Bayesian Workflow Illustrated Using BRMS

2023-06-21

BRMS: Bayesian Regression and Multilevelmodeling in Stan

BRMS package function `brm` fits multilevel models

- model specified using extended R formula syntax
- rich set of distributions
- workflow checks built-in

```
# model with treatment effect
fit1 <- brm(count ~ zAge + zBase + Trt,
            data = epilepsy, family = negbinomial(),
            prior = prior(normal(0, 1), class = b))
```

https://paul-buerkner.github.io/brms/

Bayesian Workflow (elements of)

- Exploratory data analysis / Data elicitation
- Model specification
- Model checking (simulated data)
  - Recover parameters
  - Prior predictive distribution checking
  - Calibration checking
- Model checking (simulated or observed data)
  - Sample diagnostics
  - Posterior predictive checks
- Model comparison
  - Leave-one out cross-validation

Workflow in Context

- Research priorities: innovation, pedagogy
  - Workflow is a blueprint for papers and lesson plans
- Applied concerns: accuracy, efficiency, generalizability
  - Workflow provides a principled choice between available models
- Challenge: automating the process
  - Writing (many) Stan models
  - Simulating data
  - Checking inferences
  - Generating application-specific visualizations
- BRMS meets (most of) these challenges

BRMS processing

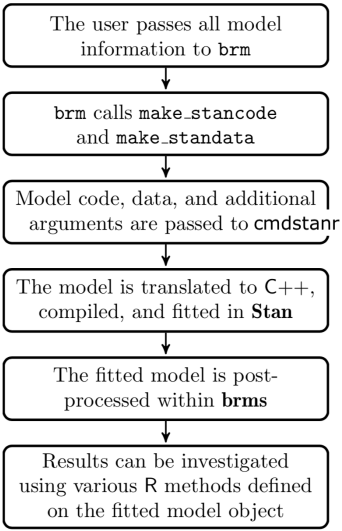
Arguments to the `brm` function provide the full specification of the model.

```
fit1 <- brm(count ~ zAge + zBase + Trt,
            data = epilepsy, family = negbinomial(),
            prior = prior(normal(0, 1), class = b))
```

From these arguments, BRMS

- generates the Stan code and input data
- uses CmdStanR to generate a sample
- returns/manages the fitted model object
- fitted model object has methods for plotting, and prediction, model checking, model comparison,

BRMS Processing



Papers Demonstrating Workflow

Visualization in Bayesian Workflow

Jonah Gabry ✉, Daniel Simpson, Aki Vehtari, Michael Betancourt, Andrew Gelman

Journal of the Royal Statistical Society Series A: Statistics in Society, Volume 182, Issue 2, February 2019, Pages 389–402, https://doi.org/10.1111/rssa.12378

The paper that launched the workflow ships

Bayesian Analysis of Tests with Unknown Specificity and Sensitivity

Andrew Gelman ✉, Bob Carpenter

Journal of the Royal Statistical Society Series C: Applied Statistics, Volume 69, Issue 5, November 2020, Pages 1269–1283, https://doi.org/10.1111/rssc.12435

Workflow in action

Gabry et al - Visualization in the Bayesian Workflow

Running example

- Air pollution from particles less than 2.5 microns in diameter ($PM_{2.5}$).
- Dataset of measurements from ground monitors plus satellites, compiled by Shaddick et al, 2017:

Data Integration Model for Air Quality: A Hierarchical Approach to the Global Estimation of Exposures to Ambient Air Pollution

Gavin Shaddick, Matthew L. Thomas, Amelia Jobling, Michael Brauer, Aaron van Donkelaar, Rick Burnett, Howard Chang, Aaron Cohen, Rita Van Dingenen, Carlos Dora, Sophie Gumy, Yang Liu, Randall Martin, Lance A. Waller, Jason West, James V. Zidek, Annette Prüss-Ustün

https://arxiv.org/pdf/1609.00141.pdf

Exploratory Data Analysis

The Model Can Often Only Be Understood in the Context of the Data

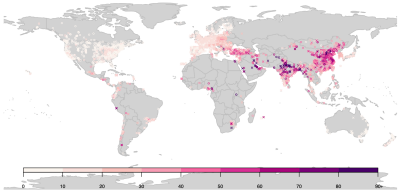
- The model describes the data generating process
- There are many ways to parameterize a multi-level model
- Large data / small data regimes require centered / non-centered parameterization
- Data tendencies require different priors

The Prior Can Often Only Be Understood in the Context of the Likelihood

Gelman et al, 2017

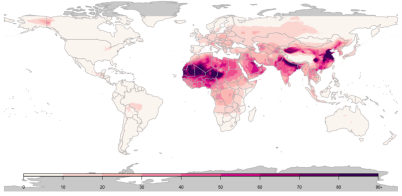
Shaddick et al - Ground-level air pollution

