Paramétrisations des lois Beta et Gamma

C'est presque un travail de journaliste de recouper toutes ces expressions. Bon...

1 Loi Gamma

1.1 Paramétrisation dans R

La loi Gamma est une loi à support dans $\mathbb{R}_+$ dont la fonction de densité f peut s'écrire sous la forme :

$$f : y \mapsto \left(\frac{1}{\Gamma(\alpha)}\right) \frac{1}{\sigma^\alpha} y^{\alpha-1} \exp(-\frac{y}{\sigma}) \mathbb{1}_{\mathbb{R}_+}(y) \tag{1}$$

avec $\alpha > 0$, $\sigma > 0$ et $\Gamma : \alpha \mapsto \int_0^{+\infty} t^{\alpha-1} e^{-t} dt$ la fonction Gamma, telle que $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ et donc $\Gamma(n + 1) = n!$ pour n entier.

Cette paramétrisation est celle qu'on trouve dans la documentation de R, notamment utile pour les fonctions *dgamma*, *rgamma*. Les paramètres α et σ sont appelés respectivement *shape* et *scale*. Sous cette paramétrisation, on a $\mathbb{E}(X) = \alpha\sigma$ et $\mathbb{V}(X) = \alpha\sigma^2$.

1.2 Formulation "classique"

Elle peut aussi s'écrire :

$$f : y \mapsto \left(\frac{1}{\Gamma(\alpha)}\right) \beta^\alpha y^{\alpha-1} \exp(-\beta y) \tag{2}$$

avec β appelé *rate*. On a $\beta = \frac{1}{\sigma}$ et $\mathbb{E}(X) = \frac{\alpha}{\beta}$ et $\mathbb{V}(X) = \frac{\alpha}{\beta^2}$. C'est cette paramétrisation qui est utilisé dans JAGS avec la commande $y \sim$ *dgamma(alpha, beta)* ou dans STAN avec la commande $y \sim$ *gamma(alpha, beta)*.

[1] [2]

1.3 Paramétrisation dans *brms*

$$f : y \mapsto \left(\frac{1}{\Gamma(\alpha)}\right) \left(\frac{\alpha}{\mu}\right)^\alpha y^{\alpha-1} \exp(-\left(\frac{\alpha}{\mu}\right) y) \tag{3}$$

soit $\beta = \frac{\alpha}{\mu}$ où α est toujours appelé *shape parameter*. $\frac{1}{\sigma} = \frac{\alpha}{\mu}$ pour joindre la paramétrisation de R. On a donc $\mathbb{E}(X) = \mu$ et $\mathbb{V}(X) = \frac{\mu}{\alpha}$. [3]

2 Loi Beta

2.1 Paramétrisation "classique"

La loi Beta est une loi à support dans $]0, 1[$ dont la fonction de densité f peut s'écrire sous la forme :

$$f : y \mapsto \left(\frac{1}{B(\alpha, \beta)}\right) y^{\alpha-1}(1 - y)^{\beta-1} \mathbb{1}_{]0,1[}(y) \tag{4}$$

avec $\alpha > 0$, $\beta > 0$ et $B : (\alpha, \beta) \mapsto \int_0^1 t^{\alpha-1}(1 - t)^{\beta-1} dt$ est la fonction Beta, qui peut aussi s'écrire $B : (\alpha, \beta) \mapsto \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ où Γ est la fonction Gamma précédemment introduite.

Cette paramétrisation est celle qu'on trouve dans la documentation de R, notamment utile pour les fonctions *dbeta*, *rbeta*. Les paramètres α et β sont appelés respectivement *shape*1 et *shape*2.

Sous cette paramétrisation, on a $\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$ et $\mathbb{V}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Cette même paramétrisation est utilisé dans JAGS à travers la commande $y \sim$ *dbeta(alpha, beta)* ou dans STAN avec la commande $y \sim$ *beta(alpha, beta)* :

$$f : y \mapsto \left(\frac{1}{B(\alpha, \beta)}\right) y^{\alpha-1}(1-y)^{\beta-1}\mathbb{1}_{]0,1[}(y) \tag{5}$$

[1] [2]

2.2 Paramétrisation dans *brms*

$$f : y \mapsto \left(\frac{1}{B(\mu\phi, \phi(1-\mu))}\right) y^{\mu\phi-1}(1-y)^{\phi(1-\mu)-1}\mathbb{1}_{]0,1[}(y) \tag{6}$$

avec $\mu = \frac{\alpha}{\phi}$ et $\beta = \phi(1-\mu)$ où ϕ appelé *precision parameter*. On a $\mathbb{E}(X) = \mu$ et $\mathbb{V}(X) = \mu(1-\mu)\left(\frac{1}{\phi+1}\right)$.

