My title*

My subtitle if needed

Hyunje Park

November 27, 2024

This paper examines the rating of childrens' books using the . By using 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.

2 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)

3 Data

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

4 Model

5 Results

6 Discussion

7 Appendix

```
# 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>
```

$$\frac{x}{5}$$

Preview of the Cleaned Dataset

cover	pages	publisher	publish_year	rating	rating_count	ra
Length:9240	Min. : 1.00	Length:9240	Min. :1896	Min. :2.000	Min. : 0	Mi
Class :character	1st Qu.: 32.00	Class :character	1st Qu.:1999	1st Qu.:3.860	1st Qu.: 67	1st
Mode :character	Median : 34.00	Mode :character	Median :2008	Median :4.070	Median : 335	Me
NA	Mean : 62.35	NA	Mean :2005	Mean :4.061	Mean : 5466	Me
NA	3rd Qu.: 48.00	NA	3rd Qu.:2014	3rd Qu.:4.250	3rd Qu.: 1408	3r
NA	Max. :1344.00	NA	Max. :2020	Max. :5.000	Max. :2055091	Ma

NA	NA's :693	NA	NA's :92	NA's :37	NA	NA

Summary Statistic of the Cleaned Dataset

8 Data

```
Warning: Removed 61 rows containing missing values or values outside the scale range
(`geom_point()`).
```

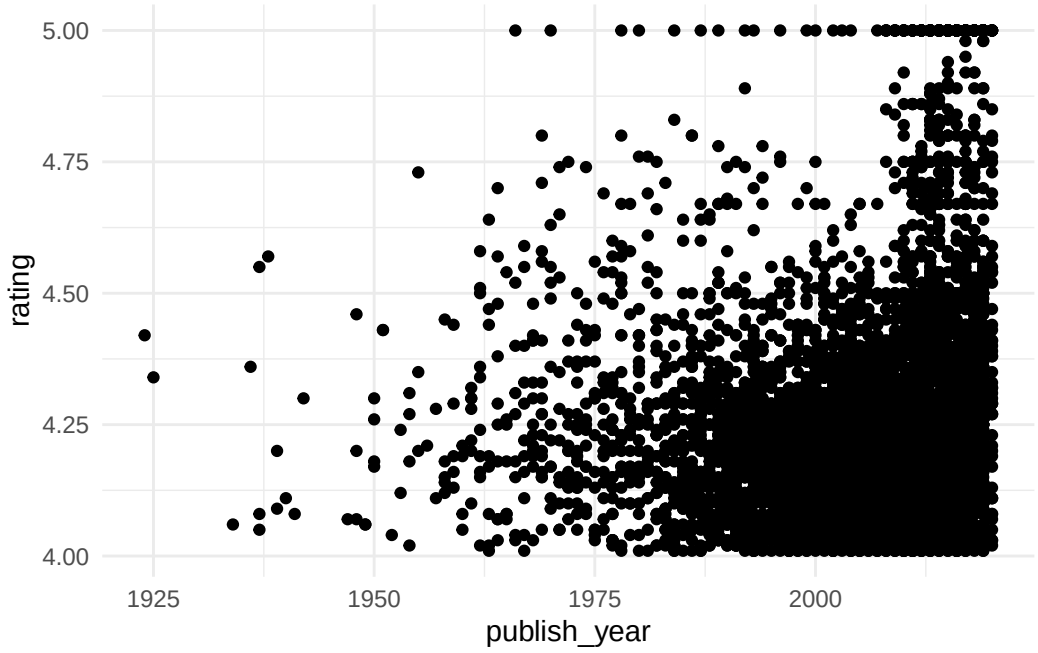


Figure 1: Summary Statistic of the Cleaned Dataset

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

