My title*

My subtitle if needed

Hyunje Park

November 27, 2024

This paper examines the rating of childrens' books using the . By using a Bayesian logistic regression and data from the Alex Cookson (CITATION), I analyzed the relationship between various characteristics of a childrens' book, and whether or not it influenced if the book had a high rating or not. The insight from this research not only clarify what makes a high rated childrens' book so high rated, but also enhance future authors to push the boundaries of their literature.

1 Introduction

For generations, children's books have been a catalyst in children's development. Through captivating fairytales and fables of stories about adventures, heroes, magical forests and magic, children gain experiences on feelings and thoughts, learning to cope with inhibitions, vulnerability and shyness (CITATION). Beyond educational purposes, children's literature can positively influence mental wellbeing, feelings and behavior. Given these significant developmental benefits, it can be said that the quality of the children's book matters greatly; choosing a well-written book can affect how it nurtures the next generation of mature, emotionally resilient individuals. This is why book rating systems hold significant importance; allowing readers to rate books on a scale from 0-5 scale (most commonly used scale) helps parents and educators assess whether a book is worth giving to children.

In this 0-5 rating system, a score of 4 or above is often seen as the benchmark for a "highly rated" book. This is influenced by central tendency bias, where people naturally gravitate towards a moderate score, avoiding extremes like 0 or 5 to appear more balanced and objective. A rating such as 4, in particular, suggests a strong endorsement without overstepping into exaggeration. Given this, a critical question arises; What factors of a book contribute to the likelihood of the book being "highly-rated" (a score above 4)?

*Code and data are available at: https://github.com/davidpxrk/childrens-book-rating. Special thanks to Rohan Alexander for his help!

In this paper, I analyzed how the characteristics of a book, such as book type, page count, publishing year, and rating counts affected the likelihood of a book being "highly-rated" using the Children's Book Ratings Data (CITATION). First, after data cleaning, I selected 9 variables on children book characteristics for my analysis in (SECTION 2). Then, a logistic regression model was created to predict the probability of the book being "highly-rated", based on the chosen book characteristic variables.

The logistic regression model showed that _____ .The findings of this research have practical implications for the writing industry, to allow future generation authors to push the boundaries of literature.

This research paper is structured as follows: (SECTION 2) contains an overview of the dataset and some tables and graphs used to illustrate the variables employed in this analysis. (SECTION 3) describes and justifies the logistic regression model that was produced in this report. (SECTION 4) highlights the result of the model, (SECTION 5) discusses some of the outcomes, weaknesses, and (SECTION APPENDIX) contains additional information on model details.

1.1 Estimand

The estimand of this paper is the probability that a book is highly-rated (has an average rating of over 4 on a 0-5 scale), based on book characteristics. It is difficult to measure the exact number as there are millions of children's books that are published and not all of them will be accessed due to various issues. For examples, children's books from different countries may have different ratings. Therefore in this paper, we attempt to estimate the estimand using a logistic regression model which is fitted using a sample from the Children's Book Rating dataset (CITATION)

2 Data

The dataset used in this paper was obtained from COOKSON (CITATION), who sourced it from Goodreads (CITATION). Cookson's dataset contained over 9,000+ records of children's books that contained information on the book such as title, author, ratings, publisher and more. In the initial step, data-cleaning was performed and relevant variables were selected. First, entries with missing information were filtered out, and variables that were irrelevant to this study was also removed, which included variables such as book title. After cleaning, there were 9240 entries in the cleaned dataset.

