Introduction à la modélisation statistique bayésienne

Ladislas Nalborczyk
GIPSA-lab, CNRS, Univ. Grenoble Alpes

Planning

Cours n°01 : Introduction à l'inférence bayésienne

Cours n°02 : Modèle Beta-Binomial

Cours n°03 : Introduction à brms, modèle de régression linéaire

Cours n°04 : Modèle de régression linéaire (suite)

Cours n°05 : Markov Chain Monte Carlo

Cours n°06 : Modèle linéaire généralisé

Cours n°07 : Comparaison de modèles

Cours n°08 : Modèles multi-niveaux

Cours n°09 : Modèles multi-niveaux généralisés

Cours n°10 : Data Hackaton

Introduction

Le modèle linéaire Gaussien qu'on a vu aux Cours n°03 et n°04 est caractérisé par un ensemble de postulats, entre autres choses :

- Les résidus sont distribués selon une loi Normale
- La variance de cette distribution Normale est constante (postulat d'homogénéité de la variance)
- Les prédicteurs agissent sur la moyenne de cette distribution
- La moyenne suit un modèle linéaire ou multi-linéaire

$$
\begin{aligned}
y_i &\sim \mathrm{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_1 \cdot \mathrm{X1}_i + \beta_2 \cdot \mathrm{X2}_i
\end{aligned}
$$

Introduction

Cette modélisation (le choix d'une distribution Normale) induit plusieurs contraintes, par exemple :

- Les données observées sont définies sur un espace continu
- Cet espace n'est pas borné

Comment modéliser certaines données qui ne respectent pas ces contraintes ? Par exemple, la proportion de bonnes réponses à un test (bornée entre 0 et 1), un temps de réponse (restreint à des valeurs positives et souvent distribué de manière approximativement log-normale), un nombre d'accidents...

Introduction

Nous avons déjà rencontré un modèle différent : le modèle Beta-Binomial (cf. Cours n°02).

- Les données observées sont binaires (e.g., 0 vs. 1) ou le résultat d'une somme d'observations binaires (e.g., une proportion)
- La probabilité de succès (obtenir 1) a priori se caractérise par une distribution Beta
- La probabilité de succès (obtenir 1) ne dépend d'aucun prédicteur

$$
\begin{aligned}
w &\sim \mathrm{Binomial}(n, p) \\
p &\sim \mathrm{Beta}(a, b)
\end{aligned}
$$

Introduction

Cette modélisation induit deux contraintes :

- Les données observées sont définies sur un espace discret
- Cet espace est borné

Comment pourrait-on ajouter des prédicteurs à ce modèle ?

Modèle linéaire généralisé

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n, p_i) \\
f(p_i) &= \alpha + \beta \cdot x_i
\end{aligned}
$$

Modèle linéaire généralisé

$$\begin{aligned} y_i &\sim \text{Binomial}(n, p_i) \\ f(p_i) &= \alpha + \beta \cdot x_i \end{aligned}$$

Objectifs :

Modèle linéaire généralisé

$$\begin{aligned} y_i &\sim \text{Binomial}(n, p_i) \\ f(p_i) &= \alpha + \beta \cdot x_i \end{aligned}$$

Objectifs :

- Rendre compte de données discrètes générées par un processus unique

Modèle linéaire généralisé

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n, p_i) \\
f(p_i) &= \alpha + \beta \cdot x_i
\end{aligned}
$$

Objectifs :

- Rendre compte de données discrètes générées par un processus unique
- Introduire des prédicteurs dans le modèle

Modèle linéaire généralisé

$$
\begin{aligned}
y_i &\sim \mathrm{Binomial}(n, p_i) \\
f(p_i) &= \alpha + \beta \cdot x_i
\end{aligned}
$$

Objectifs :

- Rendre compte de données discrètes générées par un processus unique
- Introduire des prédicteurs dans le modèle

Deux changements par rapport au modèle Gaussien :

Modèle linéaire généralisé

$$
\begin{aligned}
y_i &\sim \mathrm{Binomial}(n, p_i) \\
f(p_i) &= \alpha + \beta \cdot x_i
\end{aligned}
$$

Objectifs :

- Rendre compte de données discrètes générées par un processus unique
- Introduire des prédicteurs dans le modèle

Deux changements par rapport au modèle Gaussien :

- L'utilisation d'une distribution de probabilité Binomiale

Modèle linéaire généralisé

$$
\begin{aligned}
y_i &\sim \mathrm{Binomial}(n, p_i) \\
f(p_i) &= \alpha + \beta \cdot x_i
\end{aligned}
$$

Objectifs :