[3]

3 Lois inflated

Les modèles linéaires généralisés sont bâtis sur la bijection monotone du prédicteur linéaire $x_i^T\beta$ et d'un paramètre θ de la loi $\mathbb{P}_\theta$ par laquelle on modélise le phénomène. La loi de Poisson est à support dans $\mathbb{N}$ mais on peut infléchir la forme de sa densité modifiant le poids de la valeur 0, mais 0 appartient bien au support de la loi de Poisson. De même pour la loi Gamma dont le support est $[0, +\infty[$, elle comprend bien 0. On peut modifier le poids de la valeur 0 avec une loi inflated, mais 0 appartient bien au support de la loi Gamma.

La bijection entre l'espace $]0, +\infty[$ du paramètre λ de la loi de Poisson et l'espace du prédicteur linéaire $]-\infty, +\infty[$ est souvent assurée par le lien logarithmique. De même pour la loi Gamma, la bijection entre l'espace $]0, +\infty[$ du paramètre α de la loi Gamma et $]-\infty, +\infty[$, est souvent assurée par le lien logarithmique. Tout autre bijection monotone entre $]0, +\infty[$ et $]-\infty, +\infty[$ conviendrait.

Pour une loi telle que la loi Beta en revanche, l'appelation $inflated$ semble mauvaise puisque si on a des valeurs en 0, la loi Beta standard n'est même pas définie en 0 puisqu'elle est définie sur $]0, 1[$. On devrait plutôt parler de mon point de vue de loi beta *extended*. La loi beta extended revient à estimer une proportion de zero qui suit une loi de Bernoulli, puis à réaliser un glm sur les données restantes.

3.1 Loi *zero inflated*

La fonction de densité f peut s'écrire sous la forme dite *zero inflated* :

$$f_Z(y) = z + (1-z)f(0) \text{ if } y = 0 \text{ and } f_Z(y) = (1-z)f(y) \text{ if } y > 0.$$

Ou sous la forme dite Hurdle :

$$f_Z(y) = z \text{ if } y = 0 \text{ and } f_Z(y) = (1-z)f(y)/(1-f(0)) \text{ if } y > 0.$$

La forme Hurdle n'est pas implémentée pour la loi Beta dans *brms* au printemps 2024 [3].

3.2 Loi *zero-one inflated*

Pour une loi à support dans $]0,1[$ telle que la loi Beta, la prise en compte des deux bornes peut se faire avec la famille *zero_one_inflated_beta* qui donne la probabilité :

$$f_{\alpha,\gamma}(y) = \alpha(1-\gamma) \text{ if } y = 0 \ ; \ f_{\alpha,\gamma}(y) = \alpha\gamma \text{ if } y = 1 \ ; \ f_{\alpha,\gamma}(y) = (1-\alpha) f_{\alpha,\gamma}(y) \text{ si } y \in]0,1[.$$

4 Syntaxe *brms*

```r
brmsfamily(family = "Gamma",link="log")

brmsfamily(family = "Beta", link = "logit", link_phi = "log")

brmsfamily(family = "zero_inflated_beta",
           link = "logit",
           link_phi = "log",
           link_zi = "logit" # zi probability of 0
           )

brmsfamily(family = "zero_one_inflated_beta",
           link = "logit",
           link_phi = "log",
           link_zoi = "logit", # alpha of the formula
           link_coi = "logit" # conditional one-inflation probability gamma
           )
```

5 Cadre GLM

Soit Y une variable aléatoire à valeurs dans $\mathbb{R}$, de loi $\mathbb{P}_Y$. On dit que la loi de Y appartient à la famille exponentielle si elle peut s'écrire sous la forme :

$$\mathbb{P}_\theta(dy) = \exp\left(\left(\frac{y\theta - b(\theta)}{\phi}\right) + c(y,\phi)\right)\nu(dy)$$

où ν est une mesure de référence, b et c des fonctionnelles, $\phi > 0$ un paramètre de nuisance et $\theta \in \mathbb{R}$, dit paramètre naturel de la loi, appartient à $D_{\nu,\phi} = \{\theta, \int \exp\left(\left(\frac{y\theta - b(\theta)}{\phi}\right) + c(y,\phi)\right)\nu(dy) < \infty\}$.
Si $\theta \in \mathring{D}_{\nu,\phi}$, alors $b'(\theta) = \mathbb{E}_\theta(y)$.

On considère un phénomène dont on a n réalisations $(y_i)_{i\in[\![1,n]\!]} \in \mathbb{R}^n$ qu'on suppose issues d'une même famille de lois mais de paramètres $(\theta_i)_{i\in[\![1,n]\!]} \in \mathbb{R}^n$ potentiellement différents. On suppose ces lois indépendantes. Soient $(x_i)_{i\in[\![1,n]\!]} \in (\mathbb{R}^p)^n$ les covariables supposées expliquer le phénomène.