Data analysis is performed using statistical programming language , along with packages

2.1 Data Cleaning

```
# A tibble: 9,240 x 12
   cover     pages publisher  publish_year rating rating_count rating_5 rating_4
   <chr>     <int> <chr>             <int>  <dbl>        <int>    <int>    <int>
 1 Paperback    NA HarperCol~         2005   4.22      2055091   985699   650702
 2 Hardcover    NA Riverhead          2015   3.92      2002733   648904   764208
 3 Hardcover    NA Scholastic         2003   4.47      1734916   370456   543695
 4 Paperback    NA HarperCol~         2001   4.17      1364643   638927   422372
 5 Paperback    93 Harcourt           2000   4.31      1277979   717114   331172
 6 Hardcover   176 Harpercol~         2002   4.3       1151744   627508   320602
 7 Hardcover   451 Amy Einho~         2009   4.47      1084920   253256   615592
 8 Hardcover    64 HarperCol~         1964   4.38       876053   537395   198996
 9 Paperback    37 Red Fox            2000   4.22       788702   418885   200789
10 Hardcover    NA Margaret ~         2009   4.32       767112   416566   220754
# i 9,230 more rows
# i 4 more variables: rating_3 <int>, rating_2 <int>, rating_1 <int>,
#   rated_high <dbl>
```

Table 1: Preview of the Cleaned Children's Book Dataset

cover	pages	publisher	publish_year	rating	rating_count
Paperback	NA	HarperCollins Publishers	2005	4.22	2055091
Hardcover	NA	Riverhead	2015	3.92	2002733
Hardcover	NA	Scholastic	2003	4.47	1734916
Paperback	NA	HarperCollinsPublishers	2001	4.17	1364643
Paperback	93	Harcourt	2000	4.31	1277979

rating_5	rating_4	rating_3	rating_2	rating_1	rated_high
985699	650702	323439	68978	26273	1
648904	764208	423888	117612	48121	0
370456	543695	577239	132741	110785	1
638927	422372	227845	50232	25267	1
717114	331172	160400	46684	22609	1

Table 1 (CROSS REFERENCE) presents the first 5 rows from the cleaned dataset, that contains information on Children's book.

2.2 Variable

This analysis focuses on the following variables, with a focus on `rated_high` as the dependent variable

- `rated_high`: A binary variable telling us if a book is rated high or not (above a 4 rating)
  - `0`: The book has a rating below 4
  - `1`: The book has a rating greater or equal to 4
- `pages`: An integer that represents the number of pages for the book
- `rating_count`: An integer that represents how many ratings the book received
- `publish_year`: An integer that represents the year the book was published
- `cover`: A categorical variable that represents the type of the book, which can include:
  - `Ebook`: The book is in a digital format.
  - `Kindle`: The book is only available through Kindle e-reader tablet.
  - `Board book`: The book is a picture book designed for the youngest of children, babies, infants.
  - `Hardcover`: The book has a rigid protective cover.
  - `Paperback`: The book is has a thick paper cover.

Table 2 (CITATION) shows a summary of detailed statistics about the cleaned dataset. It showed that books that were considered "high rated" had more pages on average than low rated books. Furthermore, high rated books were on average published 2 years before low rated books, around the year 2003. In addition to this, the amount of ratings high-rated books received far exceed low rated books, at about 4x. The most common type of book cover also turned out to be hard covers. These variables are key in the analysis, and are further shown in (CITE HERE)

rated_high	avg_pages	avg_publish_year	avg_rating_count	avg_cover
0	57.63858	2005.513	1963.318	Hardcover
1	66.06849	2003.934	8040.772	Hardcover