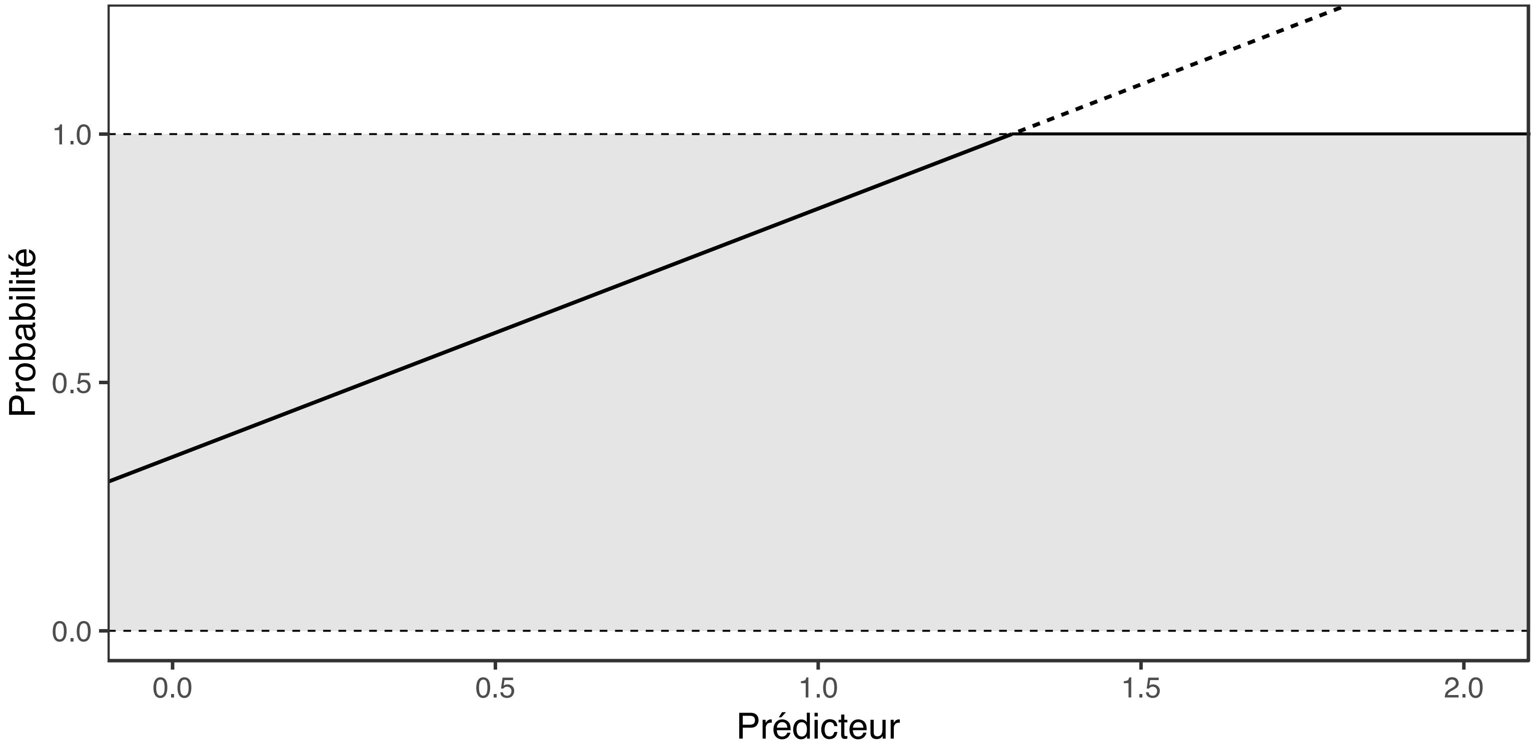
- Rendre compte de données discrètes générées par un processus unique
- Introduire des prédicteurs dans le modèle

Deux changements par rapport au modèle Gaussien :

- L'utilisation d'une distribution de probabilité Binomiale
- Le modèle linéaire ne sert plus à décrire directement un des paramètres de la distribution, mais une fonction de ce paramètre (on peut aussi considérer que le modèle Gaussien était formulé avec une fonction identité)

