Children's Books*

What factors influence the likelihood of a book being high-rated

Hyunje Park

November 29, 2024

This paper examines the factors influencing the rating of children's books using the data from Alex Cookson (Cookson 2020). By using a Bayesian logistic regression, I analyzed how various book characteristics impact the likelihood of a book receiving a high rating. The results showed that the book's cover type had the biggest positive impact on the likelihood of a book being high-rated. The insight from this research not only clarify the key elements of high-rated children's books, but also enhance future authors to push the boundaries of their literature by exploring different writing choices to enhance reader engagement.

Table of contents

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

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 (Pulimeno, Piscitelli, and Colazzo 2020). 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 (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 categorized as "high-rated" (a score above 4)?

In this paper, I analyzed how the characteristics of a book, such as cover type, page count, publish year, and rating counts affected the likelihood of a book being "high-rated" using the Children's Book Ratings Data (Cookson 2020). 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 high-rated, based on the chosen book characteristic variables.

The logistic Bayesian logistic regression model revealed that book covers significantly influenced the likelihood of a book being rated highly, with the `Kindle Edition` format standing out as particularly impactful. These findings have practical implications for the publishing industry, offering valuable insights for authors and publishers to explore innovative ways to enhance reader engagement and elevate the literary experience.

This research paper is structured as follows: Section 2 (Data) contains an overview of the dataset and some tables and graphs used to illustrate the variables employed in this analysis. Section 3 (Model) describes and justifies the logistic regression model that was produced in this report. Section 4 (Result) highlights the result of the model, Section 5 (Discussion) discusses some of the outcomes, weaknesses, and Section 6 (Appendix) contains additional information on model details.

1.1 Estimand

The estimand of this paper is the probability that a children's book is "highly-rated" (has a rating 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 there could be variations in ratings across countries. 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 (Cookson 2020).

2 Data

The dataset used in this paper was obtained from Cookson (Cookson 2020). Cookson's dataset contained over 9,000 records of children's books that contained information of books 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 were also removed, which included variables such as book title. Then, a new variable called `rated_high` was created, which was a binary variable indicating if the book had a rating over 4 (1 if it exceeded 4, 0 else). After cleaning and creating this new variable, there were 9240 entries in the cleaned dataset.

Data analysis is performed using Statistical Programming Language R (R Core Team 2023) , along with packages `tidyverse` (Wickham et al. 2019), `palmerpenguins` (Horst, Hill, and Gorman 2020), `kableExtra` (Zhu 2021), `arrow` (Richardson et al. 2024), `readr` (Wickham et al. 2023), `rstanarm` (Goodrich et al. 2022), `modelsummary` (Arel-Bundock 2022), `dplyr` (Wickham, François, et al. 2023), `here` (Müller 2020), `ggplot2` (Wickham 2016), `knitr` (Xie 2014), `scales` (Wickham, Wilke, et al. 2023).

2.1 Data Cleaning

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 presents the first 5 rows from the cleaned dataset.

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 "high-rated" 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 representing the number of ratings the book received
- `publish_year`: An integer representing the year the book was published
- `cover`: A categorical variable representing the cover 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 3 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, with publish year of 2005 and 2003 for the “low-rated” and “high-rated” books respectively. “High-rated” books also received far more ratings on average at 8,000 ratings, more than four times that of “low-rated” books at an average of 2000 reviews per book. Lastly, `Kindle Edition` covers had the highest proportion of high-rated books out of any other cover types (as well as low-rated books). These variables are key in the analysis, and are further mentioned in Section 2.

Table 3

rated_high	avg_pages	avg_publish_year	avg_rating_count	highest_proportion_cover
0	57.63858	2005.513	1963.318	Kindle Edition
1	66.06849	2003.934	8040.772	Kindle Edition