Summary Statistic of the Cleaned Dataset

2.3 Data Analysis

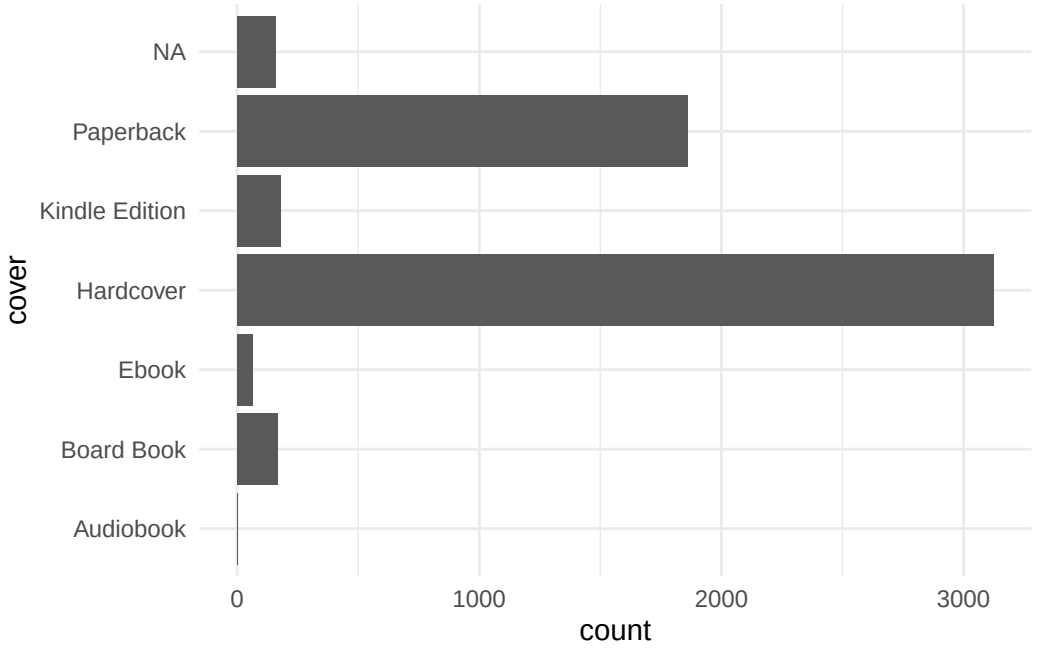


Figure 1: Summary Statistic of the Cleaned Dataset

(CITATION HERE PICTURE ABOVE) shows the number of cover types for "high-rated" books (rated_high = 1). It is evident that Paperback and Hardcover books make up the majority of "high-rated" books, which could potentially reflect the aesthetic and collectible value of Paperback/Hardcover books over something non-physical like an Ebook. This illustrates that physical books could be more appealing for readers due to ownership bias, greatly increasing their rating. This finding also reflects the result from (TABLE)

(CTATION HERE) shows that generally, "high-rated" books have more page count than low-rated books. This suggests that higher page counts could correlate with higher ratings as more pages could allow deeper storytelling, richer world-building or character development, which can result in a more immersive experience for the reader. This graph illustrates the importance for authors to use more pages for properly structured narratives.

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
Warning: Removed 91 rows containing missing values or values outside the scale range
(`geom_point()`).
```

Figure 2: Summary Statistic of the Cleaned Dataset

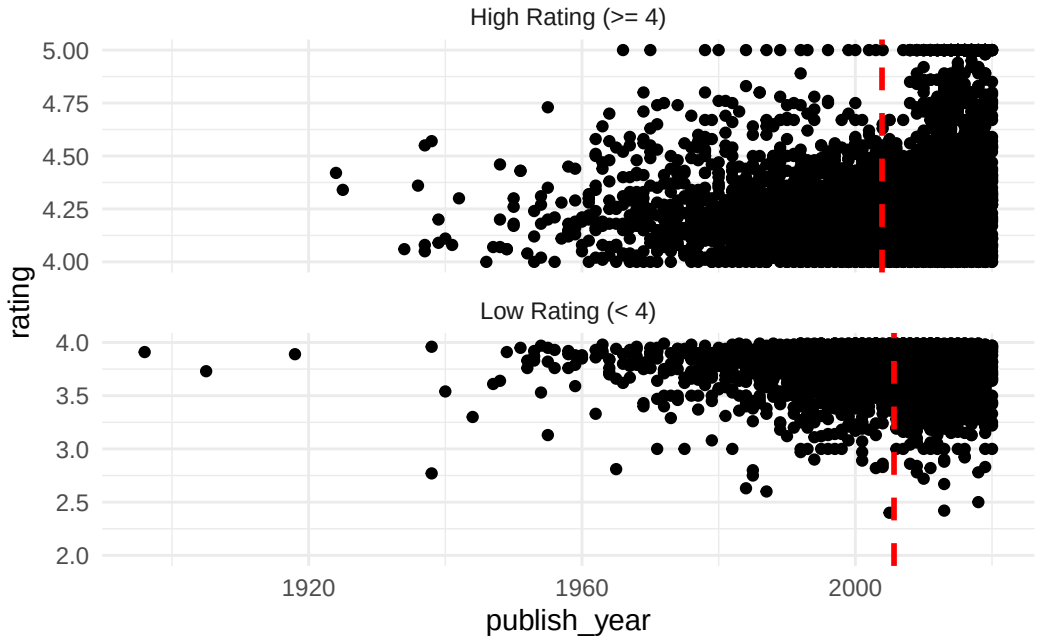


Figure 3: Summary Statistic of the Cleaned Dataset

(CITATION) shows the relationship between year of publish and the book's rating. Where the horizontal axis shows the year of publishment, and vertical axis shows the rating of the book. By looking at the red line of fit (average year of publishment), "high-rated" books tend to be published before low-rated books. This could be due to selective survival where older books that remain in circulation could be considered "classics" and have stood the test of time.This is further discussed in (DISCUSSION HERE)