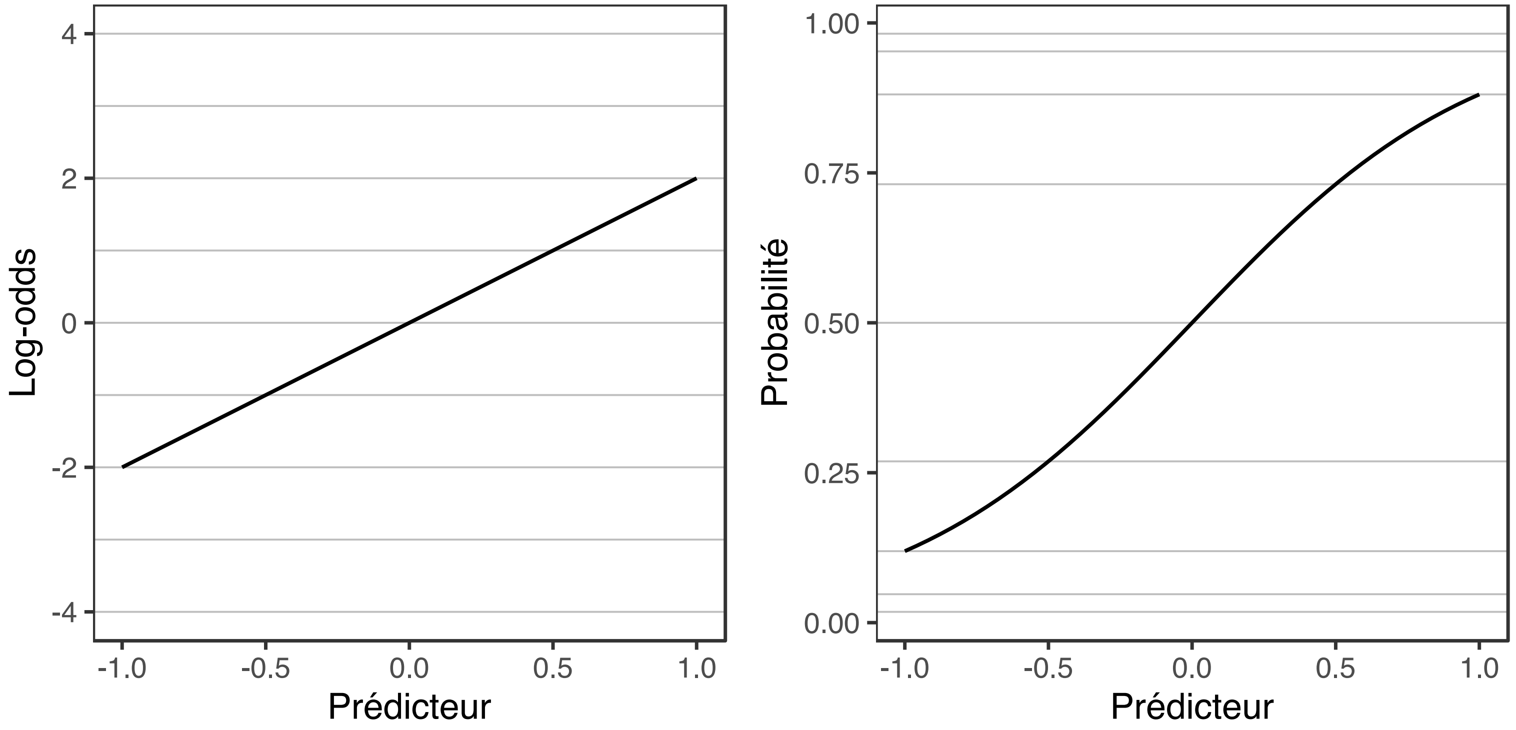
Fonction de lien

Les fonctions de lien ont pour tâche de mettre en correspondance l'espace d'un modèle linéaire (non borné) avec l'espace d'un paramètre potentiellement borné comme une probabilité, définie sur l'intervalle $[0, 1]$.



Fonction de lien

Les fonctions de lien ont pour tâche de mettre en correspondance l'espace d'un modèle linéaire (non borné) avec l'espace d'un paramètre potentiellement borné comme une probabilité, définie sur l'intervalle $[0, 1]$.



Régression logistique

La fonction Logit du GLM binomial (on parle de "log-odds") :

Régression logistique

La fonction Logit du GLM binomial (on parle de “log-odds”) :

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$$

Régression logistique

La fonction Logit du GLM binomial (on parle de “log-odds”) :

$$\mathrm{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$$

La cote d'un événement (*odds* en anglais) est le ratio entre la probabilité que l'événement se produise et la probabilité qu'il ne se produise pas. Le logarithme de cette cote est prédit par un modèle linéaire.

