Online appendix to "Bayesian multilevel modeling and its application in comparative journalism studies"

Online appendix to "Bayesian multilevel modeling and its application in comparative journalism studies"

Source Code

All source will be available on the author(s)' Github upon acceptance. The following are key code chunks used in the paper.

Example 1

Estimation of prior. Determination of prior based on the level of significance.

```
beta_d_1 <- 0.15
sig_level <- 0.001
beta_d_1 / abs(qnorm(sig_level / 2))
```

Full code. The cubic relationship can be set using the `I()` function.

```
require(tidyverse)

informative_prior <- c(prior_string("normal(0, 1)", class = "b"),
                       prior_string("normal(0, 1)", class = "Intercept"),
                       prior_string("normal(0.14, 0.045)",
                                    class = "b", coef = "EXPRNCE"),
                       prior_string("normal(0.07, 0.027)",
                                    class = "b", coef = "RANK"))

std <- function(x){
  (x - mean(x, na.rm = TRUE))/sd(x, na.rm = TRUE)
}
```

```
## eiu is Index of Democracy, gdppc is GDP per capita (log)
set.seed(12121)
wjs %>%
    select(ppa, EXPRNCE, RANK, GENDER, UNI_EDU, UNI_EDU, eiu, gdppc, COUNTRY) %>%
    mutate(gdppc = log(gdppc), COUNTRY = as.factor(COUNTRY)) %>%
    mutate_at(vars(ppa:gdppc), std) -> wjs_processed
brm(ppa ~ EXPRNCE + RANK + GENDER + UNI_EDU + eiu + gdppc + (1|COUNTRY),
    data = wjs_processed, cores = 4, iter = 3000, prior = informative_prior,
        sample_prior = TRUE) -> ppa_mod


set.seed(12121)
## iter needs to set larger to ensure convergence
brm(ppa ~ EXPRNCE + RANK + GENDER + UNI_EDU + I(eiu^3) + gdppc + (1|COUNTRY),
    data = wjs_processed, cores = 4, iter = 4000, prior = informative_prior,
    sample_prior = TRUE) -> ppa_qmod
```

Example 2

Expression of outlet-level models in lme4 and brms. As below

```
require(brms)
require(lme4)
# brms
brm(z ~ offset(log(n)) + (1|k)+log(x),
          family = negbinomial(), data = outlet_data)
# lme4
glmer.nb(z ~ offset(log(n)) + (1|k) + log(x), data = outlet_data)
```

Expression of disaggregated outlet-level model in MASS. As below

```
require(MASS)
glm.nb(z ~ offset(log(n)) + log(x), data = outlet_data)
```

Expression of article-level models in lme4 and brms

As below

```
brm(y~(1|k/j)+log(x), family = bernoulli(), data = article_data)
glmmer(y~(1|k/j)+log(x), family = binomial(), data = article_data)
```

Expression of disaggregated article-level model in MASS. As below

```
glm(y ~ log(x), family = binomial(), data = article_data)
```

Setting up weakly informative prior in the brms model. As below

```
weaklyinformative_prior <- c(prior_string("normal(0, 1)",
                                          class = "b"),
                             prior_string("normal(0, 1)",
                                          class = "Intercept"))
# The sample_prior argument is optional. But useful for further analysis.
import_brms <- brm(z~offset(log(n))+(1|k)+log(x_import),
          data = outlet_data, family = negbinomial(),
          prior = weaklyinformative_prior, sample_prior = TRUE)
import_brms
```

Finger-like shape of Example 1 and multivariate regression

The finger-like shape is a manifestation of analyzing the two ordinal variables (two Likert items) as if they were in an interval scale. The PPC shows the original analytic scheme of both Reich and Hanitzsch (2013) and Hamada (2021) cannot capture the shape of the data. Also, the psychological distances between two neighboring options in a Likert scale (e.g. "complete freedom" and "a great deal of freedom") are not the same across all pairs and one can't assume that Likert items can be summed or averaged (Bürkner & Vuorre, 2019).