```
Warning: Removed 37 rows containing missing values or values outside the scale range
(`geom_point()`).
```

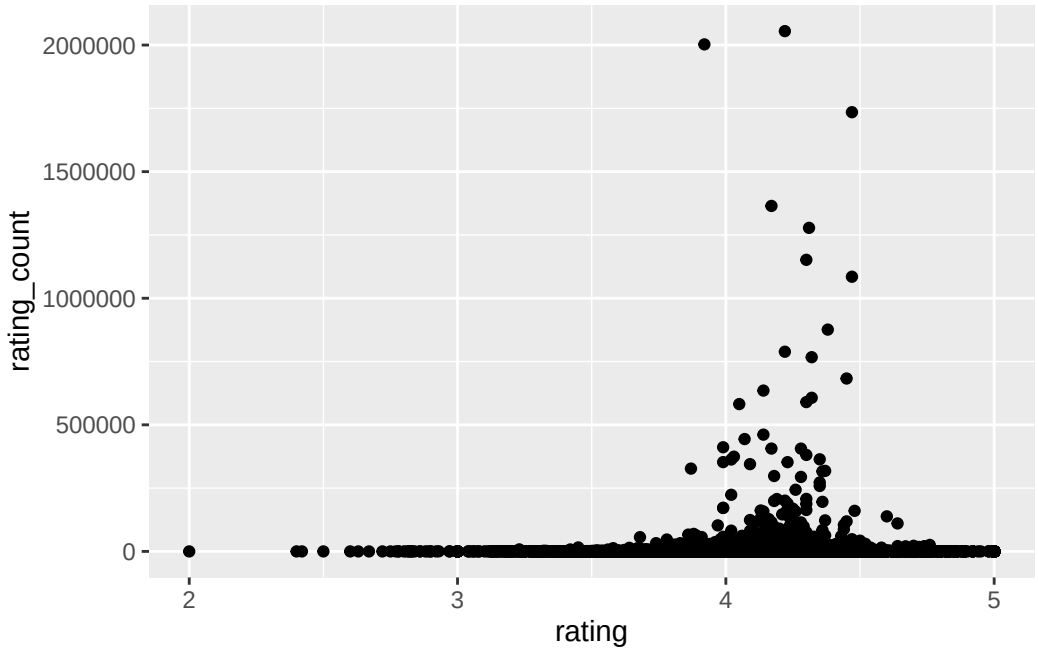


Figure 4: Summary Statistic of the Cleaned Dataset

(FIGURE 5) shows the relationship between the rating of a book and the number of ratings it received. The red line indicates the cut-off for a book to be considered "highly-rated". It is apparent that books that are considered "highly_rated" (right of the red line) have far more ratings on average than low-rated books.

3 Model

This study employed a Bayesian logistic regression model to analyze the relationship between Children's book's "highly-rated" status and the characteristic of the book. The model is as follows:

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) && (1)\\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 \times \text{pages}_i && (2)\\
&\quad + \beta_2 \times \text{rating\_count}_i + \beta_3 \times \text{publish\_year}_i && (3)\\
&\quad + \beta_4 \times \text{coverEbook}_i + \beta_5 \times \text{coverKindleEdition}_i && (4)\\
&\quad + \beta_6 \times \text{coverBoardBook}_i + \beta_7 \times \text{coverHardcover}_i && (5)\\
&\quad + \beta_8 \times \text{coverPaperback}_i && (6)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (7)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (8)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (9)\\