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

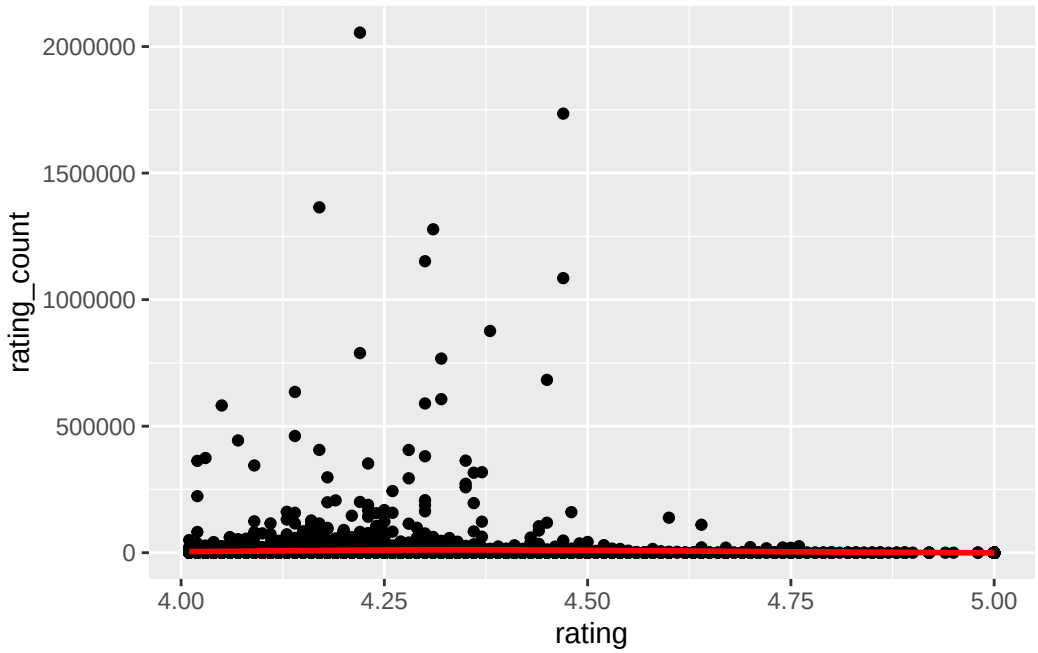


Figure 2: Summary Statistic of the Cleaned Dataset

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 476 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 476 rows containing missing values or values outside the scale range
(`geom_point()`).
```

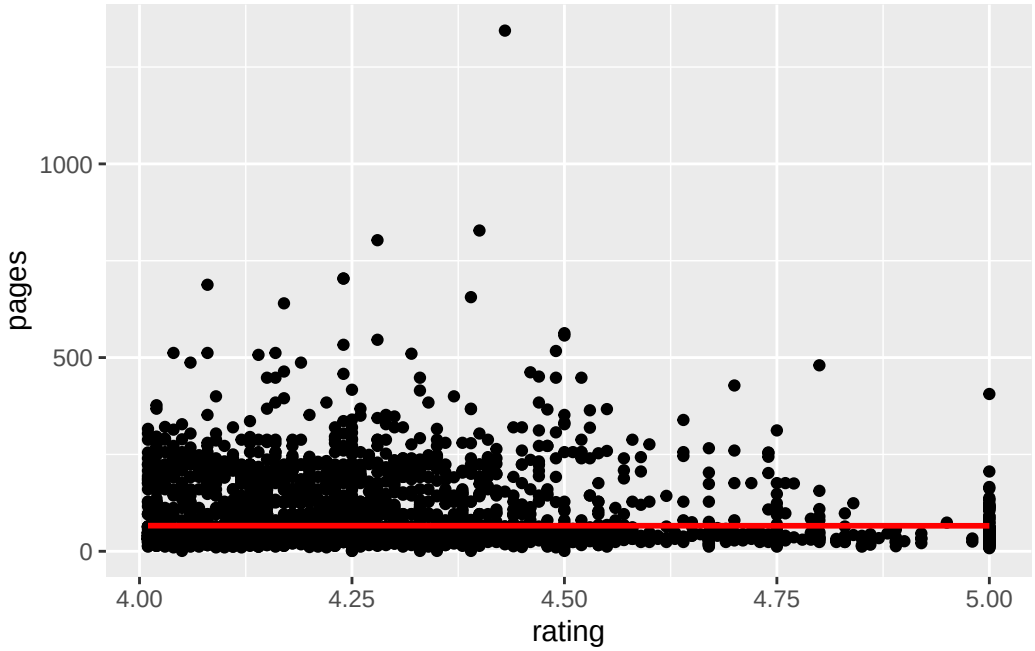


Figure 3: Summary Statistic of the Cleaned Dataset

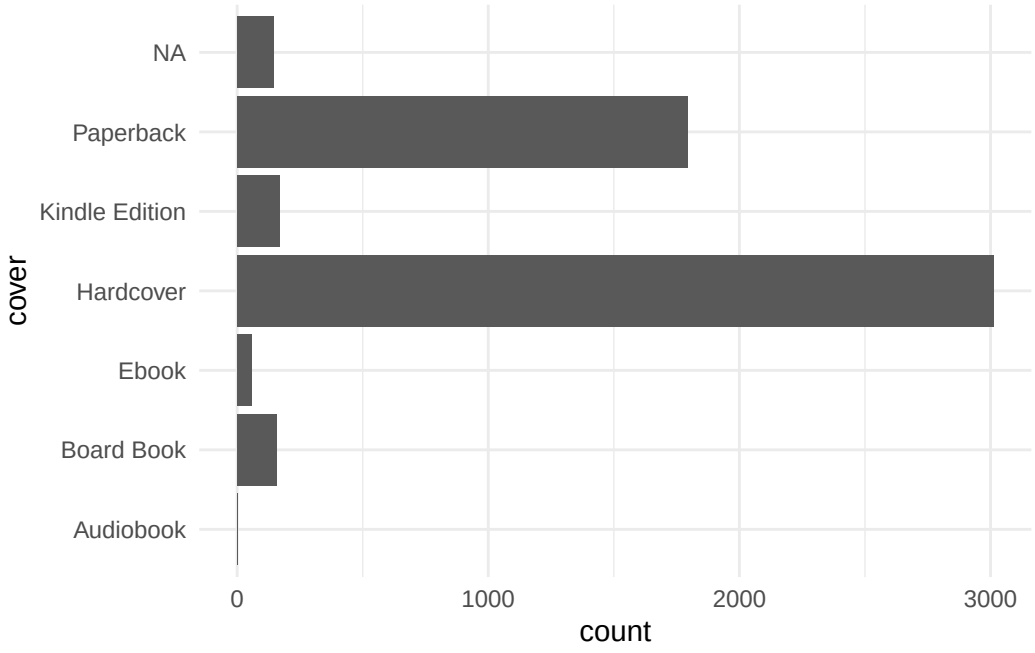


Figure 4: Summary Statistic of the Cleaned Dataset

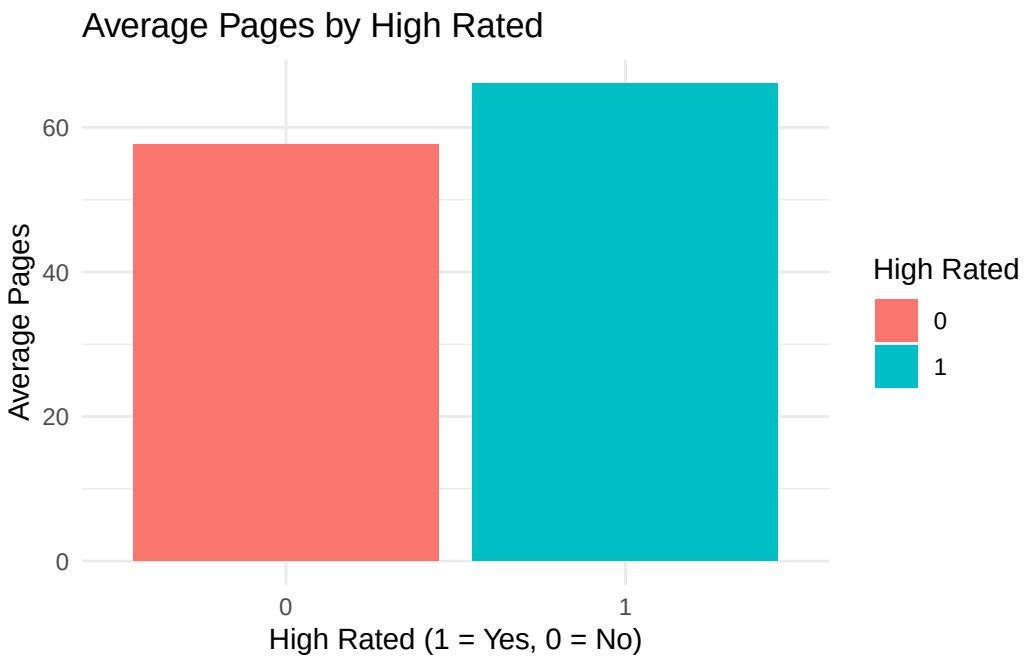


Figure 5: Summary Statistic of the Cleaned Dataset