If this were not a replication and one didn't need to accept the previous operationalization of perceived professional autonomy, a better way to answer this research question using the WJS data is to make two modifications. First, one should use the ordinal regression technique to model the responses from Likert items (Bürkner & Vuorre, 2019). Second, the two Likert items related to perceived professional autonomy can be modeled jointly using a multivariate regression model.

As an exploratory analysis, we fitted a Bayesian model with these two modifications (see Online Appendix). And indeed the posterior predictive check of the model suggests that the modified model is better at pinning down the data-generating process of the Likert items than the unmodified model: the new model correctly pins down both the range, peaks and shape of the original data (Figure 1). The bottom line is the same: democratic performance of a country is still a strong predictor of journalists' perceived professional autonomy in the modified model, although the magnitude of influence is not as great as shown by the unmodified model displayed in table 1.

As below

```
brm(mvbind(C9, C10) ~ EXPRNCE + RANK + GENDER + UNI_EDU + eiu + gdppc + (1|COUNTRY),
    family=cumulative("logit"), cores = 4, iter = 3000, sample_prior = TRUE,
```

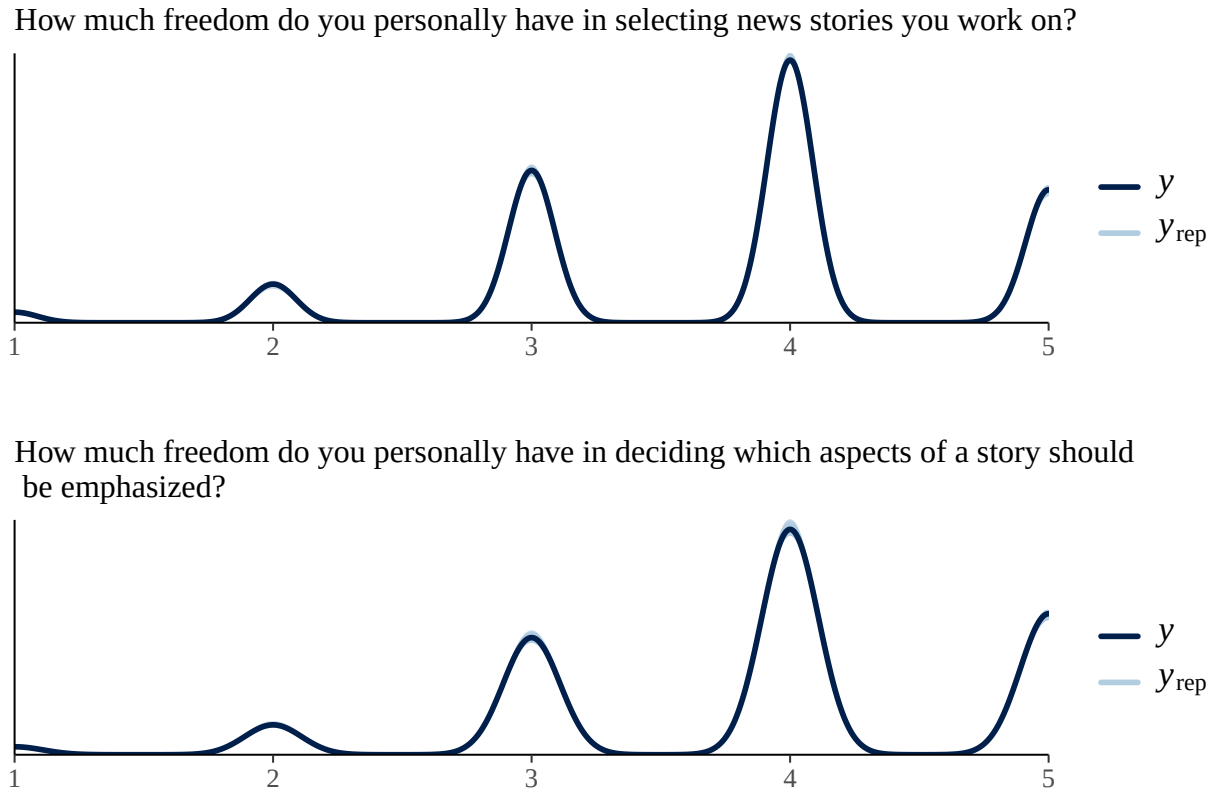

Figure 1. Posterior predictive checks with 100 sets of simulated data based on the multivariate ordinal model

```
    data = wjs, control = list(adapt_delta = 0.99), prior = informative_prior)
```

