The plots can be found in the EDA section for transformation and linearity). The third condition is randomness. In this dataset, the creator uses the Goodreads website, which is an authoritative site for books that includes many different kinds of books. In addition, in the introduction part, the creator highlights that the books

he selects are regardless of language and publication. We can also see randomness in the datasets because the publication data are ranged widely and very different from one to another. The last condition is independence. In this dataset, all the books are from different authors, different publishers, and published in different time periods. Therefore, each row does not relate to others in any way. In other words, removing one row of data does not affect other rows of data, so they are independent of each other. The AIC is 10289, and the percent reduction in deviance is 3.585%. Below is our final model:

Final Model: $log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -0.3247 + 0.09153languageCodeEng + 3.352languageCodeJpn + 0.000001764sqNum_pages + 0.01106SqrtRatings_count - 0.05160sqrtText_reviews_count - 0.00000009923SqrtRatings_count: sqrtText_reviews_count$

Analysis and Conclusion

The model we fitted shows that the language code as English and Japanese, the number of pages and the rating count positively affect the odds of a book getting an average rating score that is higher than the median score among the books in the Goodreads website, while the reviews count would negatively affect the rating score of a book. Besides, we can also see how influential these predictors are to the ratings of the book by looking at each coefficient. For example, we can see that one additional increase in the rating count would lead to an increase in the odds of getting a high score by the multiplier of e^1.106e-02. The model enables us to answer our research question - having the book with the language of English and Japanese, have the number of pages, and rating count as more as possible would contribute to a higher odds of getting a high rating score in the GoodReads Websites, where we define the “high rating score” as being higher than 3.96. We are comfortable making conclusions with this model, but we would feel better if we could have improvements in the following limitations: 1.the size of the database. As the data only covers reviews for around 8000 books at a certain time period, the conclusion we made is only applicable in that certain range. 2. The influence of predictors as publishers and authors. We know by common sense that people tend to rate high on the book published by prestige authors and publishers, but we couldn’t manage to include these two variables in our model-based what we learned in 212 class as there are too many different levels, and the omitting these two predictors could cause some impreciseness for out answer to the research questions.