```
`summarise()` has grouped output by 'rated_high'. You can override using the
`.groups` argument.
```

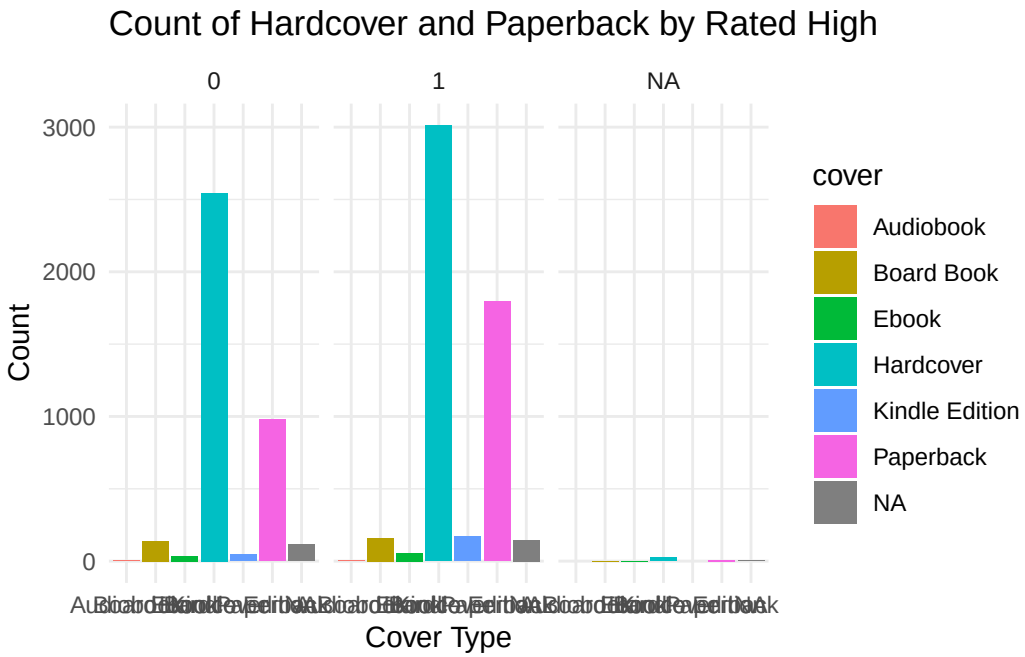


Figure 6: Summary Statistic of the Cleaned Dataset

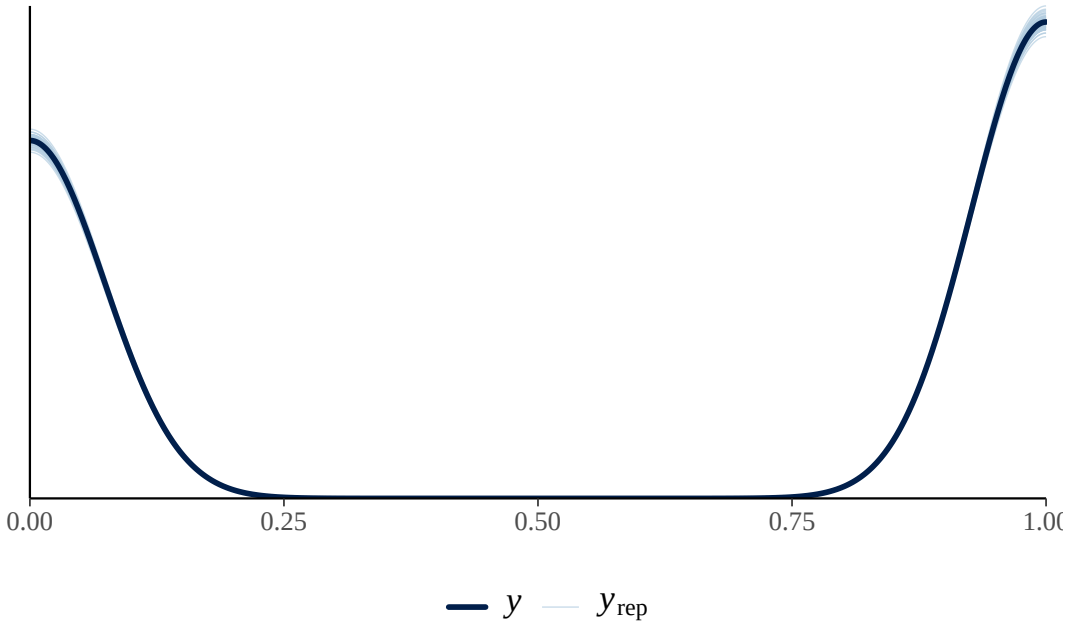