- Data collected by the World Health Organization (WHO)
- 6003 observations, 2980 distinct locations, 107 countries, 7 WHO regions
- sparse data for Africa, central Asia, Russia, no data for 70 countries
- Goal: use satellite data to fill gaps in coverage



Locations of ground monitors measuring $PM_{2.5}$ (circles) and PM_{10} (crosses)

$PM_{2.5}$ ground monitor readings

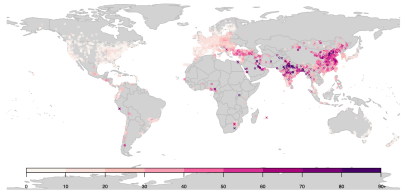


Satellite based estimates of $PM_{2.5}$ (μgm^{-3}) for 2014, by grid cell

$PM_{2.5}$ satellite readings

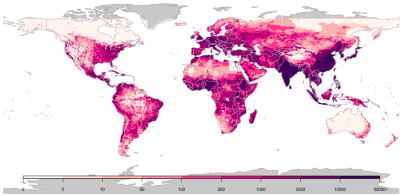
Monitor Locations and Population Density

- Monitor sites are near where people live; but not all people live near monitors.



Locations of ground monitors measuring $PM_{2.5}$ (circles) and PM_{10} (crosses)

Ground monitor locations

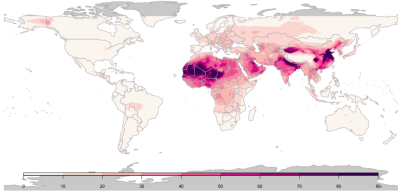


Population estimates for 2014 from the Gridded Population of the World

Population density

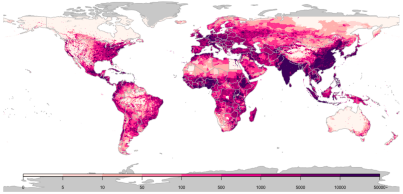
Satellite Measurements and Population Density

- Satellite data measures *optical aerosol depth* (OAD), from which $PM_{2.5}$ is estimated, confounding factors introduce artifacts.
- WHO cares about populated areas.



Satellite based estimates of $PM_{2.5}$ (μgm^{-3}) for 2014, by grid cell

$PM_{2.5}$ satellite readings

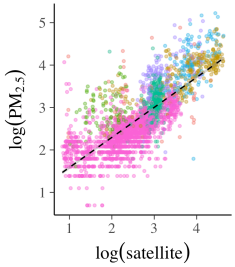


Population estimates for 2014 from the Gridded Population of the World

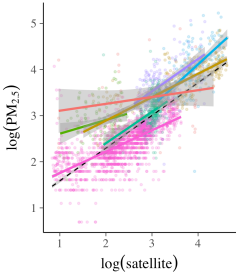
Population density

Gabry et al - Exploratory Data Analysis for Model Specification

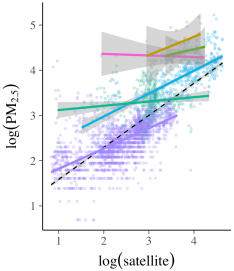
- "Three simple, but plausible, models for the data"
  - Simple linear model
  - Multilevel model, observations stratified by WHO super-region
  - Multilevel model, observations stratified by clustered-region



Simple linear model



7 groups, by WHO region



6 groups, by $PM_{2.5}$

Exploratory Data Analysis Notebooks

Notebooks:

- Python Jupyter:
  https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/eda_air
- R Jupyter:
  https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/eda_air
- R markdown (run via Rstudio):
  https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/eda_air

Multi-level Modeling with BRMS

Structure of a multi-level model: $y_i \sim \mathrm{D}(f^{-1}(\eta_i), \theta)$

- D is a distributional **family**
- θ denotes family-specific parameter(s)
- η is the linear predictor
- η can be transformed by an the inverse *link function* f
- η can be rewritten $\mathrm{X}\beta + \mathrm{Z}\,\mathrm{u}$ where
  β and u are population-level and group-level coefficients respectively,
  X, Z are the corresponding design matrices
- y is the modeled data, i.e., the observed outcome generated by the model given inputs X, Z.