Summary Statistic of the Cleaned Dataset

2.3 Data Analysis

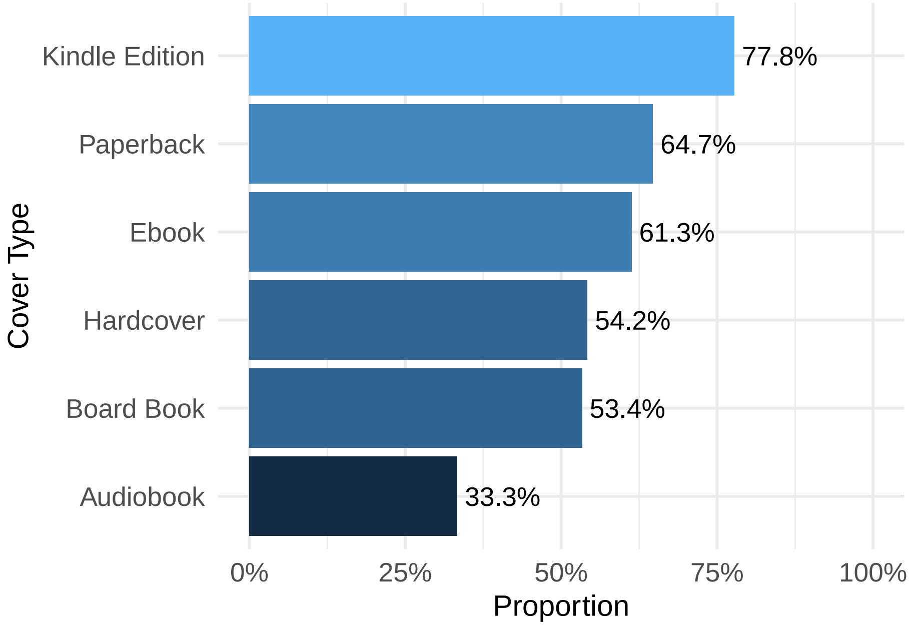


Figure 1: Proportion of High-rated Books by Cover Type

Figure 1 shows the proportion of "high-rated" books for each cover type. `Kindle Edition` books had the highest proportions of "high-rated" books at 77.8%, followed by Paperback and Ebook, at 64.7% and 61.3% respectively. This illustrates that generally, digitally formatted books had higher proportions of "high-rated" books compared to physical formats. This finding also reflects the result from Table 3.

Figure 2 shows that "high-rated" books have more page count than "low-rated" books. Highly-rated books had an average page count of 66.1, compared to the 57.6 average page count of low-rated books.

Figure 3 shows the relationship between a book's rating and the publish year. The horizontal axis shows the publish year, and vertical axis shows the rating of the book. The graph is split into two vertically, the top plot showing ratings of high-rated books and low-rated books. It showed a left-tail distribution, peaking around the mid 2010s, followed by a big drop in late 2010s. High-rated books had a lower average publish year than low-rated books, at 2003 and 2005 respectively.

Figure 4 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 "high-rated". It is apparent that books that are considered "high_rated" (right of the red line) have far more ratings on average than "low-rated" books.