Figure 7: Summary Statistic of the Cleaned Dataset

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

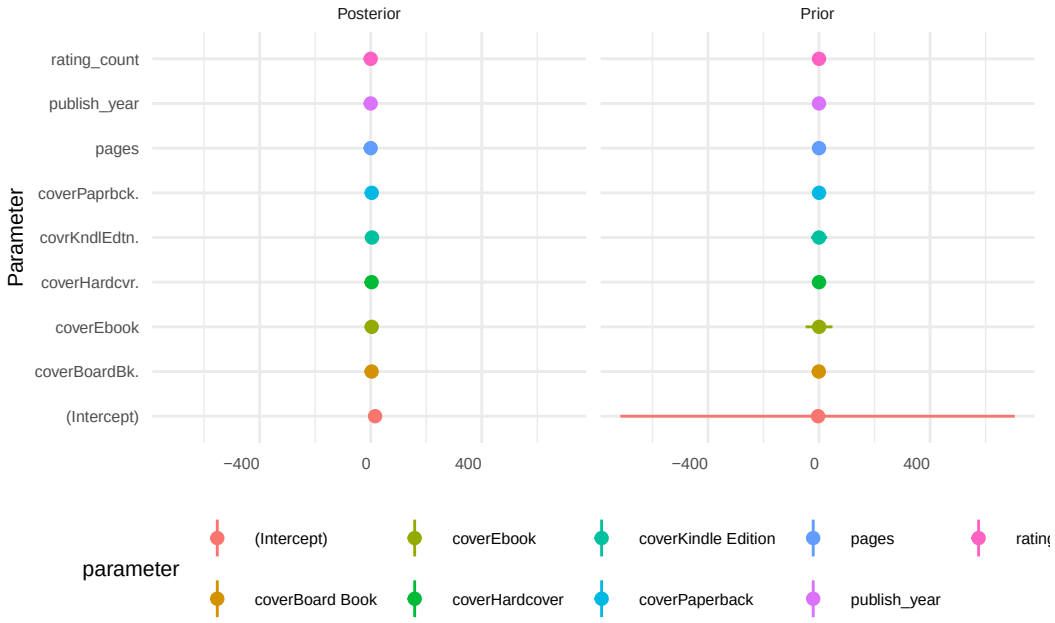


Figure 8: Summary Statistic of the Cleaned Dataset

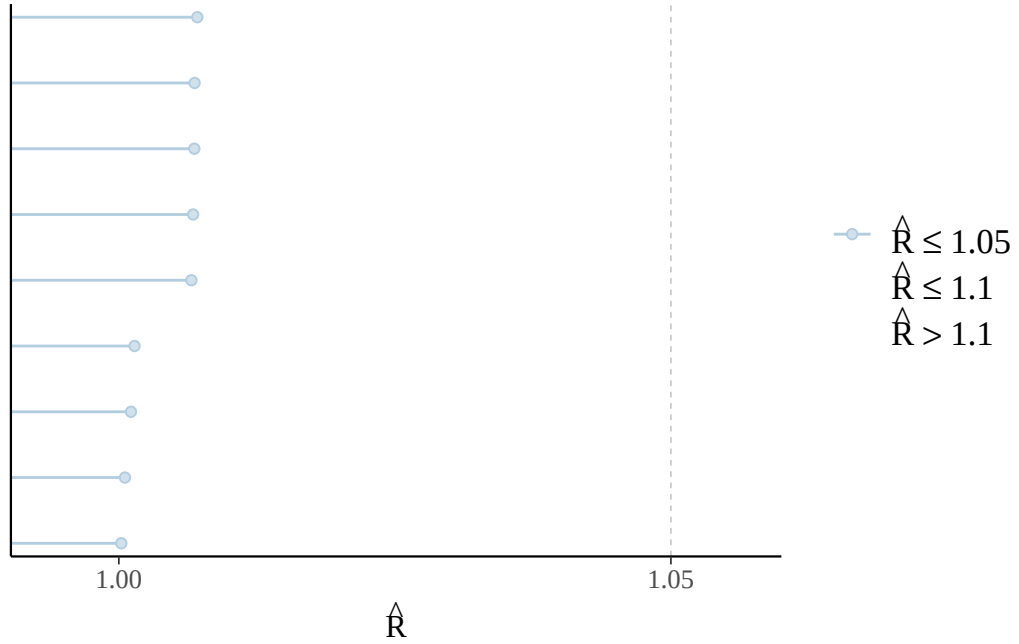
