hw8_GohBram

Bram Goh

2024-12-02

Table of contents


```
library(here)
library(readxl)  # for reading excel files
library(modelsummary)  # for summarizing data
library(cmdstanr)  # use two cores
library(posterior)
library(bayesplot)
library(brms)
library(tidyverse)
```

Research Question

> Do statistical regularities in language production (specifically, content word ratio and combination ratio) predict performance on high predictability SPiN items?

Table 1: Descriptive statistics

	Unique	Missing Pct.	Mean	SD	Min	Median	Max	Histogram
content_word_ratio	211	0	0.6	0.0	0.5	0.6	0.7	
combination_ratio	211	0	0.8	0.1	0.7	0.8	1.0	
Highspin_successes	20	0	15.3	4.3	0.0	16.0	22.0	
Highspin_trials	4	0	24.9	0.4	22.0	25.0	25.0	

	content_word_ratio	combination_ratio	Highspin_successes	Highspin_trials
content_word_ratio	1	.	.	.
combination_ratio	−0.13	1	.	.
Highspin_successes	−0.21	0.04	1	.
Highspin_trials	−0.04	0.01	0.09	1

Variables

- `Highspin_successes`: grouped number of successes on high predictability SPiN trials.
- `Highspin_trials`: number of high predictability SPiN trials
- `content_word_ratio`: average ratio of content words (nouns, verbs, adjective, and adverbs) to total words per sentence
- `combination_ratio`: average ratio of trigrams to total words per sentence

Data Import

```
frog <- read.csv("processed_frog_data_ver2.csv")
frog[1] <- NULL
frog <- frog %>% select(c(qualtrics_id, content_word_ratio, combination_ratio, Highspin_succe
```

Variable Summary

Table Table 1 shows the summary statistics for and Pearson's correlations between the variables of interest.

```
datasummary_skim(frog)
datasummary_correlation(frog)
```

Model

Let Y_i = `Highspin_successes`, N_i = `Highspin_trials`, X_1 = `content_word_ratio`, X_2 = `combination_ratio`

Model:

$$Y_i \sim Bin(N_i, \mu_i)$$
$$log(\frac{\mu_i}{1-\mu_i}) = \eta_i$$
$$\eta_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2$$

Prior:

$$\beta_0 \sim t_4(0,1)$$
$$\beta_1 \sim t_4(0,1)$$
$$\beta_2 \sim t_4(0,1)$$
$$\beta_3 \sim t_4(0,1)$$

Analysis

We used 4 chains, each with 4,000 iterations (first 2,000 as warm-ups).

Results

As shown in the rank histogram in Figure 1 below, the chains mixed well.

```
as_draws(m_logitlink) |>
    mcmc_rank_hist(pars = c("b_Intercept", "b_content_word_ratio", "b_combination_ratio", "b_
```

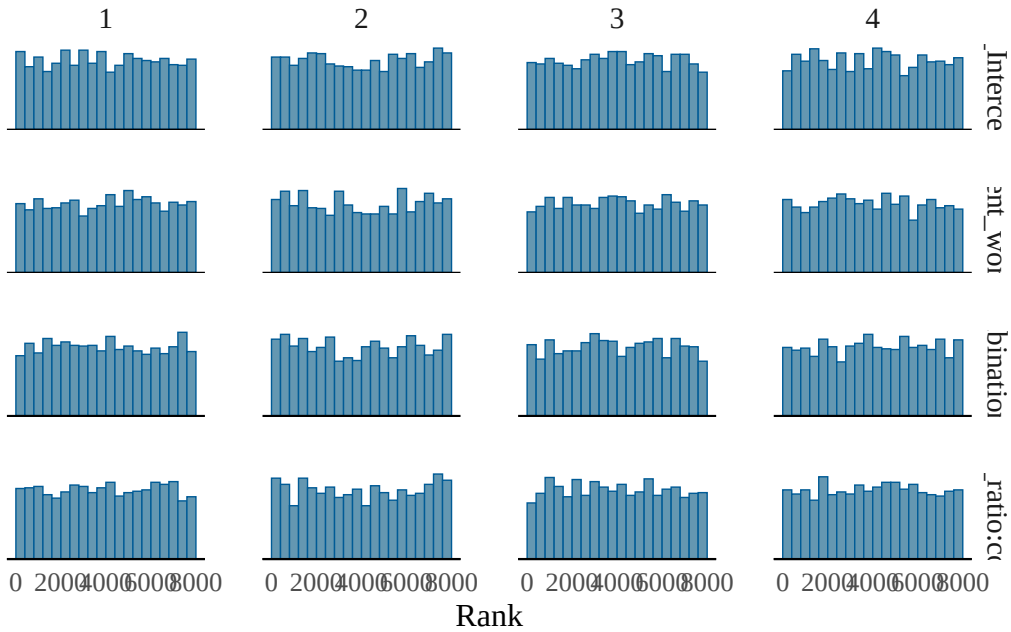


Figure 1: Rank histogram of the posterior distributions of model parameters.