Régression logistique

La fonction Logit du GLM binomial (on parle de “log-odds”) :

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

La cote d'un événement (*odds* en anglais) est le ratio entre la probabilité que l'événement se produise et la probabilité qu'il ne se produise pas. Le logarithme de cette cote est prédit par un modèle linéaire.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta \cdot x_i$$

Régression logistique

La fonction Logit du GLM binomial (on parle de "log-odds") :

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

La cote d'un événement (*odds* en anglais) est le ratio entre la probabilité que l'événement se produise et la probabilité qu'il ne se produise pas. Le logarithme de cette cote est prédit par un modèle linéaire.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta \cdot x_i$$

Pour retrouver la probabilité d'un événement, il faut utiliser la fonction de **lien inverse**, la fonction **logistique** (ou logit-inverse) :

Régression logistique

La fonction Logit du GLM binomial (on parle de “log-odds”) :

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

La cote d'un événement (*odds* en anglais) est le ratio entre la probabilité que l'événement se produise et la probabilité qu'il ne se produise pas. Le logarithme de cette cote est prédit par un modèle linéaire.

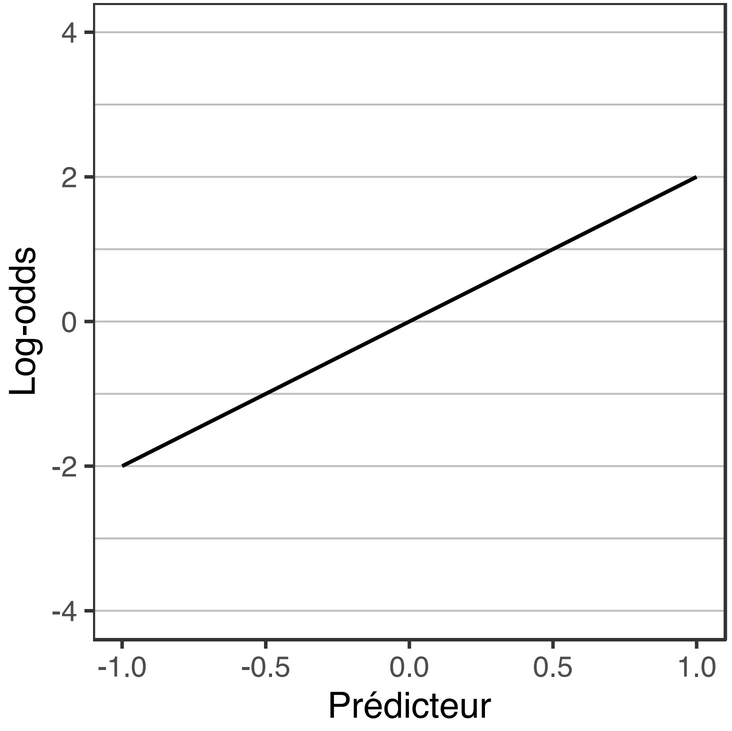
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta \cdot x_i$$

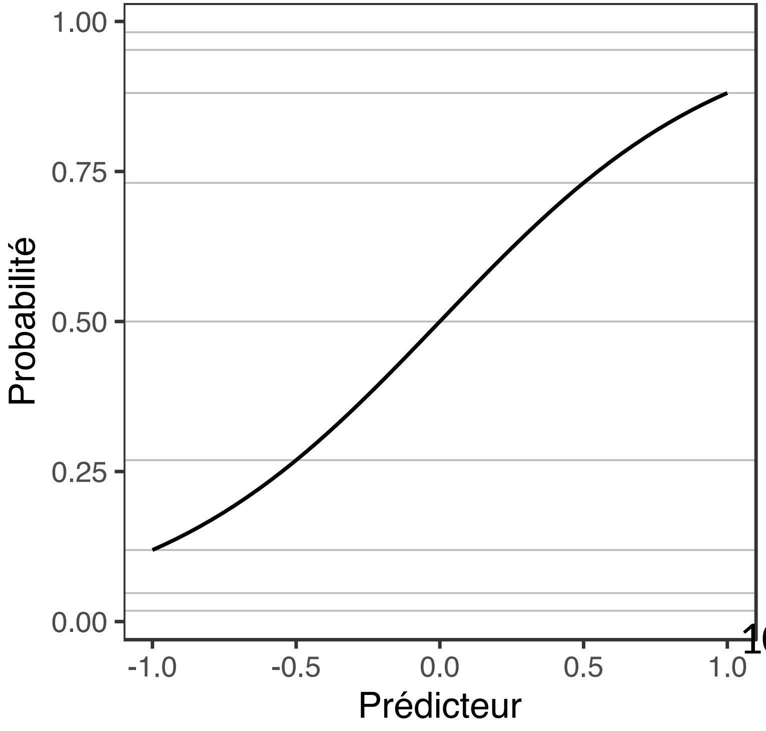
Pour retrouver la probabilité d'un événement, il faut utiliser la fonction de **lien inverse**, la fonction **logistique** (ou logit-inverse) :

$$p_i = \frac{\exp(\alpha + \beta \cdot x_i)}{1 + \exp(\alpha + \beta \cdot x_i)}$$