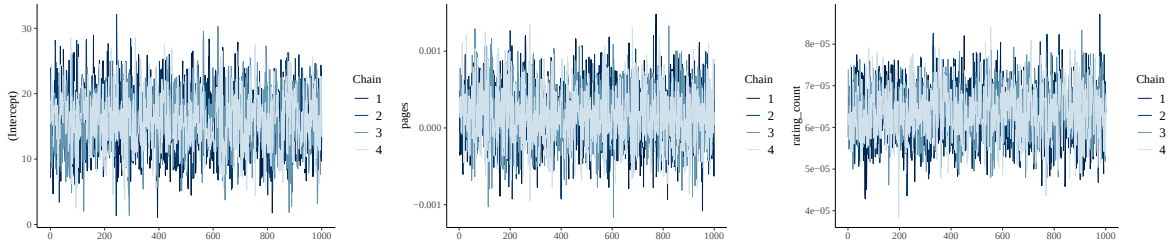
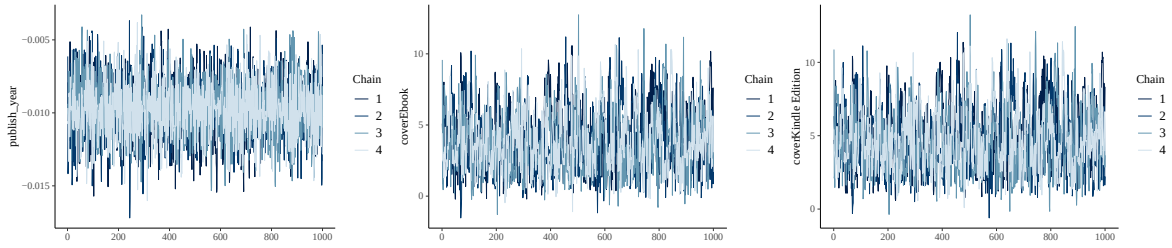


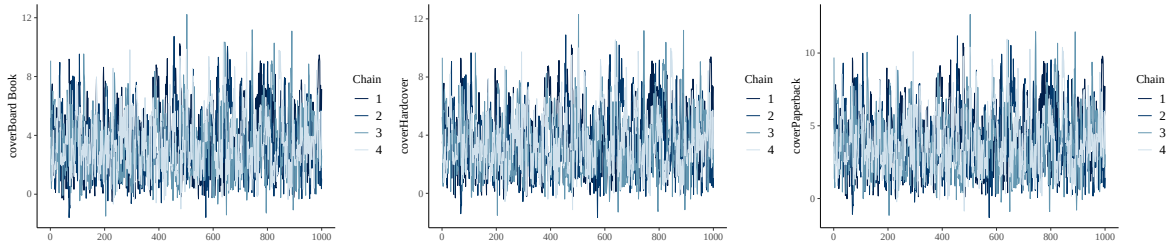
Figure 18: Summary Statistic of the Cleaned Dataset



(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