\beta_3 &\sim \text{Normal}(0, 2.5) && (10)\\
\beta_4 &\sim \text{Normal}(0, 2.5) && (11)\\
\beta_5 &\sim \text{Normal}(0, 2.5) && (12)\\
\beta_6 &\sim \text{Normal}(0, 2.5) && (13)\\
\beta_7 &\sim \text{Normal}(0, 2.5) && (14)\\
\beta_8 &\sim \text{Normal}(0, 2.5) && (15)
\end{aligned}
$$

3.1 Prior Distribution

The rstanarm package (CITATION) was used in order to run the regression model mentioned above. The default priors of the rstanarm package was used (CITATION). These default priors are designed to be weakly informative, meaning they provide enough information to regularize the model and prevent extreme estimates, while being flexible to let the data drive the inference. By default, priors for regression coefficients are centered around 0 with a standard deviation of 2.5, which is shown in equations (Equation LIST).

3.2 Model Justification

The predictor variables shown in (LIST HERE) were chosen due to the number of studies regarding their influence on ratings of books.

More specifically, I expect that

4 Results

After running the regression model aboe, we receive the coefficient values showcased in Table X.

5 Discussion

6 Appendix

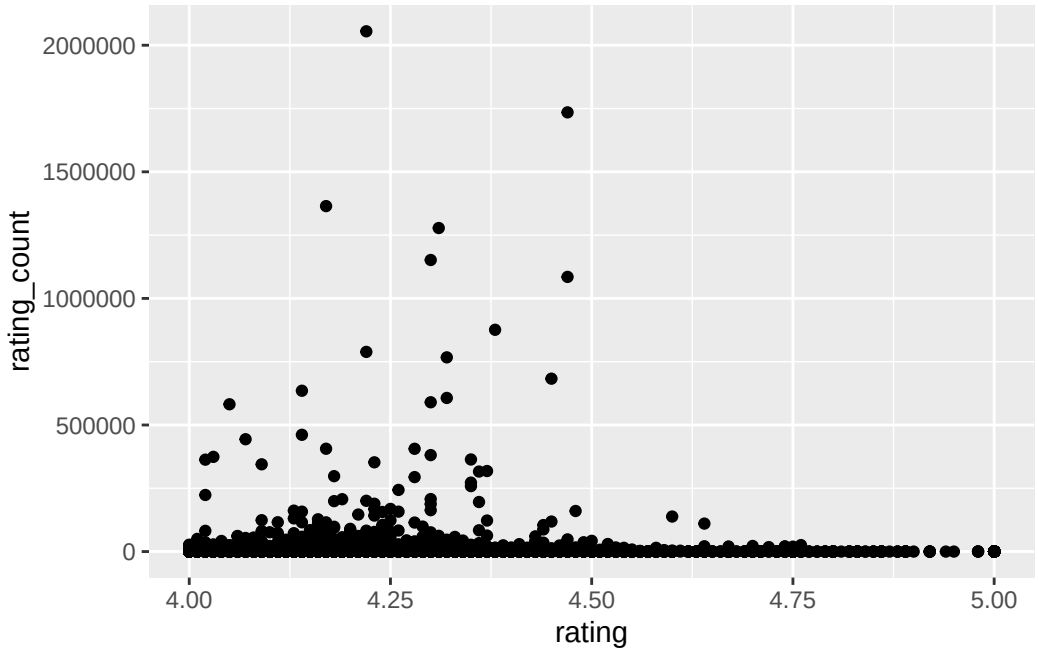


Figure 5: Summary Statistic of the Cleaned Dataset