Complications induites par la fonction de lien

Ces fonctions de lien posent des problèmes d'interprétation : Un changement d'une unité sur un prédicteur n'a plus un effet constant sur la probabilité mais la modifie plus ou moins en fonction de son éloignement à l'origine. Quand $x = 0$, une augmentation d'une demi-unité (i.e., $x = 0.5$) se traduit par une augmentation de la probabilité de 0.25. Puis, chaque augmentation d'une demi-unité se traduit par une augmentation de p de plus en plus petite…





Complications induites par la fonction de lien

Deuxième complication : cette fonction de lien force chaque prédicteur à interagir avec lui même et à interagir avec TOUS les autres prédicteurs, même si les interactions ne sont pas explicites…

Complications induites par la fonction de lien

Deuxième complication : cette fonction de lien force chaque prédicteur à interagir avec lui même et à interagir avec TOUS les autres prédicteurs, même si les interactions ne sont pas explicites…

Dans un modèle Gaussien, le taux de changement de y en fonction de x est donné par $\partial(\alpha + \beta x)$ / $\partial x = \beta$ et ne dépend pas de x (i.e., β est constant).

Complications induites par la fonction de lien

Deuxième complication : cette fonction de lien force chaque prédicteur à interagir avec lui même et à interagir avec TOUS les autres prédicteurs, même si les interactions ne sont pas explicites…

Dans un modèle Gaussien, le taux de changement de y en fonction de x est donné par $\partial(\alpha + \beta x) \ / \ \partial x = \beta$ et ne dépend pas de x (i.e., β est constant).

Dans un GLM binomial (avec une fonction de lien logit), la probabilité d'un événement est donné par la fonction logistique :

Complications induites par la fonction de lien

Deuxième complication : cette fonction de lien force chaque prédicteur à interagir avec lui même et à interagir avec TOUS les autres prédicteurs, même si les interactions ne sont pas explicites…

Dans un modèle Gaussien, le taux de changement de y en fonction de x est donné par $\partial(\alpha + \beta x) \,/\, \partial x = \beta$ et ne dépend pas de x (i.e., β est constant).

Dans un GLM binomial (avec une fonction de lien logit), la probabilité d'un événement est donné par la fonction logistique :

$$p_i = \frac{\exp(\alpha + \beta \cdot x_i)}{1 + \exp(\alpha + \beta \cdot x_i)}$$

Complications induites par la fonction de lien

Deuxième complication : cette fonction de lien force chaque prédicteur à interagir avec lui même et à interagir avec TOUS les autres prédicteurs, même si les interactions ne sont pas explicites…

Dans un modèle Gaussien, le taux de changement de y en fonction de x est donné par $\partial(\alpha + \beta x) \;/\; \partial x = \beta$ et ne dépend pas de x (i.e., β est constant).

Dans un GLM binomial (avec une fonction de lien logit), la probabilité d'un événement est donné par la fonction logistique :

$$p_i = \frac{\exp(\alpha + \beta \cdot x_i)}{1 + \exp(\alpha + \beta \cdot x_i)}$$

Et le taux de changement de p en fonction du prédicteur x est donné par :

Complications induites par la fonction de lien

Deuxième complication : cette fonction de lien force chaque prédicteur à interagir avec lui même et à interagir avec TOUS les autres prédicteurs, même si les interactions ne sont pas explicites…

Dans un modèle Gaussien, le taux de changement de y en fonction de x est donné par $\partial(\alpha + \beta x) / \partial x = \beta$ et ne dépend pas de x (i.e., β est constant).

Dans un GLM binomial (avec une fonction de lien logit), la probabilité d'un événement est donné par la fonction logistique :

$$p_i = \frac{\exp(\alpha + \beta \cdot x_i)}{1 + \exp(\alpha + \beta \cdot x_i)}$$

Et le taux de changement de p en fonction du prédicteur x est donné par :

$$\frac{\partial p}{\partial x} = \frac{\beta}{2(1 + cosh(\alpha + \beta \cdot x)}$$

Complications induites par la fonction de lien

Deuxième complication : cette fonction de lien force chaque prédicteur à interagir avec lui même et à interagir avec TOUS les autres prédicteurs, même si les interactions ne sont pas explicites…

Dans un modèle Gaussien, le taux de changement de y en fonction de x est donné par $\partial(\alpha + \beta x)$ / $\partial x = \beta$ et ne dépend pas de x (i.e., β est constant).