Results

Parameter	Median	95% CI	pd	Rhat	ESS
Intercept(1)	-3.49	(-4.63, -2.26)	100%	1.004	656.00
Intercept(2)	-1.81	(-2.94, -0.58)	99.85%	1.003	654.00
Intercept(3)	0.14	(-1.00, 1.37)	59.77%	1.003	651.00
Intercept(4)	2.29	(1.16, 3.53)	100%	1.003	652.00
EXPRNCE	0.33	(0.28, 0.38)	100%	1.000	4492.00
RANK	0.37	(0.33, 0.40)	100%	1.001	5186.00

Parameter	Median	95% CI	pd	Rhat	ESS
GENDER	-0.27	(-0.32, -0.22)	100%	1.000	5320.00
UNI_EDU	-0.13	(-0.20, -0.06)	100%	1.001	5312.00
eiu	0.21	(0.10, 0.32)	99.98%	1.006	402.00
gdppc	-0.09	(-0.23, 0.06)	87.93%	1.004	715.00
Intercept(1)	-3.02	(-4.21, -1.88)	100%	1.004	749.00
Intercept(2)	-1.31	(-2.48, -0.19)	99.02%	1.004	745.00
Intercept(3)	0.59	(-0.58, 1.72)	85.02%	1.004	745.00
Intercept(4)	2.70	(1.53, 3.82)	100%	1.004	746.00
EXPRNCE	0.36	(0.31, 0.41)	100%	1.000	5361.00
RANK	0.35	(0.31, 0.39)	100%	1.000	5773.00
GENDER	-0.26	(-0.31, -0.22)	100%	1.001	5402.00
UNI_EDU	-0.06	(-0.12, 0.01)	93.92%	1.000	5609.00
eiu	0.24	(0.13, 0.36)	99.97%	1.007	694.00
gdppc	-0.05	(-0.20, 0.09)	77.07%	1.004	762.00

Example 2: Validation of the seed dictionaries

One might argue our seed dictionaries are not validated. We deem that the criterion validity of our dictionary should be built-in. Coverage of China, by definition, should mention words about China. We also maintain that it is difficult to conduct a human validation of the seed dictionaries because the useNews dataset (Puschmann & Haim, 2020) doesn't provide access to the full-length articles. Nonetheless, we attempted to conduct such validation by randomly selecting 10 matching and 10 non-matching items for each language. In total, 160 items were selected. A coder with no knowledge of the ground truth (matching or not) annotated the full-text articles by following the URLs in the metadata provided in the useNews dataset (Puschmann & Haim, 2020). For some languages, the

coder used Google Translate to translate the full-text article to English. In total, only 138 URLs were still accessible. Comparing the ground truth and the human coding, 62 and 70 are true-positive and true-negative cases. There are also 5 false-negative and 1 false-positive cases. There are several reasons not related to our seed dictionaries that could lead to the 6 misclassified cases. The false-positive case is Korean and that could be due to incorrect tokenization at the Media Cloud's end. The 5 false-negative cases are in Norwegian, Portuguese, and Romanian. The content on the website might have changed since Media Cloud collected the data. For example, the ever-changing content in the sidebar might be counted as "main content" and recorded in the DTM. Despite all these caveats, our seed dictionaries have a precision and recall of 98.4% and 92.5% respectively. For the purpose of demonstrating regression analysis, this level of accuracy is more than sufficient.

Extensions to the useNews example

Extension #1: Hypothesis testing

Under the Bayesian framework, we can test hypotheses about the parameter estimates (Shikano, 2019). These hypotheses are similar to the null hypothesis in the Neyman-Pearson's sense. For example, we want to test the two-sided hypothesis of $\gamma_{01} = 0$ for Model A1. For this, we need to calculate the posterior probability of this hypothesis: $P(\gamma_{01} = 0|X)$. The function *hypothesis()* can be used to test such hypothesis.

```
hypothesis(import_brms, "logx_import = 0")
```

The posterior probability of this hypothesis is 0. Therefore, we can safely reject the hypothesis and conclude that $\gamma_{01} \neq 0$, i.e. log import volume of the outlet's country is associated with the frequency of China coverage at the outlet level.