```
Warning: Removed 501 rows containing missing values or values outside the scale range
(`geom_point()`).
```

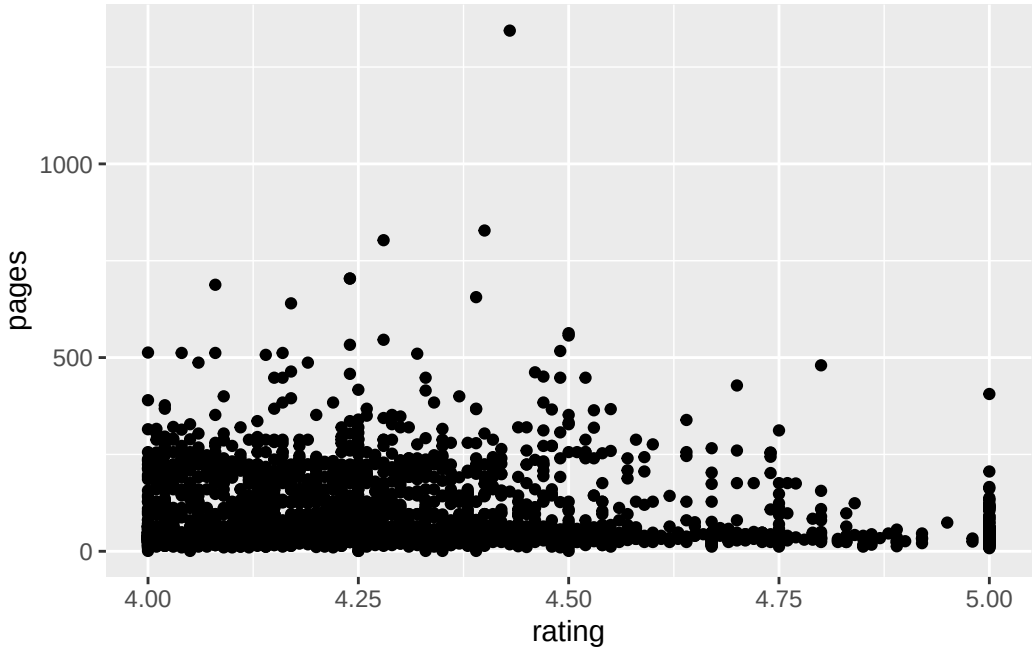


Figure 6: Summary Statistic of the Cleaned Dataset

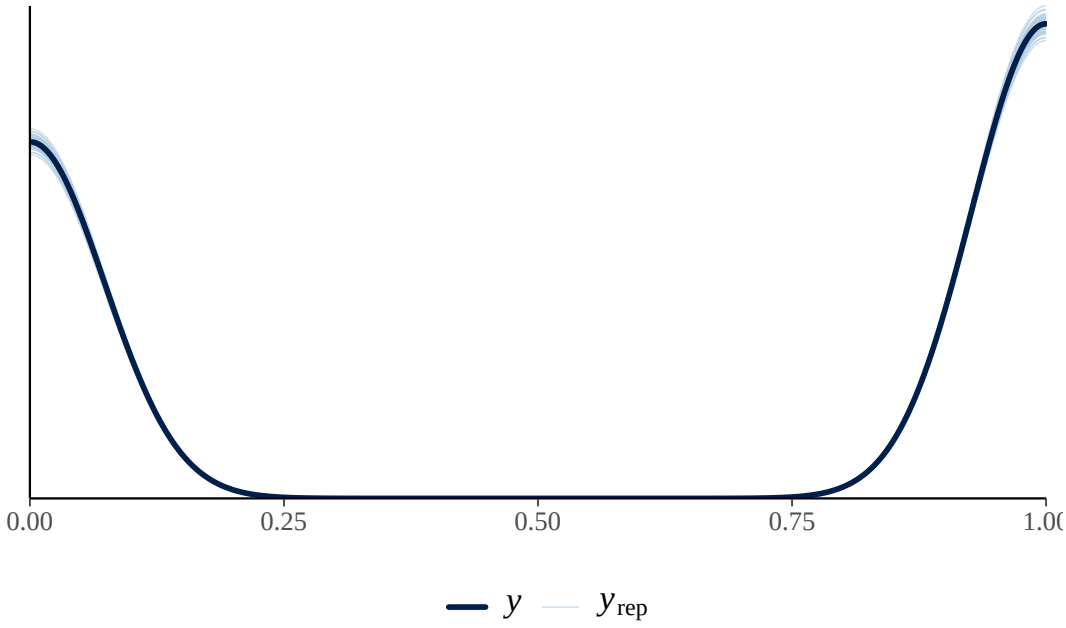


Figure 7: Summary Statistic of the Cleaned Dataset

