Factors Influencing Early Childhood Comprehension of American English*

Analyzing the impact of early childhood on English comprehension

Ziyuan Shen

November 26, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Early childhood language comprehension is a fundamental area of cognitive development that influences future literacy and communication skills. Research has identified several key predictors of language comprehension in young children. For example, a study by Taylor et al.(Zubrick, Taylor, and Christensen (2015)) found that socioeconomic factors, family background, and personal characteristics significantly influence early childhood receptive vocabulary development. As highlighted by Lawton et al.(Lawton, Araujo, and Kufaishi (2023)), the quality and quantity of language exposure in infancy are associated with brain development and long-term language achievement. In addition, Bergelson's study(Zonarich (2024)) showed that babies begin to understand common nouns as early as 6 to 7 months of age, which suggests that comprehension begins earlier than previously thought. Despite these insights, I believe there is still a need for robust research that considers the range of demographic, developmental, and environmental factors that influence the early comprehension of American English learners.

In this paper, my estimand is the number of words comprehended by children learning American English. Our focus is on quantifying how developmental factors such as age, birth order, caregiver influences such as education level, and language context such as monolingual versus bilingual status affect comprehension outcomes. Using Bayesian regression modelling, I assessed the contributions of these predictors and identified the factors that have the greatest impact on early childhood comprehension.

*Code and data are available at: Factors Influencing Early Childhood Comprehension of American English.

My findings show that developmental factors such as age, production ability of language, and caregiver education are important predictors of young children's level of comprehension. Children with more educated caregivers demonstrated better comprehension, which highlights the role of the home environment in language development. These findings are important because they provide a basis for carefully targeted early childhood interventions, which highlight the areas that need to be supported in order to promote language comprehension. By identifying the most influential factors, this study contributes to a better understanding of early language learning, which is essential for the development of effective educational practices and policies.

The structure of this paper is organized as follows: after this introduction, Section 2 details the dataset used, including the data collection and cleaning processes, and provides an overview of the key variables. **?@sec-model** introduces the regression models employed in the analysis and discusses why these models are suitable for estimating comprehension outcomes. **?@sec-result** presents the results, emphasizing the relationships between different predictors and comprehension. Finally, **?@sec-discussion** concludes the paper by evaluating the implications of the findings, discussing their importance for educational policy, and outlining potential directions for future research.

2 Data

2.1 Data Overview

I used the statistical programming language R (R Core Team 2023) to retrieve, simulate, clean, analyze, and test early childhood language comprehension data. The dataset used for this analysis is from the Wordbank database of children's vocabulary growth (Frank, n.d.), specifically focusing on American English children. It contains comprehension and a variety of predictors, which include developmental, demographic, and environmental factors. Following the methodology discussed in "Telling Stories with Data" (Alexander 2023), I explore how different predictors affect language comprehension in early childhood. Also, the following packages were used in this study:(Please refer to **?@sec-data-cleaning** for detailed data processing steps.)

The dataset used in this study provides different data on the language comprehension skills of young children among American English language learners. The dataset consists of 14,826 rows and 22 columns covering a range of early childhood developmental attributes such as comprehension, output and parental education. To ensure the validity of the analysis, I filtered the data to include children whose comprehension assessments and caregiver information were complete, which would ensure a comprehensive, unbiased representation of the target population. This cleaning allowed us to focus on high-quality, well-integrated cases, which provided reliable results in young children's comprehension skills.

2.2 Data Measurement

The Early Childhood Language Development Assessment Measurement Data Set was collected using a standardized instrument such as the Communicative Development Inventory (CDI). Specifically, comprehension and vocabulary data are obtained through parent reports. These CDIs are survey instruments provided to parents or caregivers that allow them to document their child's language skills at specific developmental stages.

For example, let your child learn the meaning of different objects at home. This process naturally occurs through interaction with caregivers and is translated into quantifiable data when parents marked items on the CDI checklist that their child understood or said. The resulting scores become items in our dataset, which reflect each child's language comprehension skills in numerical form.