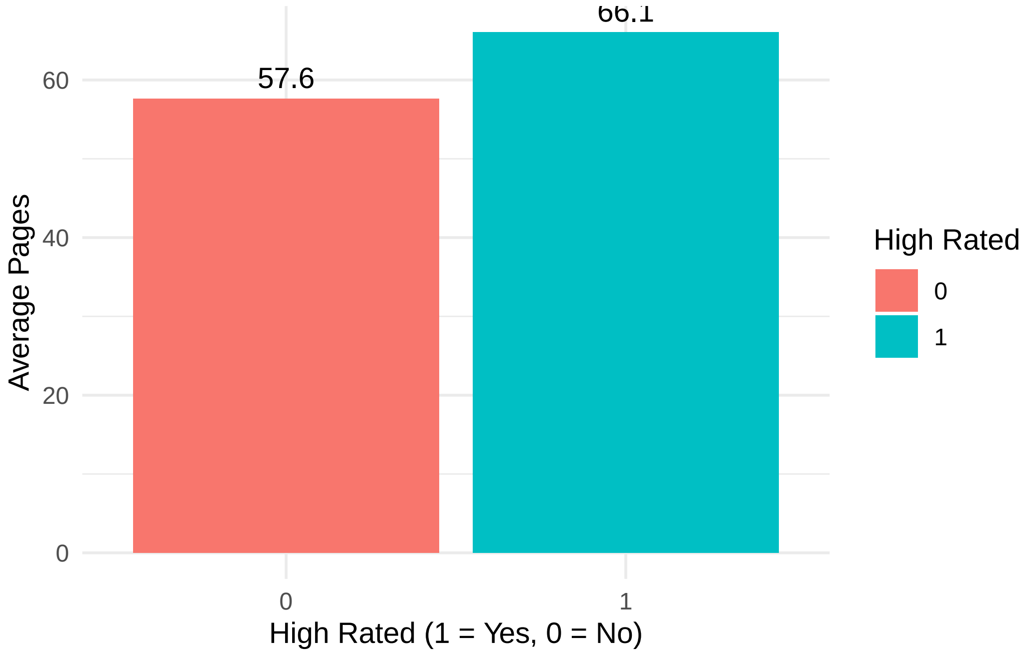


Figure 2: Average Page Counts of High-rated and Low-rated Books

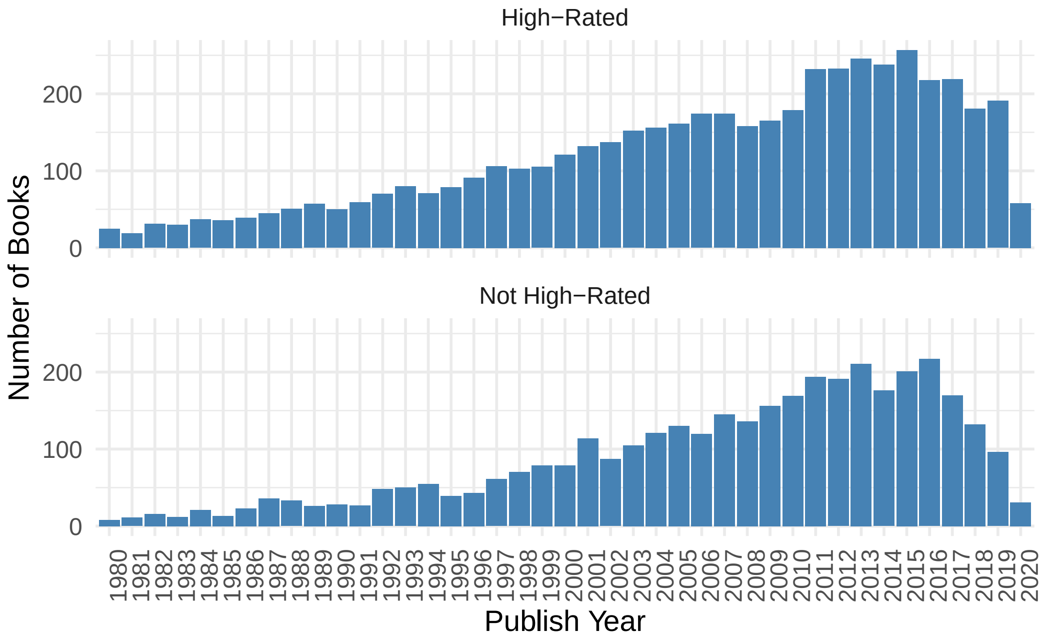


Figure 3: Number of Books by Publish Year and Rating Status

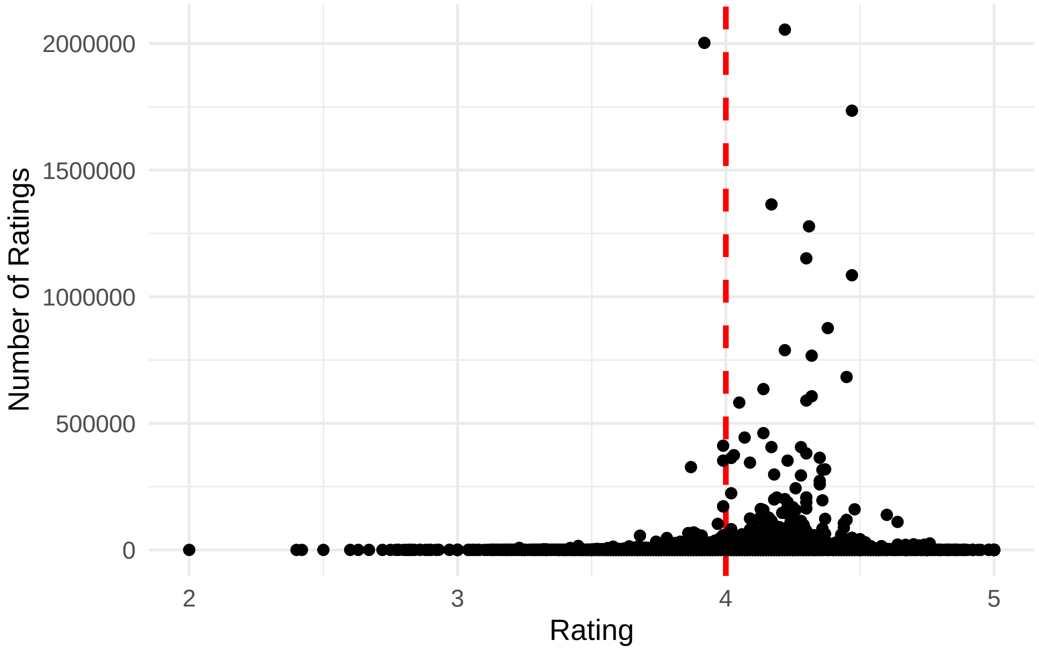


Figure 4: Rating Counts by the Rating of the Book

3 Model

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

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

3.1 Prior Distribution

The rstanarm package (Goodrich et al. 2022) was used in order to run the regression model mentioned above. The default priors of the rstanarm package was used (Goodrich et al. 2022). 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 7, 8, 9, 10, 11, 12, 13, 14, 15.

3.2 Model Justification

The Bayesian logistic regression model was chosen because it is well-suited for binary outcome variables, making it appropriate for analyzing the likelihood of books being high-rated. Additionally, Bayesian methods enables the incorporation of prior knowledge and uncertainty into the analysis, which provides more robust estimates of the model parameters. (Yu 2014)

4 Results

After running the regression based on the model shown above, we get the coefficient values showcased in Table 4.

Figure 9 represent the 90% probability intervals of the estimates, where a coefficient is considered statistically significant if the range excludes 0. In addition to this, these coefficients are expressed in log-odds, meaning it suggests a positive association with a book being high-rated when positive and low-rated when negative (Yu 2014).

With this in mind, we can see that, the Kindle Edition cover type had the highest correlation with a book being "high-rated", assuming all else held constant, reflecting the analysis that was seen in Section 2. In contrast, the number of pages of a book, and the number of ratings had no statistically significant influence on the likelihood of a book being "high-rated".

