My title*

My subtitle if needed

First author Another author

November 20, 2024

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

Table of contents

*Code and data are available at: .

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data (Toronto Shelter & Support Services 2024)…. Following Alexander (2023), we consider…

Overview text

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (**?@fig-bills**), from Horst, Hill, and Gorman (2020).

Talk more about it.

And also planes (**?@fig-data**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.4 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.

3 Model

3.1 Model Selection

To investigate the relationship between children's vocabulary acquisition and their demographic and linguistic characteristics, we constructed a logistic regression model. By examining key demographic and linguistic predictors, we aim to identify how characteristics like age, norming status, and word categories influence vocabulary development. The dependent variable, high_vocabulary, is a binary outcome indicating whether a child's average production and comprehension score (denoted as prod_comp_mean) exceeds 350. This threshold was chosen to distinguish children with relatively advanced vocabulary levels. More background details and diagnostics are included in Appendix B.

3.2 Logistic Regression Model Overview

- High Vocabulary: The outcome variable, high_vocabulary, is a binary indicator that takes the value of 1 if the average production and comprehension score (prod_comp_mean) exceeds 350 and 0 otherwise. This threshold was selected to represent children with relatively advanced vocabulary skills, determined through exploratory data analysis.

- Scaled Age (age_scaled): This continuous variable represents the child's age, standardized to ensure the model coefficients reflect changes per standard deviation in age. Standardization improves numerical stability and aids interpretability.
- Norming Status (is_norming): A binary indicator denoting whether a child is part of the norming dataset (TRUE) or not (FALSE). This variable accounts for potential differences in data collection or assessment protocols.
- Broad Category (broad_category): A categorical variable grouping words into four broad linguistic categories: nouns, verbs, adjectives, and function_words. The reference category is function_words.

The model takes the form:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \text{age\_scaled}_i + \beta_2 \cdot \text{is\_normingTRUE}_i \tag{1}$$

$$+ \beta_3 \cdot \text{broad\_categoryNouns}_i \tag{2}$$

$$+ \beta_4 \cdot \text{broad\_categoryFunction\_words}_i \tag{3}$$

$$+ \beta_5 \cdot \text{broad\_categoryVerbs}_i \tag{4}$$

Where:

- p_i represents the probability that person i has a high vocabulary
- β_0 is the intercept, capturing the baseline log-odds when all predictors are at their reference or mean levels
- β_1: Effect of age (standardized)
- β_2: The effect of whether the individual belongs to the norming group
- β_3, β_4, β_5: The effects of being in the respective broad word categories (nouns, function words, or verbs), compared to the reference category (likely "adjectives").

3.3 Model Assumpations

- Linearity of the Logit: The model assumes a linear relationship between the log-odds of the outcome and the independent variables. Standardizing age ensures this assumption is more likely to hold.
- Independent Observations: The data points are assumed to be independent, which is valid given the dataset structure.
- Categorical Variable Encoding: The broad category variable was encoded using sum contrasts to test the deviation of each category from the overall mean effect.

3.4 Interpretation of Coefficients

The logistic regression coefficient (β) represents the change in the log-odds of the dependent variable (high vocabulary) for a one-unit change in the predictor variable, holding all other variables constant.

- Intercept (β_0): • Represents the log-odds of high vocabulary when all predictors are at their reference or mean levels. • For example, if $\beta_0 > 0$, the baseline odds of high vocabulary are greater than 50%.
- Scaled Age (β_1): Interpretation: For each one standard deviation increase in age, the log-odds of high vocabulary increase by β_1. Example: If $\beta_1 = 0.5$, then $\exp(0.5) \approx 1.65$, meaning the odds increase by 65% for every one standard deviation increase in age.
- Norming Status (β_2): If a child belongs to the norming group, the log-odds of high vocabulary increase by β_2 compared to non-norming children. If $\beta_2 = 0.1$, then $\exp(0.1) \approx 1.11$, meaning being in the norming group increases the odds of high vocabulary by 11%.
- Broad Category (β_3, β_4, β_5): The coefficients for broad_category represent the difference in log-odds compared to the reference category ("adjectives"). Interpretation:
  - β_3 (Function Words): A positive β_3 indicates higher odds of high vocabulary for function words compared to adjectives
  - β_4 (Nouns): A negative β_4 indicates lower odds of high vocabulary for nouns compared to adjectives
  - β_5 (Verbs): If $\beta_5 = 0$, there is no difference in odds between verbs and adjectives.