While accepting a null hypothesis is not a feasible option in a frequentist framework, as such a framework has an "inability to gain evidence for the null" (Rouder, Speckman,

Sun, Morey, & Iverson, 2009, p. 226) hypothesis, it is practically possible to accept a null hypothesis in a Bayesian framework (Kruschke, 2018). Bayes factor or the region of practical equivalence (ROPE) could be used to decide to accept the null hypothesis. Kruschke (2018), for example, suggests using $\sigma = \pm 0.1$ as half of small effect size to establish a ROPE. However, there is no fixed rule on how to define a ROPE. If now 95% of the HDI for an estimate falls within the ROPE, researchers could accept the null hypothesis for practical purposes. A practical application could be with manual content analysis or survey research in communication research if some variables are challenging and costly to measure. In a frequentist framework, failing to reject the null hypothesis could wrongly inform researchers to collect more data in subsequent studies to discern between insufficient power[1] or the null is true. Instead, ROPE could be used to decide whether, in future studies, more data should be collected to retest the same hypothesis. When ROPE provides a strong signal that the null hypothesis is true, it is no longer worthy to collect more of the costly data just to boost the statistical power.

The package bayestestR (Makowski, Ben-Shachar, & Lüdecke, 2019) can be used to calculate and visualize the ROPE.

```
library(bayestestR)
rope(import_brms)
plot(rope(import_brms))
```

The proportion of 95% HDI falling inside the ROPE is 0%. Figure 2 also displays that the 95% HDI (the pink region) has no overlap with the ROPE (the shaded region). Using the decision rules by Kruschke (2018), there are 3 possible scenarios: reject the null (when most of the 95% HDI falls outside the ROPE), accept the null (when most of the 95% HDI falls inside the ROPE) and undecided (when it can't be decided). In the current

[1] Statistical power is a frequentist concept specifying the long-run probability of correctly rejecting the null hypothesis. It is usually not relevant for Bayesian inference.

