Parameterizations of families in brms

Paul-Christian Bürkner

The purpose of this vignette is to discuss the parameterizations of the families (i.e., response distributions) used in **brms**. For a more general overview of the package see `vignette("brms")`.

Notation

Throughout this vignette, we use y to refer to the response variable and η to refer to the (non-)linear predictor (see `help(brmsformula)` for details on supported predictor terms). We write y_i and η_i for the response and linear predictor of observation i. Furthermore, we use f for a density function and g for an inverse link function.

Location shift models

The density of the **gaussian** family is given by

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y_i - g(\eta_i)}{\sigma}\right)^2\right)$$

where σ is the residual standard deviation. The density of the **student** family is given by

$$f(y_i) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}\sigma} \left(1 + \frac{1}{\nu}\left(\frac{y_i - g(\eta_i)}{\sigma}\right)^2\right)^{-(\nu+1)/2}$$

Γ denotes the gamma function and ν are the degrees of freedom. As $\nu \to \infty$, the student distribution becomes the gaussian distribution. For location shift models, y_i can be any real value.

Binary and count data models

The density of the **binomial** family is given by

$$f(y_i) = \binom{N_i}{y_i} g(\eta_i)^{y_i} (1 - g(\eta_i))^{N_i - y_i}$$

where N_i is the number of trials and $y_i \in \{0, ..., N_i\}$. When all N_i are 1 (i.e., $y_i \in \{0, 1\}$), the bernoulli distribution for binary data arises. **binomial** and **bernoulli** families are distinguished in **brms** as the bernoulli distribution has its own implementation in **Stan** that is computationlly more efficient.

For $y_i \in \mathbb{N}_0$, the density of the **poisson** family is given by

$$f(y_i) = \frac{g(\eta_i)^{y_i}}{y_i!} exp(-g(\eta_i))$$

The density of the **negative binomial** family is

$$f(y_i) = \binom{y_i + \phi - 1}{y_i} \left(\frac{g(\eta_i)}{g(\eta_i) + \phi}\right)^{y_i} \left(\frac{\phi}{g(\eta_i) + \phi}\right)^{\phi}$$

where ϕ is a positive precision parameter. For $\phi \to \infty$, the negative binomial distribution becomes the poisson distribution. The density of the **geometric** family arises if ϕ is set to 1.

Survival models

With survival models we mean all models that are defined on the positive reals only, that is $y_i \in \mathbb{R}^+$. The density of the **lognormal** family is given by

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\left(\frac{\log(y_i) - g(\eta_i)}{\sigma}\right)^2\right)$$

where σ is the residual standard deviation on the log-scale. The density of the **Gamma** family is given by

$$f(y_i) = \frac{(\alpha/g(\eta_i))^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} \exp\left(-\frac{\alpha y_i}{g(\eta_i)}\right)$$

where α is a positive shape parameter. The density of the **weibull** family is given by

$$f(y_i) = \frac{\alpha}{g(\eta_i/\alpha)} \left(\frac{y_i}{g(\eta_i/\alpha)}\right)^{\alpha-1} \exp\left(-\left(\frac{y_i}{g(\eta_i/\alpha)}\right)^\alpha\right)$$

where α is again a positive shape parameter. The **exponential** family arises if α is set to 1 for either the gamma or weibull distribution.

Beta models

The density of the **beta** family for $y_i \in (0, 1)$ is given by

$$f(y_i) = \frac{y_i^{g(\eta_i)\phi-1}(1-y_i)^{(1-g(\eta_i))\phi)-1}}{B(g(\eta_i)\phi, (1-g(\eta_i))\phi)}$$

where B is the beta function and ϕ is a positive precision parameter.

Circular models

The density of the **von_mises** family for $y_i \in (-\pi, \pi)$ is given by

$$f(y_i) = \frac{\exp(\kappa \cos(y_i - g(\eta_i)))}{2\pi I_0(\kappa)}$$

where I_0 is the modified Bessel function of order 0 and κ is a positive precision parameter.

Ordinal and categorical models

For ordinal and categorical models, y_i is one of the categories $1, ..., K$. The intercepts of ordinal models are called thresholds and are denoted as τ_k, with $k \in \{1, ..., K-1\}$, whereas η does not contain a fixed effects intercept. Note that the applied link functions h are technically distribution functions $\mathbb{R} \to [0, 1]$. The density of the **cumulative** family (implementing the most basic ordinal model) is given by

$$f(y_i) = g(\tau_{y_i+1} - \eta_i) - g(\tau_{y_i} - \eta_i)$$

The densities of the **sratio** (stopping ratio) and **cratio** (continuation ratio) families are given by

$$f(y_i) = g(\tau_{y_i+1} - \eta_i) \prod_{k=1}^{y_i} (1 - g(\tau_k - \eta_i))$$

and

$$f(y_i) = (1 - g(\eta_i - \tau_{y_i+1})) \prod_{k=1}^{y_i} g(\eta_i - \tau_k)$$

respectively. Note that both families are equivalent for symmetric link functions such as logit or probit. The density of the **acat** (adjacent category) family is given by

$$f(y_i) = \frac{\prod_{k=1}^{y_i} g(\eta_i - \tau_k) \prod_{k=y_i+1}^{K} (1 - g(\eta_i - \tau_k))}{\sum_{k=0}^{K} \prod_{j=1}^{k} g(\eta_i - \tau_j) \prod_{j=k+1}^{K} (1 - g(\eta_i - \tau_j))}$$

For the logit link, this can be simplified to

$$f(y_i) = \frac{\exp\left(\sum_{k=1}^{y_i} (\eta_i - \tau_k)\right)}{\sum_{k=0}^{K} \exp\left(\sum_{j=1}^{k} (\eta_i - \tau_j)\right)}$$

The linear predictor η can be generalized to also depend on the category k for a subset of predictors. This leads to so called category specific effects (for details on how to specify them see `help(brm)`). Note that **cumulative** and **sratio** models use $\tau - \eta$, whereas **cratio** and **acat** use $\eta - \tau$. This is done to ensure that larger values of η increase the probability of *higher* reponse categories.

The **categorical** family is currently only implemented with the logit link function and has density

$$f(y_i) = \frac{\exp(\eta_{iy_i})}{\sum_{k=1}^{K} \exp(\eta_{ik})}$$

Note that η does also depend on the category k. For reasons of identifiability, η_{i1} is set to 0.

Zero-inflated and hurdle models

Zero-inflated and **hurdle** families extend existing families by adding special processes for responses that are zero. The densitiy of a **zero-inflated** family is given by

$$f_z(y_i) = g_z(\eta_{zi}) + (1 - g_z(\eta_{zi}))f(y_i) \quad \text{if } y_i = 0 f_z(y_i) = (1 - g_z(\eta_{zi})f(y_i) \quad \text{if } y_i > 0$$

The zero-inflation part has its own linear predictor η_{zi} combined with the logit link g_z. Currently implemented families are **zero_inflated_poisson**, **zero_inflated_binomial**, **zero_inflated_negbinomial**, and **zero_inflated_beta**. The density of a **hurdle** family is given by

$$f_z(y_i) = g_z(\eta_{zi}) \quad \text{if } y_i = 0 f_z(y_i) = (1 - g_z(\eta_{zi}))f(y_i)/(1 - f(0)) \quad \text{if } y_i > 0$$

Currently implemented families are **hurdle_poisson**, **hurdle_negbinomial**, and **hurdle_gamma**. Both linear predictors (i.e., η and η_z) can be specified within the `formula` argument. See `help(brmsformula)` for instructions on how to do that.