Although the publish year was not statistically significant according to the Figure 9, it exhibited a slight negative relationship (coefficient -0.01), consistent with previous findings from Figure 3 that "high-rated" books tend to have earlier publication years compared to "low-rated" ones.

5 Discussion

5.1 Rating Count and Page Count

As shown by Section 4, the number of ratings a book receives and its page length had no statistical significance on a book being "high-rated". In particular, these factors are likely independent of how readers perceive its quality or enjoyment. Short books can be impactful and concise, whereas longer books can offer richer detail.

Furthermore, the irrelevance of the number of ratings a book receives could be explained by herd effect; where readers rely on the ratings and reviews of others to make their own rating decisions (Chen 2007). Therefore, once a book has many reviews, its average rating stabilizes, making the number of ratings irrelevant to the overall evaluation.

5.2 Type of Cover

The results in Section 4, indicated a big positive relationship between book cover types and its `rated_high` status. In particular, the "Kindle Edition" book cover had the highest coefficient (4.239). A potential explanation could be that ebooks (such as Kindle Edition) are often preferred over physical books for their convenience, affordability and accessibility. Additionally ebooks are often cheaper due to saving material costs and saving physical space. In Middle-Eastern countries, 74.6% students preferred e-books in terms of easy to carry and 80.6% of them spent more time reading from e-books than printed books (Amirtharaj, Raghavan, and Arulappan 2023). Thus, it could appear that e-books have an advantage in attracting those seeking convenience and personalization in their reading, potentially explaining its big relationship compared to any other book cover types.