In the data cleaning process, we prepared the raw election data for analysis by applying transformations, filtering, and restructuring using several R packages, including `dplyr` (Wickham et al. 2023), `tidyverse` (Wickham et al. 2019), `arrow` (Richardson et al. 2024), and `rsample` (**citersample?**). The detailed cleaning processes are in **?@sec-data-cleaning**.

2.3 Outcome variables

2.3.1 Comprehension(Number of Words Understand By Children)

The primary outcome variable was the child's comprehension of American English, measured by the number of words the child understood. These data include responses from children of different ages (months) and allow for age-specific and general measures of language comprehension in early development.

Using the package ggplot2 (Wickham 2016), Figure 1 shows the distribution of comprehension, which is the number of words children understand in early childhood. Each bar in the histogram represents the count of children who comprehend a specific range of words.

The histogram shows a right-skewed distribution, with a higher frequency of children having fewer understood words. The count gradually decreases as the number of comprehended words increases, suggesting that most children in the dataset have relatively lower comprehension levels, while fewer children understand a larger vocabulary.

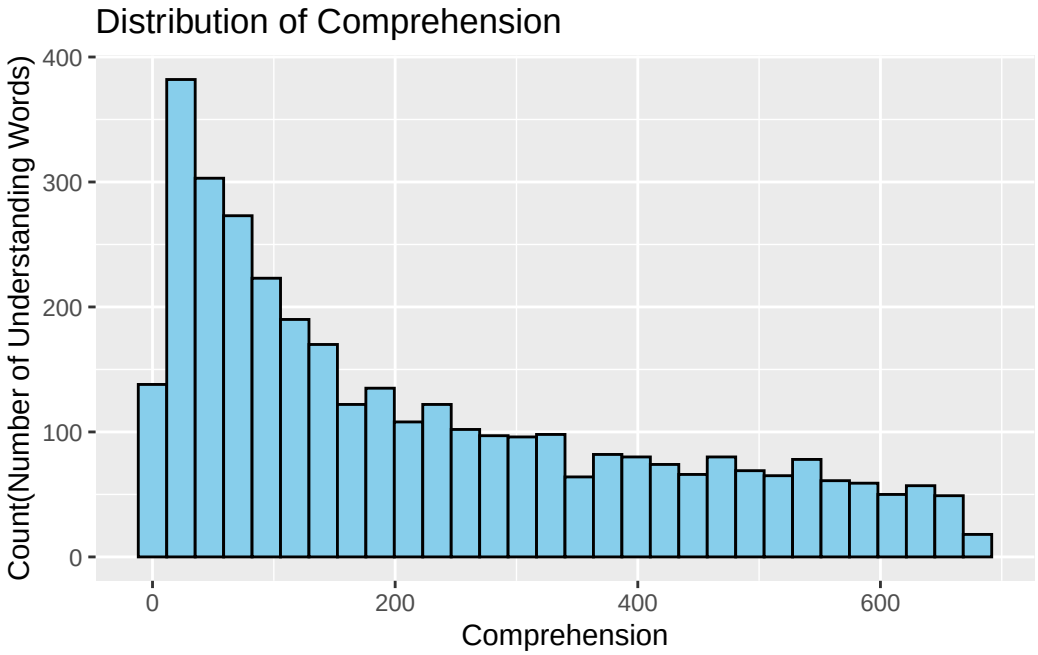


Figure 1: Distribution of comprehension among children. Each bar represents the count of children who understand a specific range of words, illustrating the varying levels of vocabulary comprehension in the dataset.

2.4 Predictor variables

There are eight predictor variables for my study after cleaning. Table 1 shows the descriptions of each predictor variable.

Table 1: Variable Descriptions

Variable	Description
age	The age of the child in months.
production	The number of words a child can produce.
is_norming	A binary variable indicating if the child is included in the norming sample (1 = Yes, 0
birth_order	The birth order of the child (e.g., 1 = first-born, 2 = second-born).
caregiver_education	The education level of the primary caregiver (e.g., Graduate, Some College).
race	The race of the child (e.g., White, Asian).
sex	The sex of the child (0 = Female, 1 = Male).
monolingual	A binary variable indicating if the child is monolingual (1 = Yes, 0 = No).