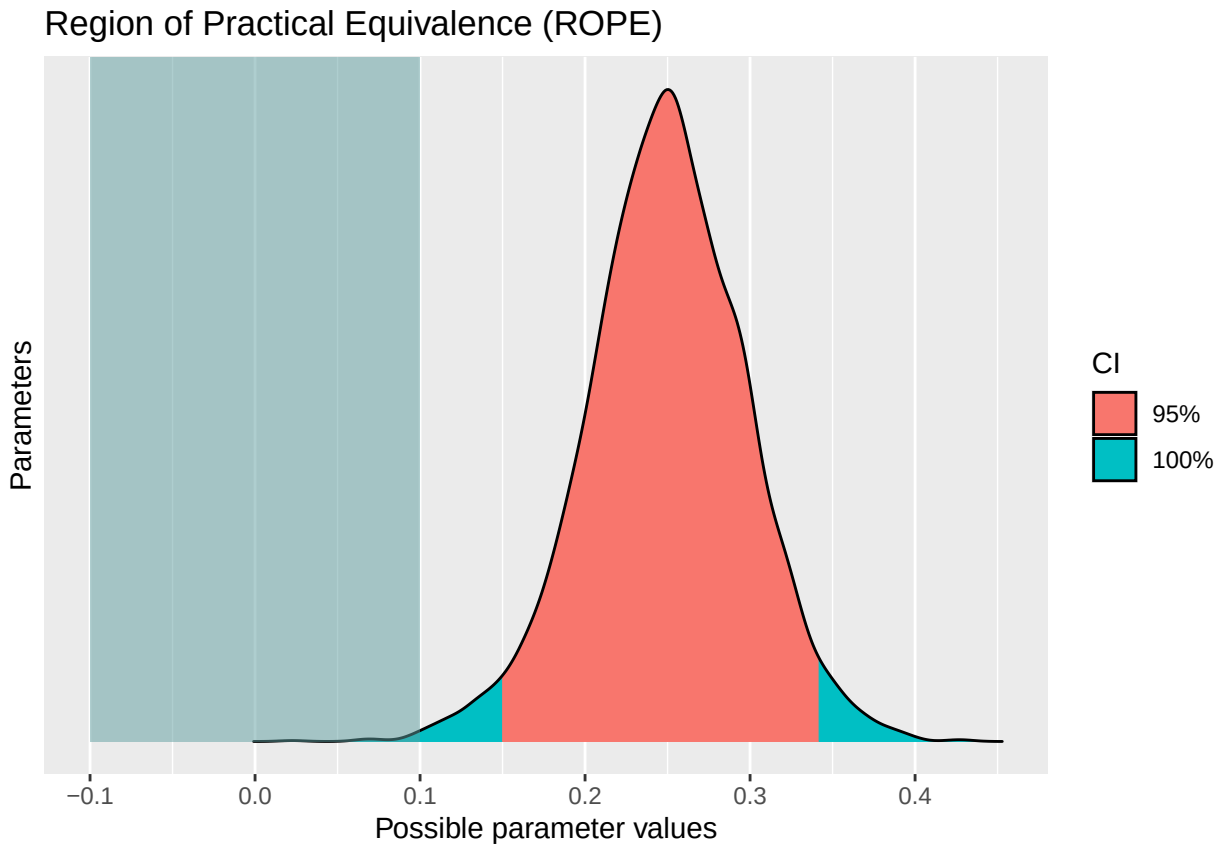

Figure 2. The Region of Practical Equivalence and the 95% High Density Interval

situation, it is very clear that the 95% HDI falls outside of the ROPE as there is 0% overlap. Similar to the result from *hypothesis()*, there is a practical effect and it is correct to reject the null hypothesis.

However, in another situation when we have a simulated independent variable *noise* that is basically random noise.

```
outlet_data$noise <- rnorm(nrow(outlet_data))


noise_brms <- brm(z~offset(log(n))+(1|k)+noise,
                  data = outlet_data, family = negbinomial(),
                  control = list(adapt_delta = 0.99),
                  prior = weaklyinformative_prior,
```

```
                  sample_prior = TRUE)
rope(noise_brms)
```

In this situation, the proportion of 95% HDI falls inside the ROPE is 68%. The proportion is quite high, but it is not decisive. Kruschke (2018) suggests accepting the null when more than 95% of the 95% HDI falls inside the ROPE. Accordingly, the current situation of 68% is "undecided". It might be worthwhile to retest the hypothesis with a larger amount of data.

One should note that the decision rules by Kruschke (2018) depend on many methodological decisions, e.g. how to determine the ROPE, how wide the ROPE should be, and which level of HDI to use. ROPE is also sensitive to the scale of measurement. Nonetheless, ROPE can provide much richer information to inform how future studies should be designed.

Extension #2: Cross-level interaction

We can also study the interaction between a macro- and a micro-level variable. Cross-level interaction is useful to study how the effect of context variables manifest differently in individuals with different characteristics. An example of cross-level interaction in our example is whether the relationship between import volume and increase in China coverage manifests differently between public broadcasters and for-profit media.

We can model a cross-level interaction by adding an interaction term between the macro- and micro-level variables, as well as a varying-slope component for the micro-level variable (Heisig & Schaeffer, 2019). With *brms*, it can be done with this:

```
# suppose "public" is a binary variable
import_brms_public <- brm(
      z~offset(log(n))+(1 + public|k)+log(x_import)*public,
```

```
      data = outlet_data, family = negbinomial(),
      prior = weaklyinformative_prior, sample_prior = TRUE)
import_brms_public
```

The regression coefficients for the main effect of public broadcasting and the interaction terms have a wide 95% HDI that covers zero. It can be visualized with a conditional effects plot (Figure 3, using the *conditional_effects()* function). The trajectories of public broadcasters and for-profit media outlets are similar, indicating no interaction.