Table 4: Modeling the likelihood of a book being High-rated

	Book is High-Rated
(Intercept)	16.353
	(4.415)
coverBoard Book	3.007
	(1.889)
coverEbook	3.426
	(1.918)
coverHardcover	3.087
	(1.896)
coverKindle Edition	4.239
	(1.884)
coverPaperback	3.418
	(1.893)
rating_count	0.000
	(0.000)
publish_year	−0.010
	(0.002)
pages	0.000
	(0.000)
Num.Obs.	8329
R2	0.039
Log.Lik.	−5500.135
ELPD	−5519.1
ELPD s.e.	47.1
LOOIC	11 038.1
LOOIC s.e.	94.1
WAIC	11 042.2
RMSE	0.48

5.3 Publish Year

While the result in Section 4 showed that the publish year was statistically insignificant, Figure 3 showed a small negative correlation. In particular, “high-rated” books had an older publish year than “low-rated” books, and we observed with a left-tail distribution with a peak around the mid 2010s, followed by a sharp drop in the late 2010s.

The first explanation could be that older books have had more time to be appreciated, analyzed and critiqued, which could allow their work to be better established than current generation books. Many older books that receive high ratings have stood the test of time, becoming classics that made its name for generations.

The reason for the sudden drop in the late 2010s could be explained by the COVID-19 pandemic, where lockdowns could have put more emphasis on other forms of entertainment such as Netflix, TV and more.

5.4 Weaknesses and Next Steps

Regional differences were not considered in this paper. A book that appeals a reader in a country might not have the same appeal in another due to the differences in cultural context, historical experiences or social norms. Furthermore, regional differences in terms of accessibility of books (such as quality of translated versions or availability) could also impact ratings.

In terms of the dataset, the number of entries were relatively low (a little over 9,000 entries). This limited number of entries could have affected the accuracy of the results, where certain factors such as book cover types might have been significantly overrepresented or underrepresented, skewing the analysis.

Future research should aim to collect more representative data, potentially including datasets from diverse sources. Additionally, regional factors related to the book should be investigated. By exploring these regional influences, we can get an broader insight as to how cultural context, historical experience and social norms play a role in book ratings. This can help make improvements for a broader international appeal.

6 Appendix

6.1 Posterior Predictive Check

We used a posterior predictive check in Figure 5 to evaluate how well the model fits the data. This compares the observed data $\mathtt{y}$ against the replicated data y_i which was generated by our model in Section 3. The posterior distribution fitting perfectly suggests that the Bayesian logistic regression model is a good model fit.