General Example for Broad Categories:

If $\beta_3 = 0.01$, $\exp(0.01) \approx 1.01$, meaning function words increase the odds by 1% compared to adjectives.

3.5 Model Justification

The model was trained on 80% of the dataset, with the remaining 20% reserved for testing. Model performance was assessed using confusion matrices and overall accuracy. Further evaluation included the analysis of residual deviance, AIC, and interpretation of individual coefficients.

Table 1: Explanatory models of flight time based on wing width and wing length

	Logistic model
(Intercept)	−0.10
	(0.01)
age_scaled	1.12
	(0.00)
is_normingTRUE	−0.17
	(0.01)
broad_categoryfunction_words	0.01
	(0.01)
broad_categorynouns	0.00
	(0.01)
broad_categoryverbs	0.00
	(0.01)
Num.Obs.	651 712
AIC	752 197.0
BIC	752 265.3
Log.Lik.	−376 092.486
RMSE	0.44

4 Results

Our results are summarized in .

```
      Predicted
Actual     0     1
     0 59837 25062
     1 26778 51251
```

```
[1] "Accuracy: 0.68182264558578"
```

The density plot highlights distinct patterns in predicted probabilities, with nouns and verbs showing higher peaks, indicating a stronger likelihood of high vocabulary acquisition in these categories. This suggests that broad categories contribute differently to vocabulary development, with nouns and verbs potentially playing a more significant role.

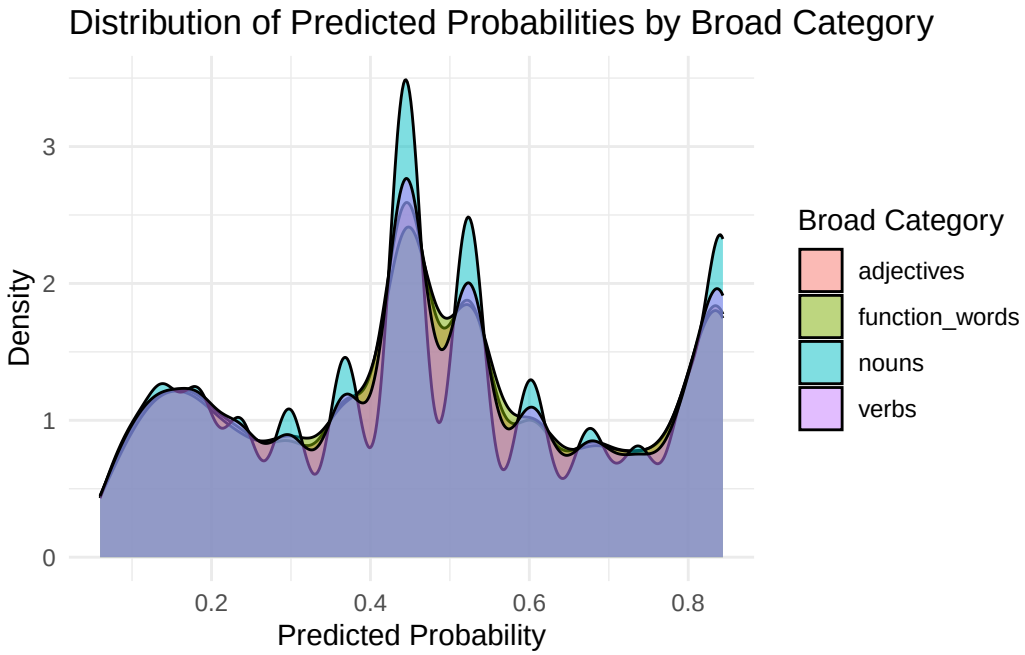

Figure 1: Distribution of predicted probabilities for high vocabulary acquisition across broad broad categories (adjectives, function words, nouns, and verbs).

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…

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…

Checking the convergence of the MCMC algo-
rithm

References

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

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

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

Toronto Shelter & Support Services. 2024. *Deaths of Shelter Residents*. https://open.toronto.ca/dataset/deaths-of-shelter-residents/.