Model Specification in BRMS

```
fit1 <- brm(count ~ zAge + zBase + Trt,
            data = epilepsy, family = negbinomial(),
            prior = prior(normal(0, 1), class = b))
```

- Arguments to the `brm` function provide the full specification of the model.
  - first argument: the extended-syntax R formula for η, the linear predictor
  - second argument (`data`) - a dataframe containing all data for y, X, and Z.
  - arg `family` - distributional family D,
    subarguments specify the link function and θ, default is `guassian`.
  - arg `prior` - priors on η and θ (specifics vary by distribution)

BRMS formula syntax

The BRMS formula syntax for a multilevel model has the general form

response $\sim$ pterms $+$ (gterms $|$ group)

- The pterms contain population-level effects, assumed to be the same across observations.
- The gterms contain so called group-level effects, assumed to vary across the grouping variables specified in group.
- The intercept term is either 1 or 0 for no intercept; if unspecified, default is 1.

This is a highly expressive syntax, see:
https://paul-buerkner.github.io/brms/reference/brmsformula.html

BRMS Model 1: Simple linear Model, Air Pollution Dataset

The baseline model in Gabry et al is the complete pooling model

$$\log\_\mathrm{pm25}_i \sim \mathrm{N}(\alpha + \beta \log\_\mathrm{sat}_i, \sigma)$$

The corresponding BRMS code is:

```
fit_complete_pool = brm(log_pm25 ~ 1 + log_sat, data=sites)
```

Model Correctness Checks

Given a single model and data, either collected or simulated, the most basic correctness checks are:

- Can the sampler fit the data satisfactorily?
  - Grossly misspecified models fail to initialize, or chains fail to converge.
  - Difficult to fit models may require reparameterization
- Are the parameter estimates reasonable?

Basic Model Correctness Checks in BRMS

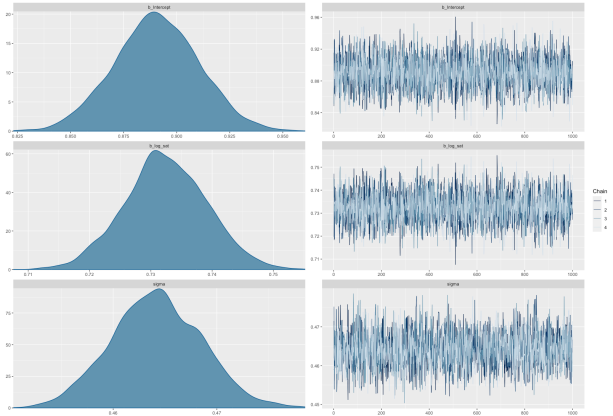
- The print method defined on fitted model object provides a summary:

```
> fit_complete_pool
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: log_pm25 ~ 1 + log_sat
   Data: sites (Number of observations: 6003)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1; total post-warmup draw
Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     0.89      0.02     0.85     0.93 1.00     4074     2825
log_sat       0.73      0.01     0.72     0.75 1.00     4044     3050
Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.46      0.00     0.46     0.47 1.00     3529     2487
```

Basic Model Correctness Checks in BRMS

- The `plot` method shows chain mixing, posterior densities.

```
> plot(fit_complete_pool)
```



Notebook

Jupyter notebook - section 1

- https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/fit_air_
- https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/fit_air_

Multi-level Models

- Full (and approximate) Bayesian methods *quantify uncertainty*
- Multilevel models have per-level variance parameters
- Prior choice is *very* important

Gabry et al - Prior predictive checks

Visualize (or summarize) the prior marginal distribution of the data

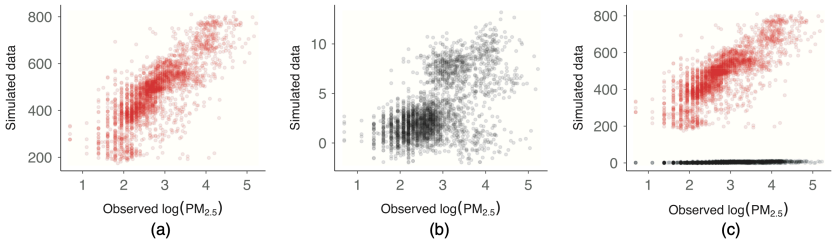


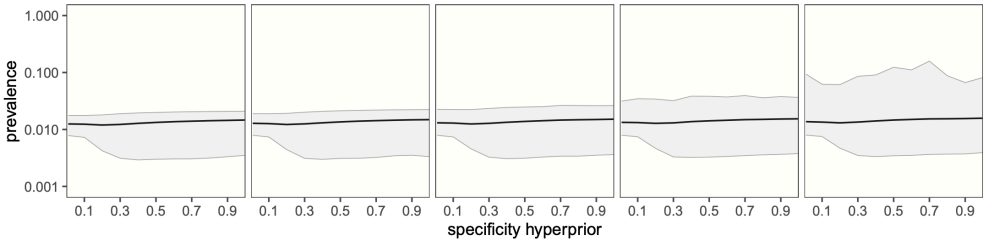
Fig. 4. Visualizing the prior predictive distribution: (a) and (b) show realizations from the prior predictive distribution using priors for the βs and τs that are vague and weakly informative respectively; the same $N_+(0,1)$ prior is used for σ in both cases; simulated data are plotted on the y-axis and observed data on the x-axis; because the simulations under the vague and weakly informative priors are so different, the y-axis scales used in panels (a) and (b) also differ dramatically; (c) emphasizes the difference in the simulations by showing the red points from (a) and the black points from (b) plotted with the same y-axis

Note: histogram of just the simulated data also informative.