Descriptions of Each Predictor Variable

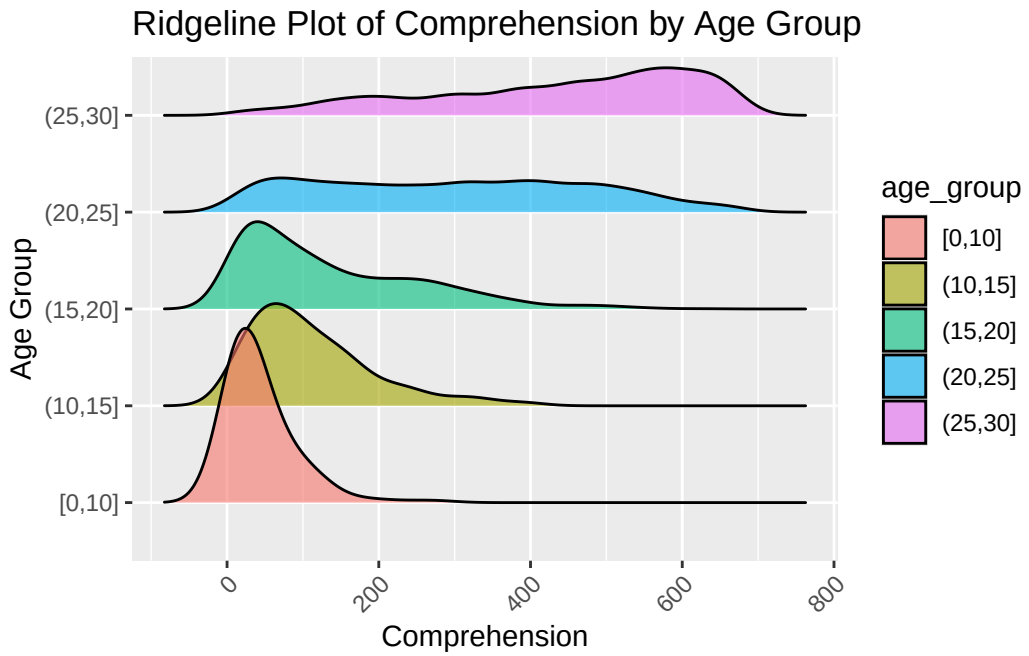


Figure 2: Ridgeline Plot of Comprehension by Age Group

Figure 2 illustrates the relationship between age and comprehension. Each grey point represents an individual observation of comprehension at a specific age, while the blue trend line provides a smoothed estimate of how comprehension changes with age.

The plot shows an upward trend in comprehension as age increases, and especially after about 10 months.There is a clear upward trend after 20 months, which suggests that children's vocabulary comprehension improves dramatically as they grow up. The shaded area around the blue line represents the confidence interval, which is an estimate of the range within which the true trend is likely to occur, taking into account the variability of the data.