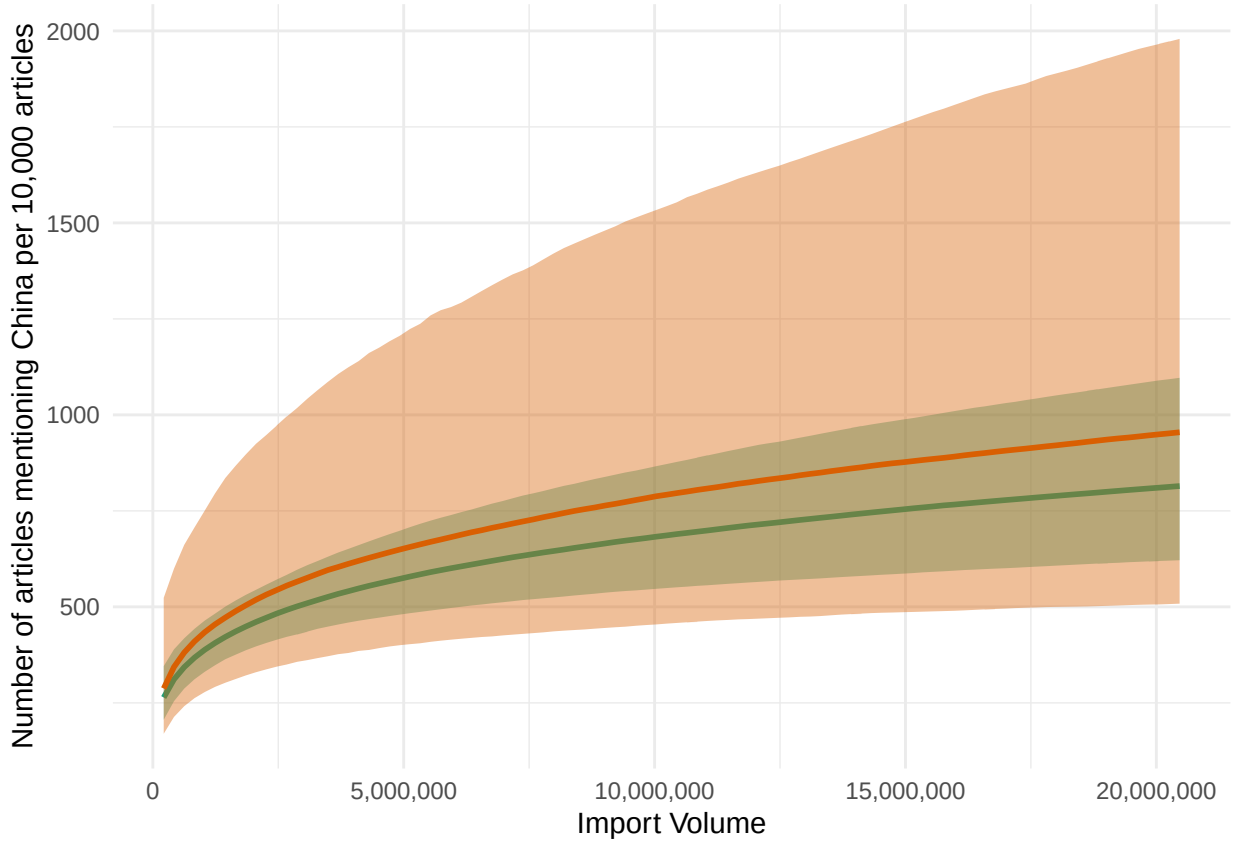

Figure 3. Conditional effects plot showing no interaction among public broadcasting, import volume and China coverage.

Extension #3: Current study as the informative prior for the next study

Suppose we would like to replicate the study and confirm the same relationship with the 2020 useNews data. This time, we do not need to (and actually, we should not) use the weakly informative prior again. It is because we have *prior* understanding about the outlet-level determinants of China coverage: We know from the 2019 data that γ_{01} should be around 0.25 with a standard deviation of 0.05. This can be entered into our new analysis as the informative prior.

```
informative_prior <- c(prior_string("normal(0.25, 0.05)", class = "b",
                                    coef = "logx_import"),
                       prior_string("normal(-6.69, 0.66)",
                                    class = "Intercept"))
import_2020 <- brm(z_2020~offset(log(n_2020))+(1|k)+log(x_import),
                   data = outlet_data2020, family = negbinomial(),
                   prior = informative_prior, sample_prior = TRUE)
```

The new analysis showed that γ_{01} is 0.22 with a 95% HDI of (0.15 to 0.29) for the year 2020. This analysis displays two things: 1) selection of prior is not always, as many authors have suggested, as controversial and subjective; and 2) Bayesian analysis also makes a strong case for replication studies, which open science hopes to enable (Dienlin et al., 2020). The use of informative prior in replication studies enables us to update our prior beliefs with the new evidence from the replications. Priors, replicable or not, will get updated.