Dans un GLM binomial (avec une fonction de lien logit), la probabilité d'un événement est donné par la fonction logistique :

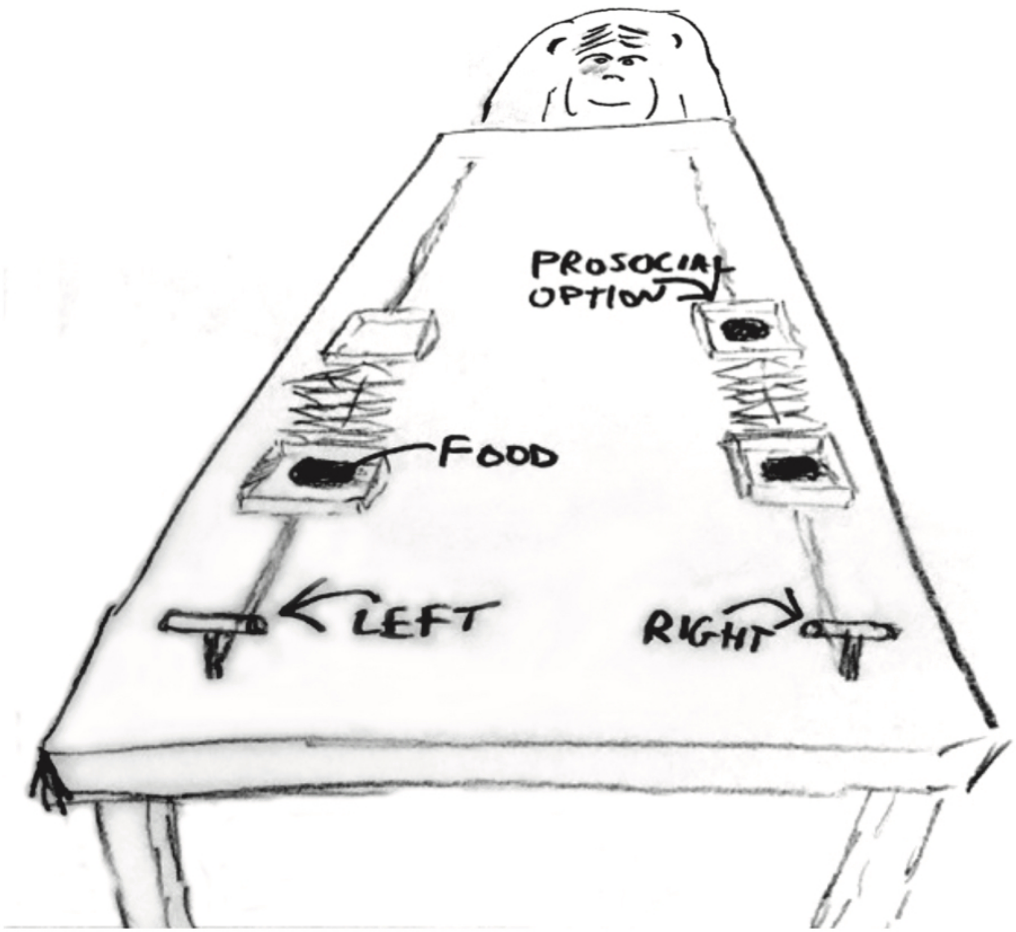
$$p_i = \frac{\exp(\alpha + \beta \cdot x_i)}{1 + \exp(\alpha + \beta \cdot x_i)}$$

Et le taux de changement de p en fonction du prédicteur x est donné par :

$$\frac{\partial p}{\partial x} = \frac{\beta}{2(1 + cosh(\alpha + \beta \cdot x)}$$

On voit que la variation sur p due au prédicteur x est fonction du prédicteur x… !

Exemple de régression logistique : La prosocialité chez le chimpanzé



Régression logistique

```r
library(tidyverse)
library(rethinking)

data(chimpanzees) # see ?chimpanzees for more information on the dataset
df1 <- chimpanzees
str(df1)
```

```
'data.frame':	504 obs. of  8 variables:
 $ actor       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ recipient   : int  NA NA NA NA NA NA NA NA NA NA ...
 $ condition   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ block       : int  1 1 1 1 1 1 2 2 2 2 ...
 $ trial       : int  2 4 6 8 10 12 14 16 18 20 ...
 $ prosoc_left : int  0 0 1 0 1 1 1 1 0 0 ...
 $ chose_prosoc: int  1 0 0 1 1 1 0 0 1 1 ...
 $ pulled_left : int  0 1 0 0 1 1 0 0 0 0 ...
```