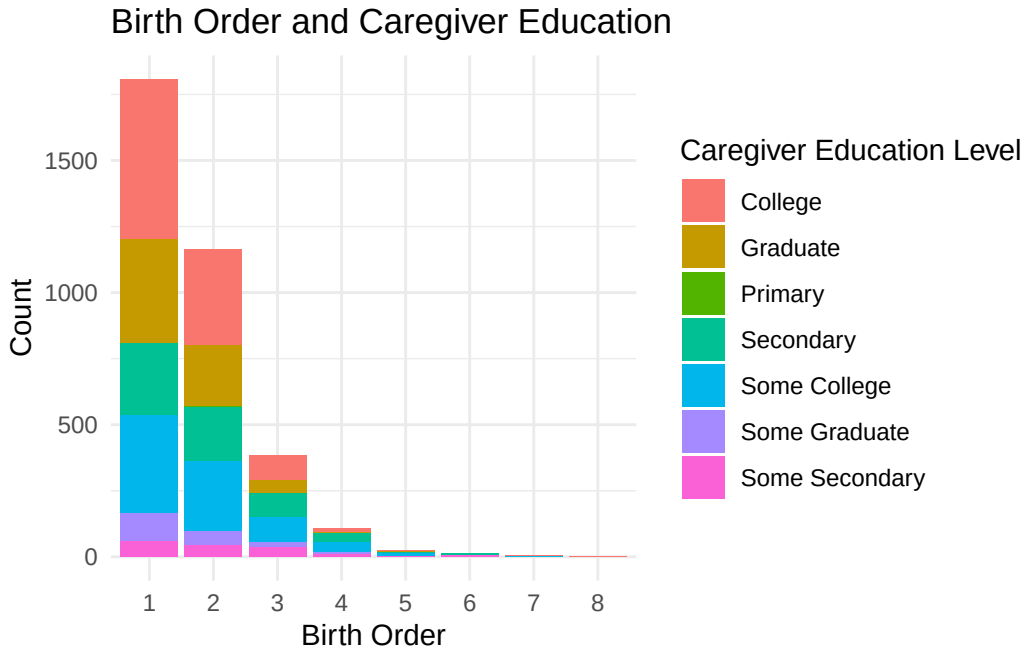


Figure 3: The relationship between age and comprehension in children. The plot includes individual observations represented by grey points, with a blue trend line and shaded confidence interval showing the general trajectory of comprehension as children age.