Extension #4: Temporal changes

Bayesian multilevel regression analysis is also useful to study temporal changes, because observations from the same subject collected from different time points are

naturally a cluster. The same strategy for handling the clustering of observations in the multilevel modeling can be used. We model that observations from the same media outlet are nested within the media outlet and then the media outlet is nested within its countries. Suppose we would like to study if there is a systematic increase in z across all outlets, maybe due to the COVID-19 pandemic or the presidential election in the US. It can be analyzed with this three-level model.

```
## Suppose yr is a binary dummy variable indicating the year 2020.
import_long <- brm(z~offset(log(n))+(1|k/i)+yr, data = outlet_long,
                   family = negbinomial(),
                   prior = weaklyinformative_prior,
                   sample_prior = TRUE)
import_long
```

The regression coefficient for the year dummy variable is 0.49 (95% HDI 0.38, 0.60). There is a global increase in China coverage from 2019 to 2020.

Leave-one-out cross-validation of Example 2

Another way to compare the cubic model and the parsimonious model is to conduct a leave-one-out cross-validation (LOO). The basic idea of LOO is to study the out-of-sample accuracy of the model by removing one data point at a time. Given the rest of the data and the prediction model, one evaluates how accurate the model is in predicting the outcome of the removed data point. The Bayesian LOO estimate of out-of-sample predictive fit ($elpd_{loo}$) can be used to compare the out-of-sample accuracy of models (Vehtari, Gelman, & Gabry, 2016).

In R, the LOO cross-validation of the two models (parsimonious model: *ppa_mod*; cubic model *ppa_qmod*) can be launched with the function `loo`.

```
loo(ppa_qmod, ppa_mod)
```

The difference in $elpd_{loo}$ between the cubic model and the parsimonious model is only -0.3 with a larger standard error of 0.5. One would say there is a difference in out of sample model fit, when $elpa_{loo}$ is many times larger than its standard error.

Poisson regression

The useNews example could also be modeled by Poisson regression.

```
weaklyinformative_prior <- c(prior_string("normal(0, 1)",
                                          class = "b"),
                             prior_string("normal(0, 1)",
                                          class = "Intercept"))
import_brms_poisson <- brm(z~offset(log(n))+(1|k)+log(x_import),
                           data = outlet_data, family = poisson(),
                           prior = weaklyinformative_prior,
                           sample_prior = TRUE)
```

However, the PPCs reveal a property of the Poisson model. Compared with the PPCs of the negative binomial model (the same as Figure 4 in the main text), the Poisson model tends to produce stimulated data with an unusual high level of certainty (both models were based on 61 data points). It is an indication of shrinkage (reduction of standard error, i.e. more false positives) due to overdispersion (the variance of the data is higher than the expectation of the model, see Gardner, Mulvey, & Shaw, 1995). The Poisson model assumes the data to have only one parameter (λ), which represents both mean and variance. The current data clearly violates this assumption (the variance is far larger than the mean).