- pulled_left : 1 lorsque le chimpanzé pousse le levier gauche, 0 sinon
- prosoc_left : 1 lorsque le levier gauche est associé à l'option prosociale, 0 sinon
- condition : 1 lorsqu'un partenaire est présent, 0 sinon

Régression logistique

LE PROBLÈME

On cherche à savoir si la présence d'un singe partenaire incite le chimpanzé à appuyer sur le levier prosocial, c'est à dire l'option qui donne de la nourriture aux deux individus. Autrement dit, est-ce qu'il existe une interaction entre l'effet de la latéralité et l'effet de la présence d'un autre chimpanzé sur la probabilité d'actionner le levier gauche.

LES VARIABLES

- Observations (`pulled_left`) : Ce sont des variables de Bernoulli. Elles prennent comme valeur 0/1.
- Prédicteur (`prosoc_left`) : Est-ce que les deux plats sont sur la gauche ou sur la droite ?
- Prédicteur (`condition`) : Est-ce qu'un partenaire est présent ?

Régression logistique

$$
\begin{aligned}
L_i &\sim \text{Binomial}(1, p_i) \\
\text{(equivalent to)} \quad L_i &\sim \text{Bernoulli}(p_i) \\
\text{logit}(p_i) &= \alpha \\
\alpha &\sim \text{Normal}(0, \omega)
\end{aligned}
$$

Modèle mathématique sans prédicteur. Comment choisir une valeur pour ω... ?

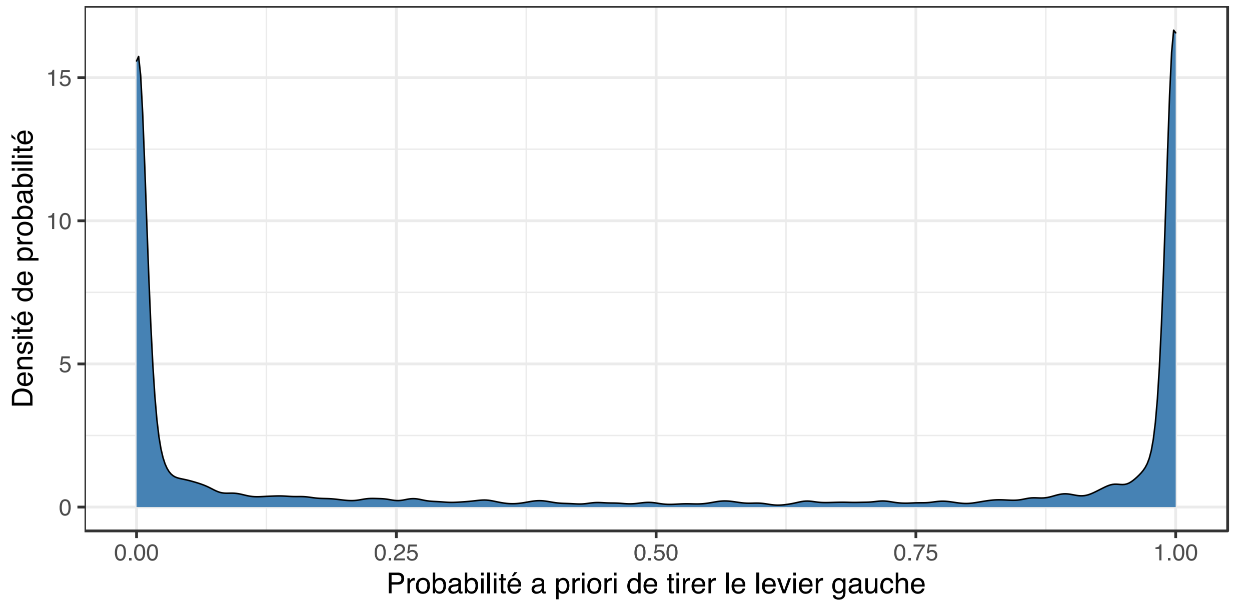
Prior predictive check

On écrit le modèle précédent avec `brms` et on échantillonne à partir du prior afin de vérifier que les prédictions du modèle (sur la base du prior seul) correspondent à nos attentes.