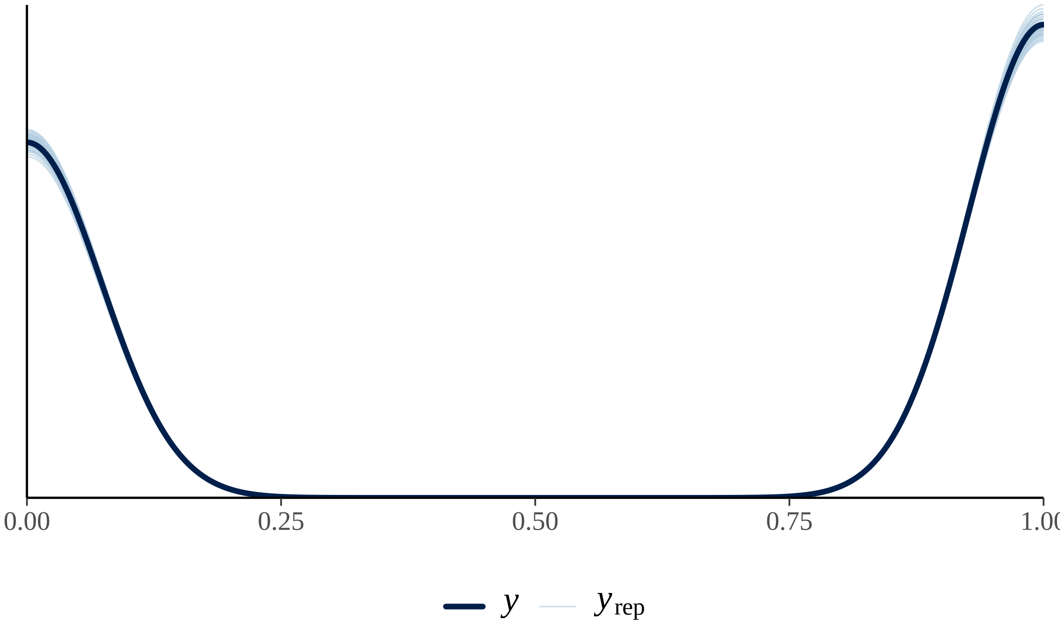


Figure 5: Posterior predictive check

6.2 Comparison of the Posterior vs Prior

In Figure 6 we compared the posterior with the prior, to examine how the estimates change once data is taken into account (Alexander 2023). Most of the variables do not vary even after data was taken into account, showing that the observed data matches the expectations of a "high-rated" book.

6.3 Markov Chain Monte Carlo Convergence Check

Figure 7 and Figure 8 are the trace plot of the model and Rhat plot of the model respectively. The trace plot shows oscillating horizontal lines, with overlaps between chains, showing no signs of issues with the model. Similarly, the Rhat plot doesn't indicate any issues with the model, as all the values remain close to 1 (Alexander 2023)

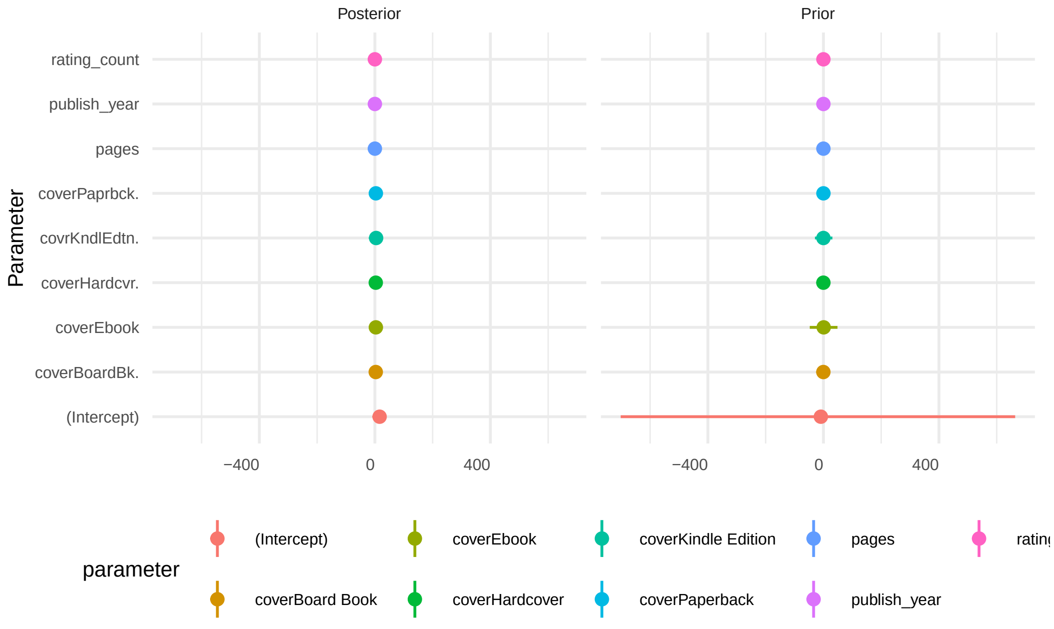


Figure 6: Comparison of the prior and posterior

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Amirtharaj, Anandhi Deva, Divya Raghavan, and Judie Arulappan. 2023. *Preferences for Printed Books Versus e-Books Among University Students in a Middle Eastern Country*. https://pmc.ncbi.nlm.nih.gov/articles/PMC10248253/#:~:text=The%20study%20concluded%20that%2074.6,is%20easy%20to%20make%20notes.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