Gelman and Carpenter - sensitivity analysis of hyperpriors

Goal: estimate prevalence of Covid-19 in the general population using test of unknown specificity (γ) and sensitivity (δ)

Sensitivity analysis: The scale of variation σ_γ, σ_δ is drawn from a half-Normal with hyperparameters location 0 and scale τ_γ, τ_δ, respectively.



Gelman and Carpenter - sensitivity analysis of hyperpriors

Fig. 2 shows how these hyperprior parameters τ_γ and τ_δ affect inferences for the prevalence π. The posterior median of π is not sensitive to the scales of the hyperpriors, but the uncertainty in that estimate [is sensitive]. ...

The steps that were taken in Sections 2 and 3 show the basic workflow: we start with a simple model; then add hierarchical structure. For the hierarchical model we started with weak priors on the hyperparameters and examined the inferences, which made us realize that we had prior information (that specificities and sensitivities of the tests should not be so variable), which we then incorporated in the next iteration of the model.

BRMS Models 2, 3: Multi-level Model, Air Pollution Dataset

The multilevel model introduces a country-level grouping factor.

$$\log\_\mathrm{pm25}_{ij} \sim \mathrm{N}(\alpha + \alpha_j + (\beta + \beta_j) \log\_\mathrm{sat}_{ij}, \sigma)$$

For this dataset, available grouping factors are either the WHO super-region or the hierarchical clustering based on per-country average $PM_{2.5}$.

```
fit_super_region =
    brm(log_pm25 ~ 1 + log_sat + (1 + log_sat | super_region),
    data=sites)
fit_cluster_region =
    brm(log_pm25 ~ 1 + log_sat + (1 + log_sat | cluster_region),
    data=sites)
```

Notebook

Jupyter notebook - section 2

- https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/fit_air_
- https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/fit_air_

Checking Predictions

- Simple idea:

  "if a model is a good fit we should be able to use it to generate data that resemble the data that we observed"

- Can also simulate data from just the priors - omit likelihood
  - BRMS options `sample_prior="only"` (`sample_prior=TRUE`)
- BRMS/bayesplot package function `pp_check` plots the simulated data.
  - argument `prefix="ppc"` (default) compares observed data to simulated data
  - argument `prefix="ppd"` plots the just simulated density, (use for prior predictive checks).

Gabry et al - Posterior predictive checks

Plot observed data against simulated data

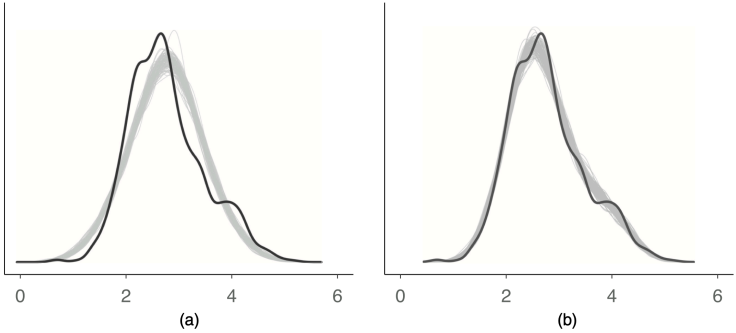


Fig. 6. Kernel density estimate of the observed data set y (dark curves), with density estimates for 100 simulated data sets y_{rep} drawn from the posterior predictive distribution (thin, lighter curves) (these plots can be produced using ppc_dens_overlay in the bayesplot package): (a) model 1; (b) model 2

Notebook

Jupyter notebook - section 3

- https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/fit_air_
- https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/fit_air_

Model Comparison

- Pointwise plots for predictive model comparison - find outliers and points with high leverage

Leave One Out Cross-validation: LOO

- Calibration of predictions - LOO cross-validation of predictive cdf
- Challenge of generaliziability

Notebook

Jupyter notebook - section 4

- https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/fit_air_
- https://github.com/mitzimorris/StanCon2023_brms_tutorial/blob/main/fit_air_

References

- BRMS documentation: https://paul-buerkner.github.io/brms/
- An Introduction to Bayesian Data Analysis for Cognitive Science
- Cross-Validation FAQ
- LOO Vignette