2.5 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

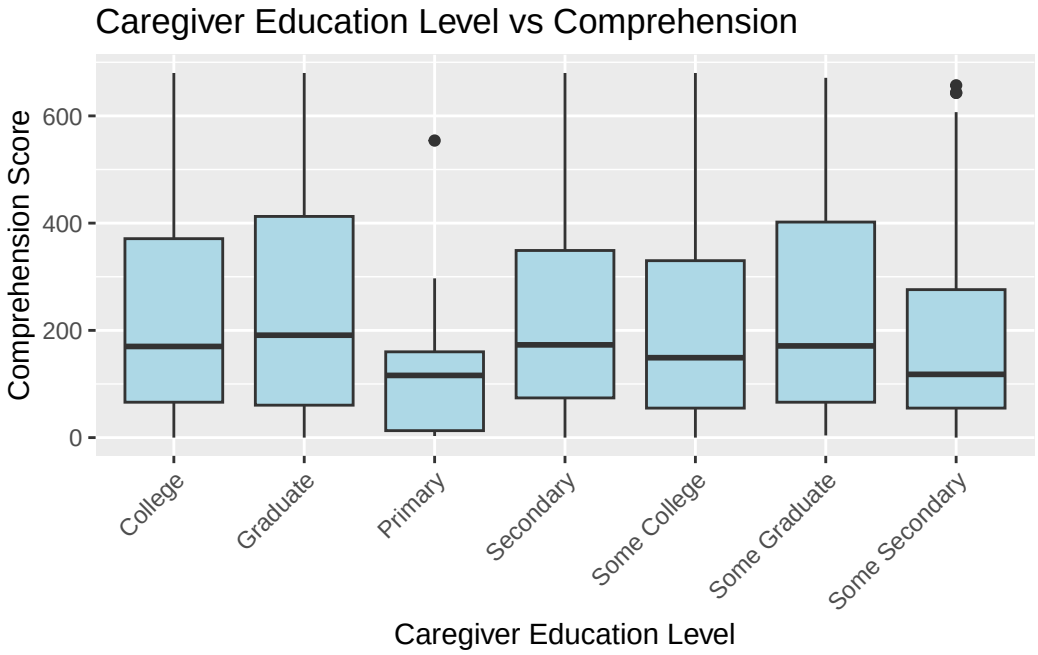


Figure 4: comprehension vs. caregiver_education

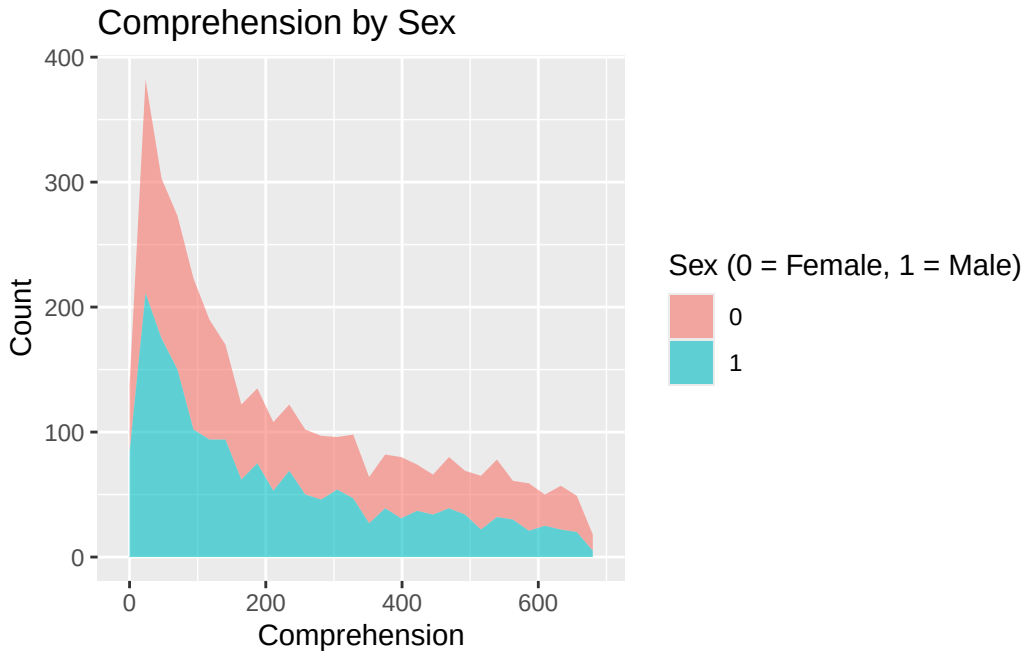


Figure 5: Comprehension by Sex

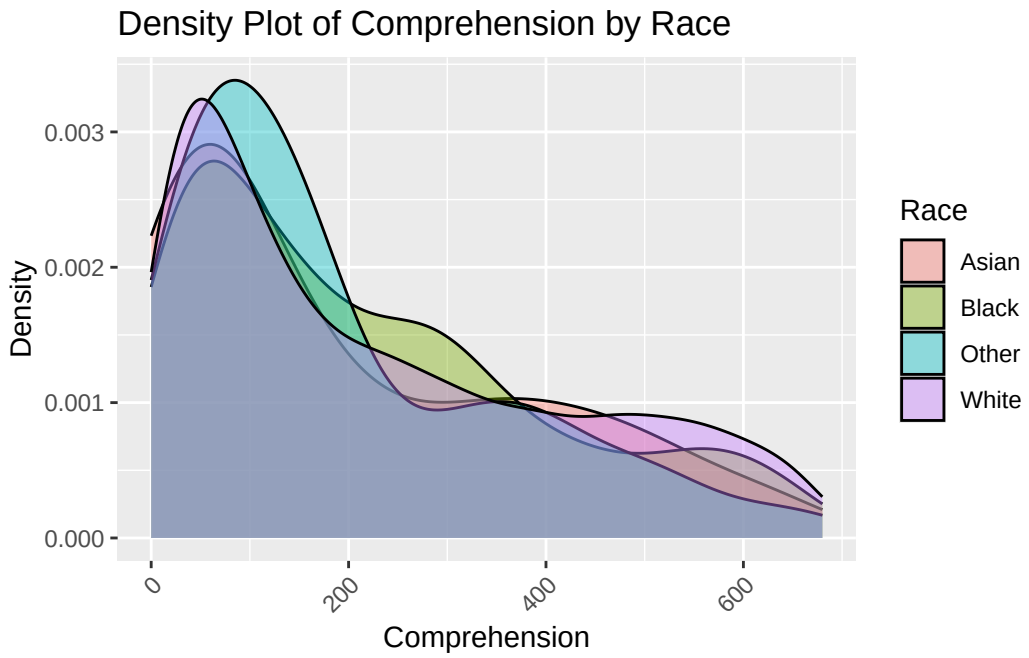


Figure 6: Comprehension by Sex

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

#| echo: false #| eval: true #| label: tbl-modelresults #| tbl-cap: "Explanatory models of flight time based on wing width and wing length" #| warning: false

modelsummary::modelsummary(list("First model" = first_model), statistic = "mad", fmt = 2)