Un modèle linéaire généralisé (*Generalized Linear Model*) pour données $(x_i, y_i)_{i\in[\![1,n]\!]} \in (\mathbb{R}^p \times \mathbb{R})^n$ est déterminé par la donnée d'une famille exponentielle $((\mathbb{P}_{\theta_i})_{\theta_i\in\mathbb{R}})_{i\in[\![1,n]\!]}$, de paramètres $\beta \in \mathbb{R}^p$ et d'une fonction $\gamma : \mathbb{R} \longrightarrow \Theta$, où Θ est l'espace du paramètre de la loi, telle que :

$$\mathbb{E}(y_i) = \gamma\left(x_i^T\beta\right) \Longleftrightarrow \gamma^{-1}(\mathbb{E}(y_i)) = x_i^T\beta$$

γ^{-1} est appelée la fonction de lien. Elle relie θ_i à $x_i^T\beta$ puisque $\mathbb{E}(y_i) = b'(\theta_i)$.

Si $\gamma = b'()$, on parle de lien canonique.

5.1 Modèle Poisson

Si $(y_i)_{i\in[\![1,n]\!]} \in \mathbb{N}^n$, alors on peut envisager de modéliser le phénomène par une loi de Poisson de paramètre λ. Cette loi s'écrit :

$$\mathbb{P}_\theta(dy) = e^{-\lambda}\frac{\lambda^y}{y!}\nu(dy)$$

où ν est la mesure de comptage sur $\mathbb{N}$. Nous avons :

$$\begin{aligned}\mathbb{P}_\theta(dy) &= \exp(-\lambda + y\log(\lambda))\frac{1}{y!}\nu(dy)\\ &= \exp(y\theta - b(\theta))\mu(dy)\end{aligned}$$

avec

- $\mu(dy) = \frac{1}{y!}\nu(dy)$
- $\theta = \log(\lambda)$ ($\Longleftrightarrow \lambda = e^\theta$)
- $b(\theta) = \lambda = e^\theta$ ($\Longrightarrow b'(\theta) = e^\theta$)

L'espérance de la loi de Poisson est : $\gamma(x_i^T\beta) = \mathbb{E}(y_i) = \lambda_i = e^{\theta_i}$.
Si $\theta_i = x_i^T\beta$, alors on a $\gamma(x_i^T\beta) = b'(\theta_i)$, c'est-à-dire le lien canonique.
Le modèle de régression Poisson consiste alors en la modélisation :

$$\mathbb{E}(y_i) = e^{x_i^T\beta}$$

5.2 Modèle logit

Dans le cas où $(y_i)_{i\in[\![1,n]\!]} \in \{0,1\}^n$, on peut évidemment appliquer un modèle linéaire généralisé. La seule loi modélisant un phénomène à valeurs dans $\{0,1\}$ est la loi de Bernoulli. La loi de Bernoulli peut s'écrire :

$$\mathbb{P}_\theta(dy) = p^y(1-p)^{1-y}\nu(dy)$$

où ν est la mesure de comptage sur $\{0,1\}$. On peut exprimer f sous la forme :

$$\begin{aligned}\mathbb{P}_\theta(dy) &= p^y(1-p)^{1-y}\nu(dy)\\ &= \exp(y\log(p) + (1-y)\log(1-p))\nu(dy)\\ &= \exp(y\log(\frac{p}{1-p}) + \log(1-p))\nu(dy)\\ &= \exp(y\theta - b(\theta))\nu(dy)\end{aligned}$$

avec

- $\theta = \log\left(\frac{p}{1-p}\right)(\Longleftrightarrow p = \frac{e^\theta}{1+e^\theta})$
- $b(\theta) = -\log(1-p) = \log(1+e^\theta)$ ($\Longrightarrow$ $b'(\theta) = \frac{e^{\theta_i}}{1+e^{\theta_i}}$)

L'espérance de la loi de Bernoulli est : $\gamma(x_i^T\beta) = \mathbb{E}(y_i) = p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$.
Si $\theta_i = x_i\beta$, alors on a $\gamma(x_i^T\beta) = b'(\theta_i)$, c'est-à-dire le lien canonique.
Le modèle de régression logit consiste alors en la modélisation :

$$\mathbb{E}(y_i) = \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)}$$

5.3 GLM Gamma

Dans le cadre *glm*, le paramètre de la loi Gamma qu'on ajuste à une combinaison linéaire de paramètres est le paramètre *rate* β (ou *scale* σ). Le paramètre α est supposé identique pour toutes les observations.

5.4 GLM Beta

C'est μ qui est estimé, la valeur de ϕ étant considérée commune à toutes les observations. Nous avons compris qu'il était logique que *brms*, spécialement conçu pour les régressions, utilise cette paramétrisation telle que ce soit le paramètre μ égal à l'espérance de la loi Beta qui soit ajusté, contrairement à $STAN$ qui utilise l'expression classique **??**.

6 Outils en ligne sympas

https://distribution-explorer.github.io/continuous/gamma.html
https://distribution-explorer.github.io/continuous/beta.html https://stackoverflow.com/questions/43615260/running-a-glm-with-a-gamma-distribution-but-data-includes-zeros

References

[1] Martyn Plummer. *JAGS user manual 4.30*. 2017.

[2] Stan Development Team. *Stan Functions Reference*. en. 2024.

[3] Paul Bürkner. *Parameterization of Response Distributions in brms*. en.