Table 3 shows the summary output for β_0, β_1, β_2, and β_3, as well as model performance statistics.

```
msummary(m_logitlink, estimate = "{estimate} [{conf.low}, {conf.high}]",
         statistic = NULL, fmt = 2)
```

```
Warning:
`modelsummary` uses the `performance` package to extract goodness-of-fit
statistics from models of this class. You can specify the statistics you wish
to compute by supplying a `metrics` argument to `modelsummary`, which will then
push it forward to `performance`. Acceptable values are: "all", "common",
"none", or a character vector of metrics names. For example: `modelsummary(mod,
metrics = c("RMSE", "R2")` Note that some metrics are computationally
expensive. See `?performance::performance` for details.
 This warning appears once per session.
```

```
pp_check(m_logitlink)
```

Table 2: Posterior summary of the model with convergence statistics.

```
m_logitlink
```

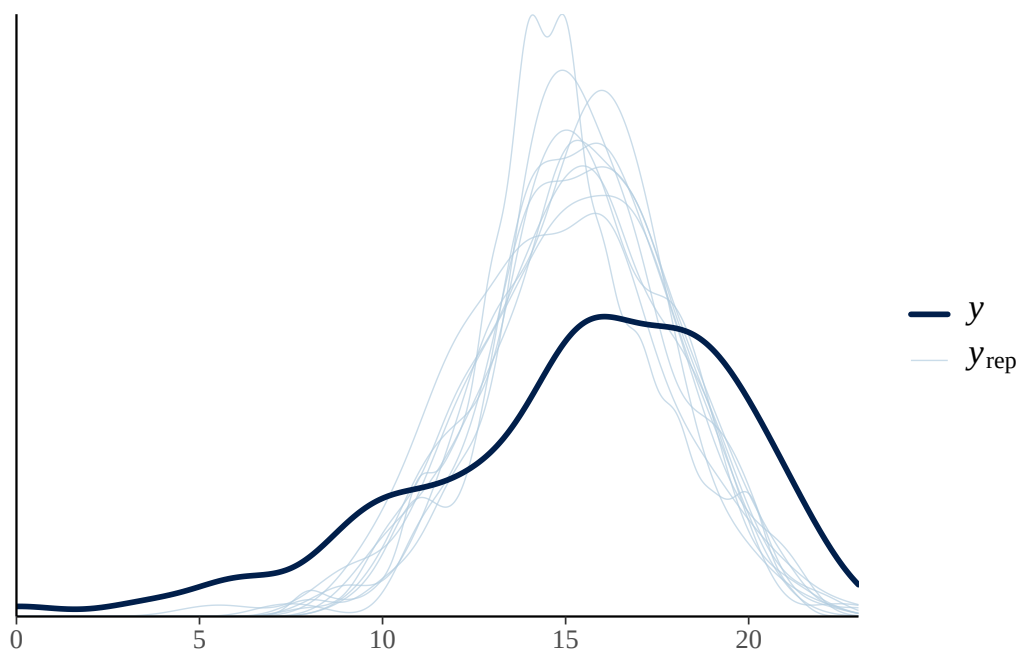
```
 Family: binomial
  Links: mu = logit
Formula: Highspin_successes | trials(Highspin_trials) ~ content_word_ratio * combination_rati
   Data: frog (Number of observations: 211)
  Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
         total post-warmup draws = 8000

Regression Coefficients:
                                     Estimate Est.Error l-95% CI u-95% CI Rhat
Intercept                                2.00      1.16    -0.17     4.38 1.00
content_word_ratio                      -2.94      1.83    -6.73     0.57 1.00
combination_ratio                        0.94      1.27    -1.64     3.46 1.00
content_word_ratio:combination_ratio    -1.36      2.03    -5.39     2.78 1.00
                                     Bulk_ESS Tail_ESS
Intercept                                3075     3472
content_word_ratio                       2921     3271
combination_ratio                        2915     3059
content_word_ratio:combination_ratio     2860     2977

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Table 3: Posterior summary of the model estimates.

	(1)
b_Intercept	1.97 [−0.17, 4.38]
b_content_word_ratio	−2.87 [−6.73, 0.57]
b_combination_ratio	0.96 [−1.64, 3.46]
b_content_word_ratio × combination_ratio	−1.37 [−5.39, 2.78]
Num.Obs.	211
R2	0.044
ELPD	−707.8
ELPD s.e.	38.3
LOOIC	1415.7
LOOIC s.e.	76.7
WAIC	1415.6
RMSE	4.20



“‘

The analysis showed that on average, the content word ratio (M = -2.87, 90% CI [-6.73, 0.57])

and combination ratio (M = 0.96, 90% CI [-1.64, 3.46]) were not significantly associated with performance on high predictability SPiN items. The interaction effect was also non-significant (M = -1.37, 90% CI [-5.39, 2.78]).