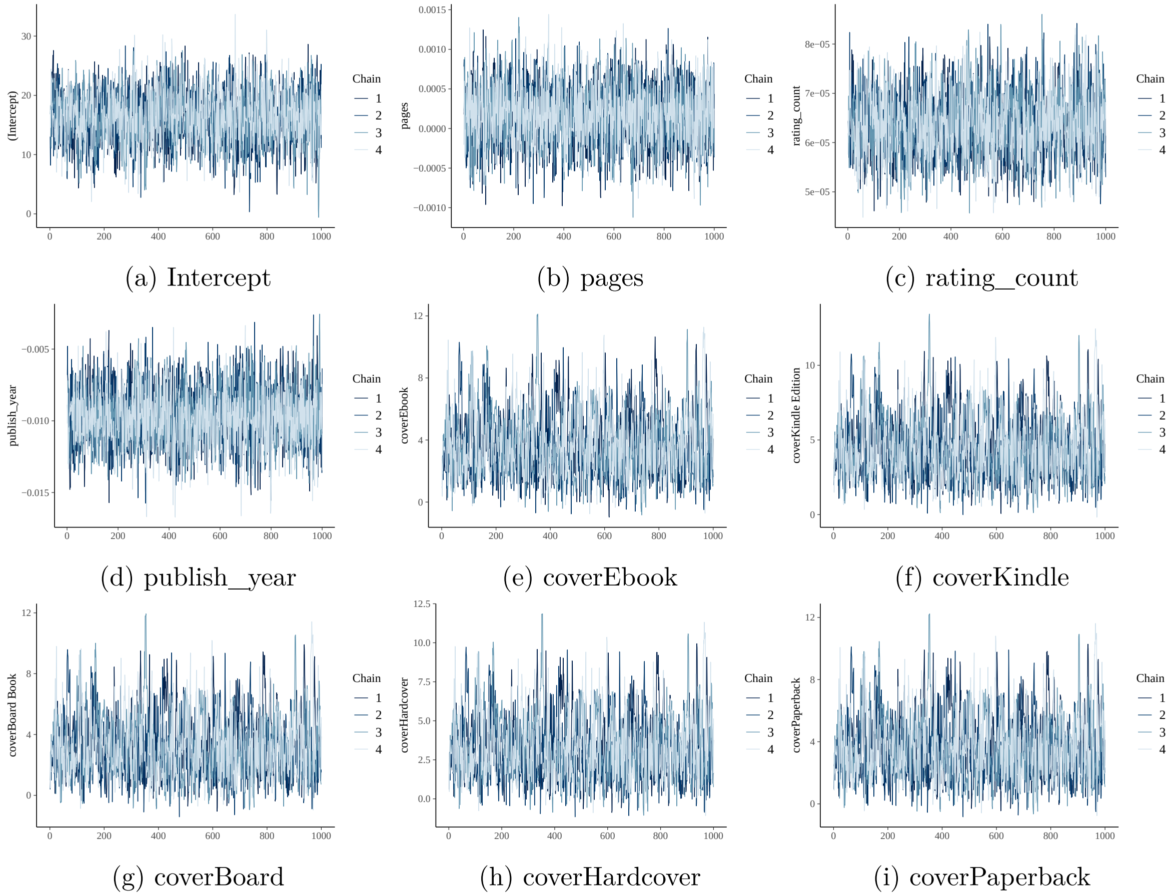
Chen, Yi-Fen. 2007. *Herd Behavior in Purchasing Books Online*. https://doi.org/10.1016/jW.chb.2007.08.004.

Cookson, Alex. 2020. *Children's Book Ratings*. https://github.com/tacookson/data/tree/master/childrens-book-ratings.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.