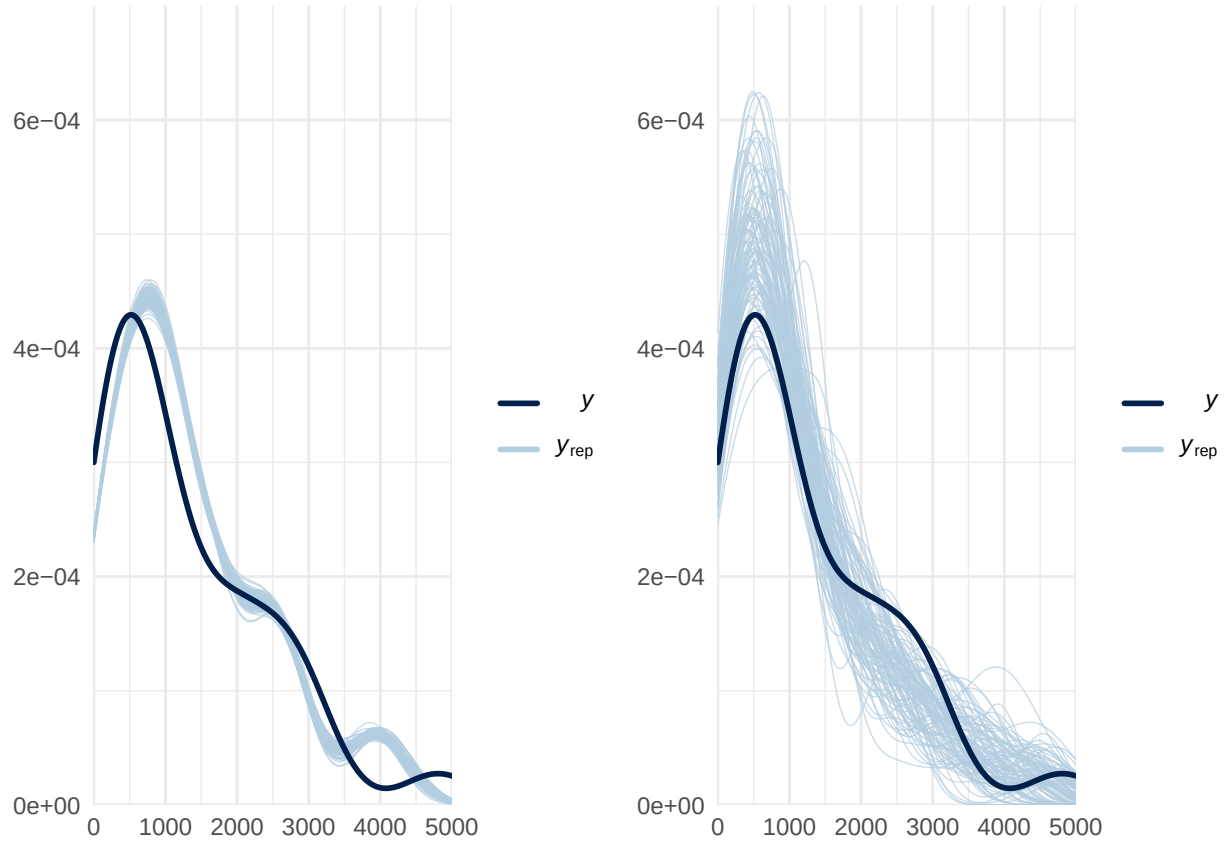

Figure 4. Posterior predictive checks of the Poisson model (Left) and the negative binomial model (Right) with 100 sets of simulated data

References

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., … al. (2020). An agenda for open science in communication. *Journal of Communication*, *71*(1), 1–26. https://doi.org/10.1093/joc/jqz052

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychological Bulletin*, *118*(3), 392–404.

https://doi.org/10.1037/0033-2909.118.3.392

Hamada, B. I. (2021). Determinants of journalists' autonomy and safety: Evidence from the Worlds of Journalism Study. *Journalism Practice*, 1–21. https://doi.org/10.1080/17512786.2021.1871861

Heisig, J. P., & Schaeffer, M. (2019). Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. *European Sociological Review*, *35*(2), 258–279. https://doi.org/10.1093/esr/jcy053

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Puschmann, C., & Haim, M. (2020). *UseNews*. Open Science Framework. https://doi.org/10.17605/OSF.IO/UZCA3

Reich, Z., & Hanitzsch, T. (2013). Determinants of journalists' professional autonomy: Individual and national level factors matter more than organizational ones. *Mass Communication and Society*, *16*(1), 133–156. https://doi.org/10.1080/15205436.2012.669002

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/pbr.16.2.225

Shikano, S. (2019). Hypothesis testing in the Bayesian framework. *Swiss Political Science Review*, *25*(3), 288–299. https://doi.org/10.1111/spsr.12375

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4