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

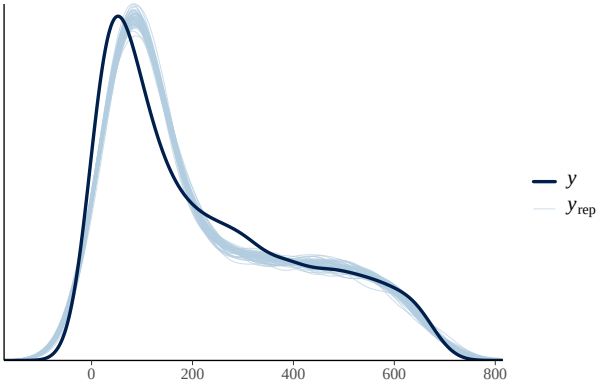
A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...



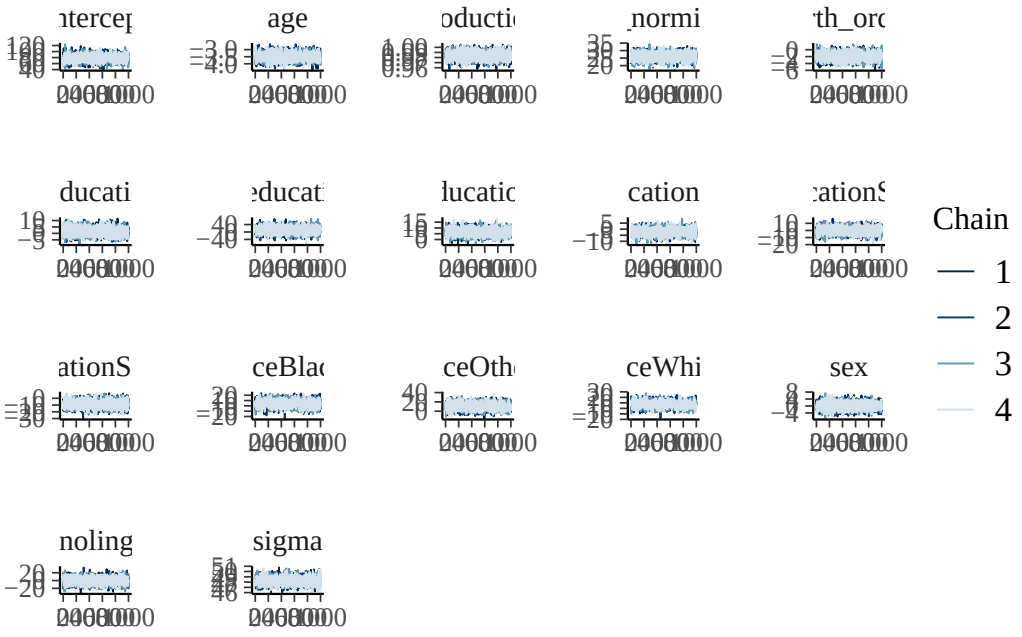
(a) Posterior prediction check

Figure 7: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...



(a) Trace plot

Figure 8: Checking the convergence of the MCMC algorithm

References

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

Frank, Mika et al., Braginsky. n.d. "Wordbank Database." *Wordbank*. https://wordbank.stanford.edu/data/?name=admin_data.

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

Lawton, Will, Ozzy Araujo, and Yousif Kufaishi. 2023. "Language Environment and Infants' Brain Structure." *Journal of Neuroscience* 43 (28): 5129–31. https://doi.org/10.1523/JNEUROSCI.0787-23.2023.

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://github.com/apache/arrow/.

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

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://dplyr.tidyverse.org.

Zonarich, Elizabeth. 2024. "Why Do Some Kids Learn to Talk Earlier Than Others?" *Harvard Gazette*, October. https://news.harvard.edu/gazette/story/2024/01/why-do-some-kids-learn-to-talk-earlier-than-others-childhood-development-linguistics/?utm_source=chatgpt.com.

Zubrick, Stephen R., Catherine L. Taylor, and Daniel Christensen. 2015. "Patterns and Predictors of Language and Literacy Abilities 4-10 Years in the Longitudinal Study of Australian Children." *PLOS ONE*, September. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0135612&utm_source=chatgpt.com#abstract0.