(a) Intercept

(b) pages

(c) rating_count

(d) publish_year

(e) coverEbook

(f) coverKindle

(g) coverBoard

(h) coverHardcover

(i) coverPaperback

Figure 7: Trace plot

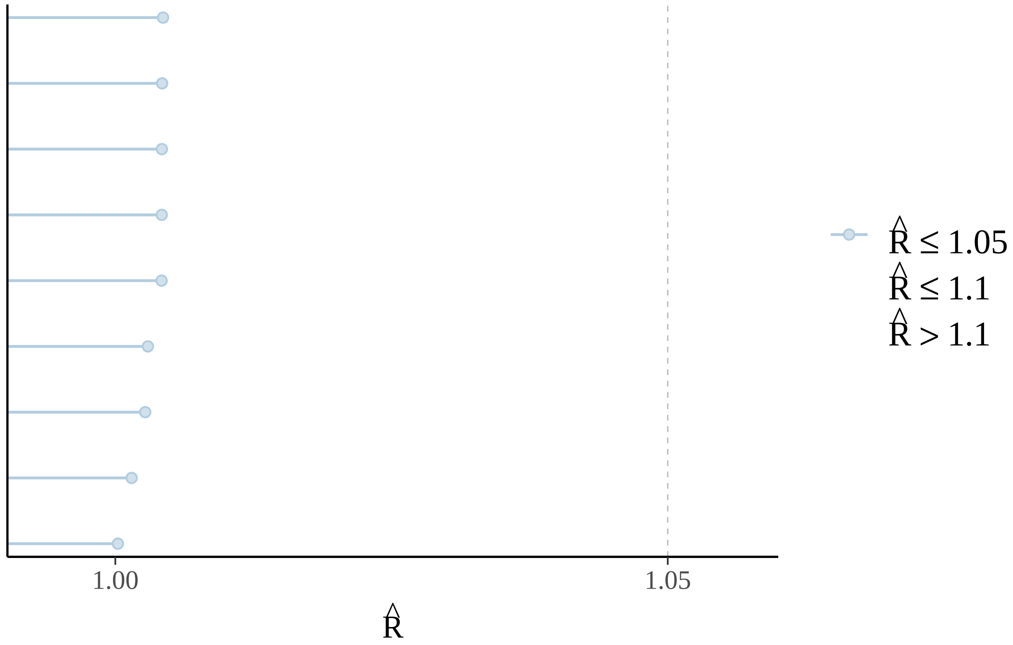


Figure 8: Rhat Plot

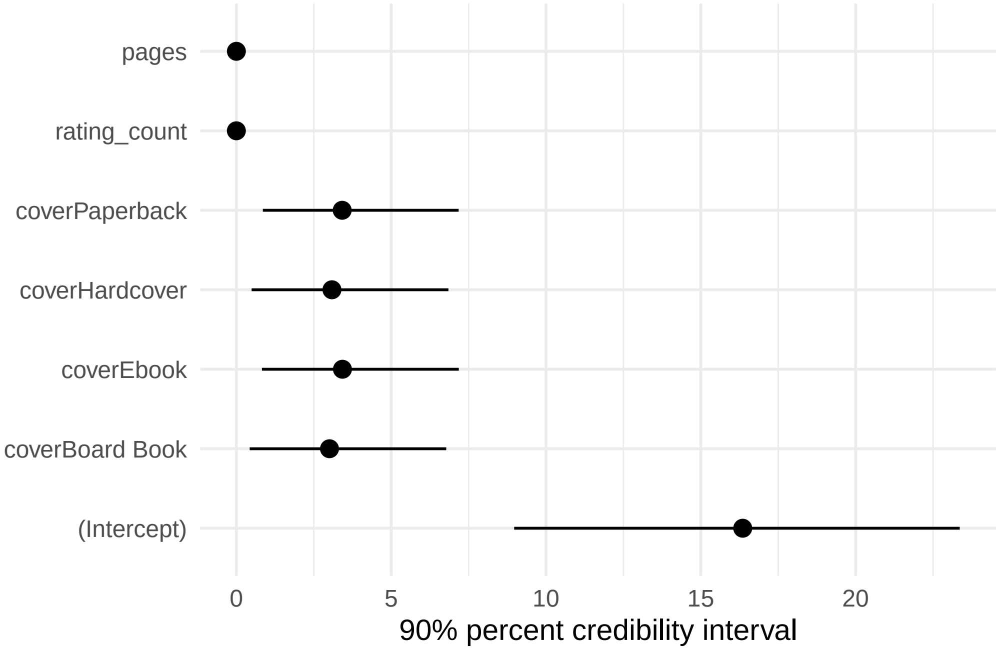


Figure 9: 90% Credibility Intervals for Predictors of High-rated Books

Pulimeno, Manuela, Prisco Piscitelli, and Salvatore Colazzo. 2020. *Children's Literature to Promote Students' Global Development and Wellbeing*. https://pmc.ncbi.nlm.nih.gov/articles/PMC7036210/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley et al. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Claus Wilke, et al. 2023. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Yu, Hannah. 2014. *Fake News Vs Fox News: The Influence of Media Preferences on Voting Behavior in the 2020 u.s. Presidential Election Among Party Voters*. https://github.com/hannahyu07/Fox-News.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.