```
Drawing from prior...
```

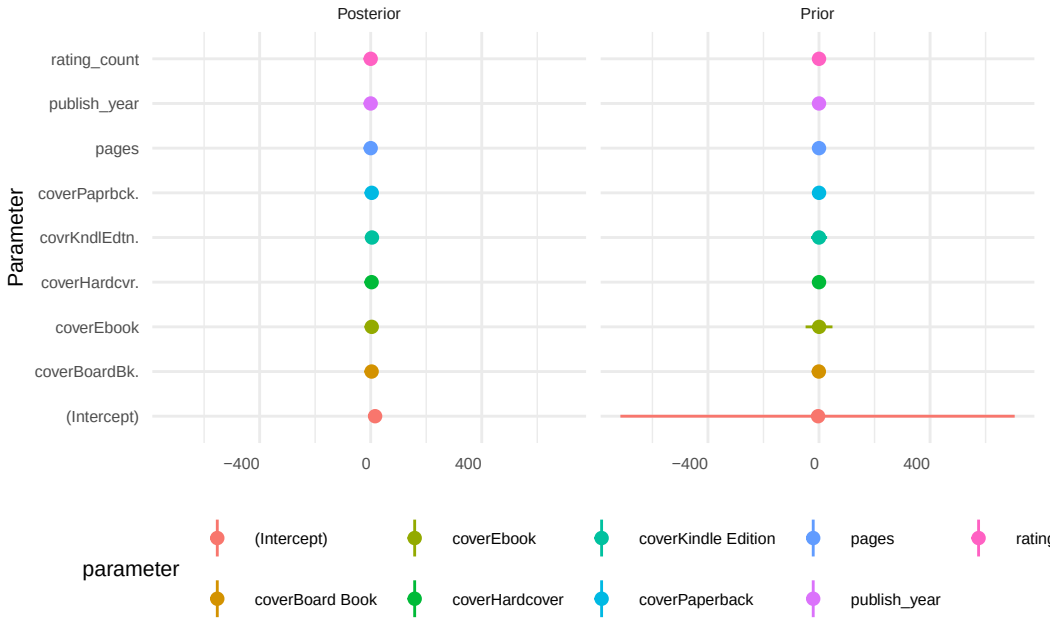


Figure 8: Summary Statistic of the Cleaned Dataset

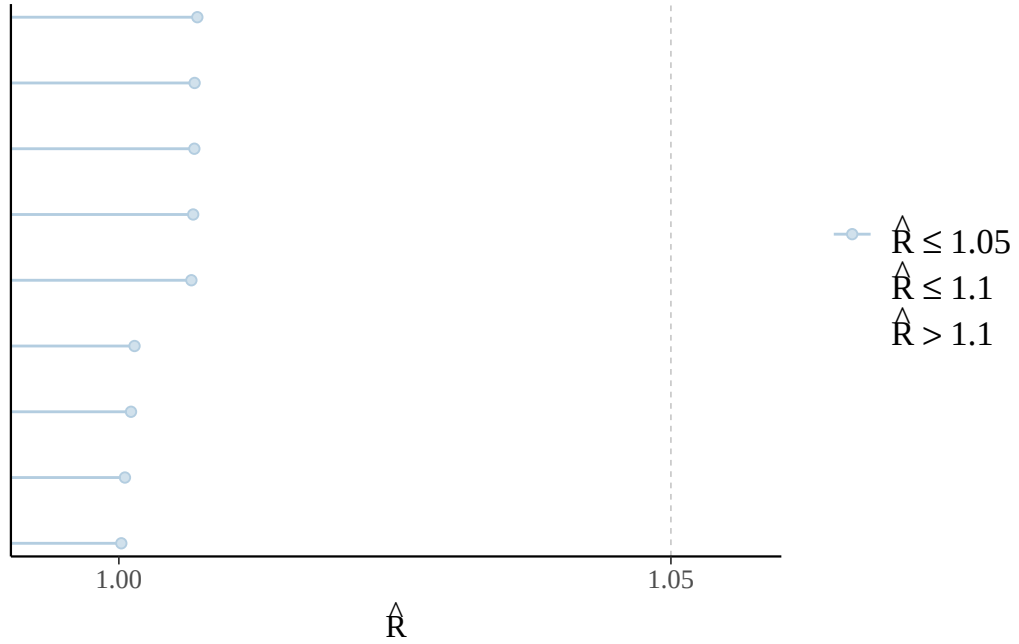
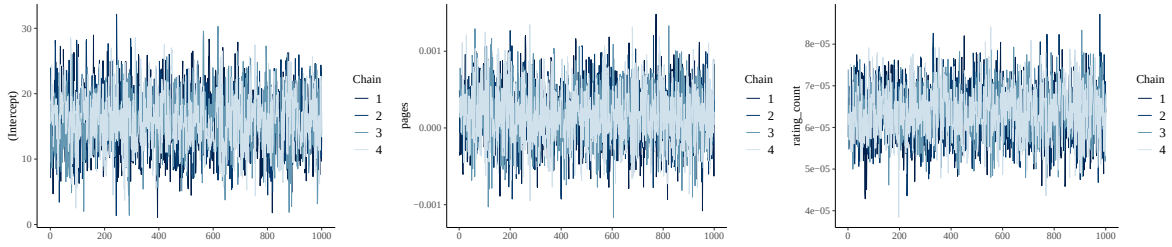
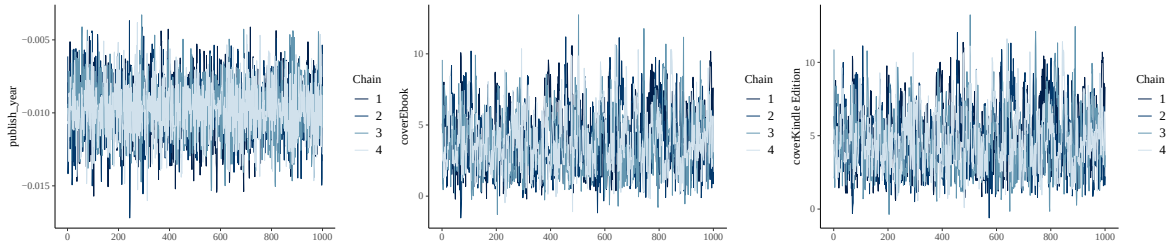


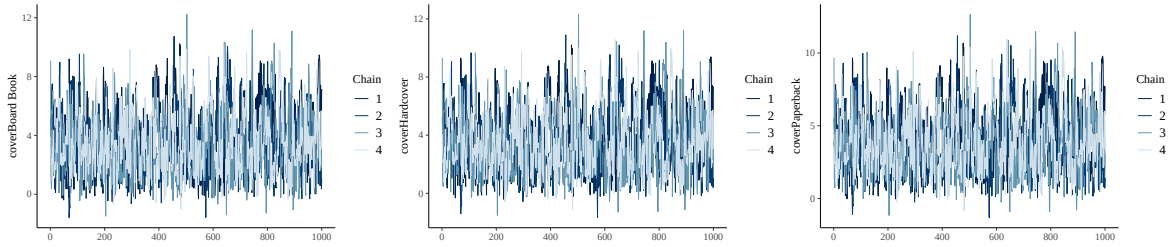
Figure 18: Summary Statistic of the Cleaned Dataset



(a) Intercept (a) pages (a) rating_count



(a) publish_year (a) coverEbook (a) coverKindle



(a) coverBoard (a) coverHardcover (a) coverPaperback

Summary Statistic of the Cleaned Dataset

–> # References