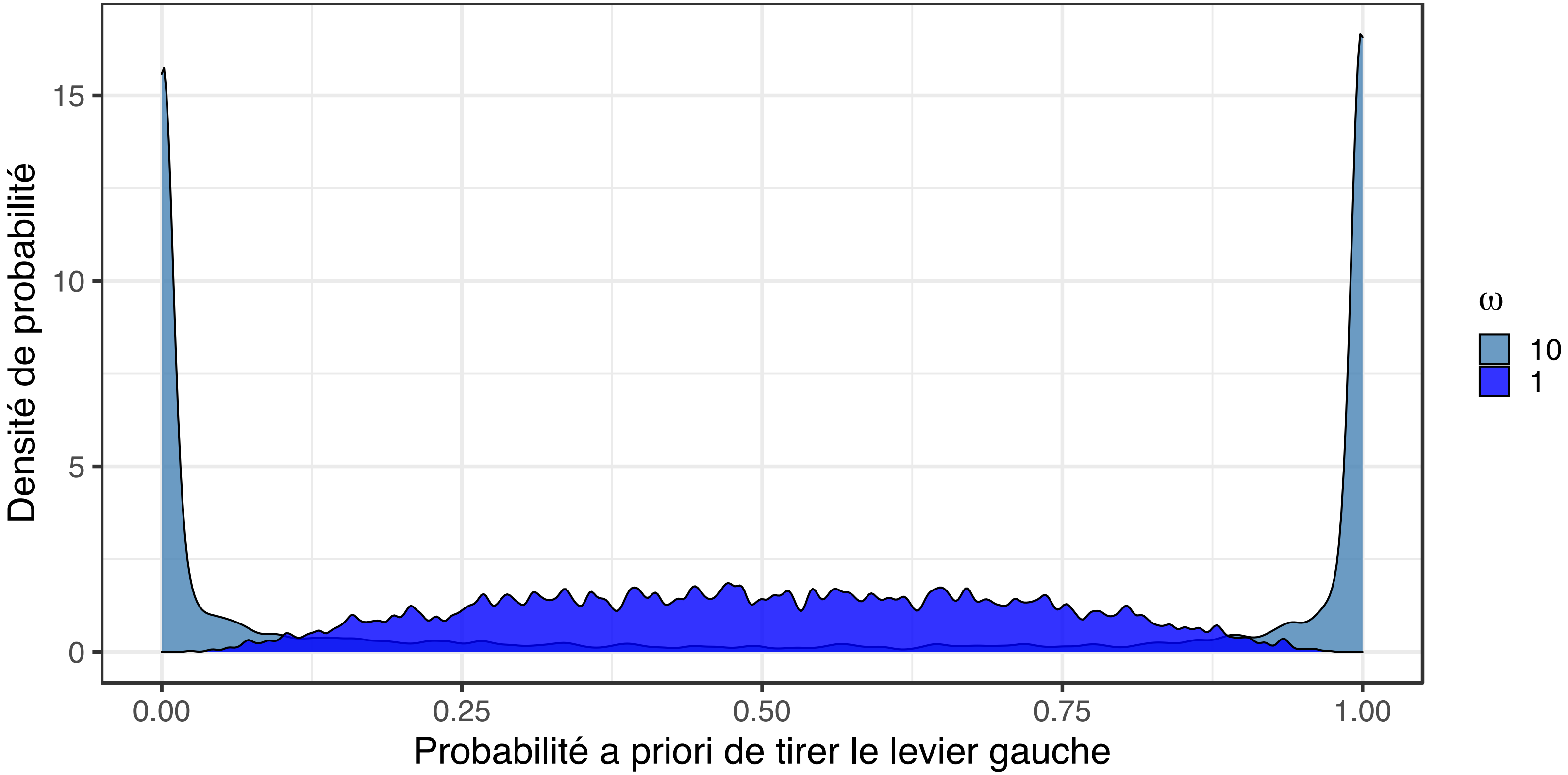
```r
library(brms)

mod1.1 <- brm(
  formula = pulled_left | trials(1) ~ 1,
  family = binomial,
  prior = set_prior("normal(0, 10)", class = "Intercept"),
  data = df1,
  # stores prior samples
  sample_prior = "yes"
  )
```

Prior predictive check

```r
# extracts prior samples
prior_samples(mod1.1) %>%
  # applies the inverse link function
  mutate(p = brms::inv_logit_scaled(Intercept) ) %>%
  ggplot(aes(x = p) ) +
  geom_density(fill = "steelblue", adjust = 0.1) +
  theme_bw(base_size = 20) +
  labs(x = "Probabilité a priori de tirer le levier gauche", y = "Densité de probabilité")
```



Prior predictive check



Régression logistique

L'intercept s'interprète dans l'espace des log-odds… pour l'interpréter comme une probabilité, il faut appliquer la fonction de lien inverse. On peut utiliser la fonction `rethinking::logistic()` ou la fonction `plogis()`.

Régression logistique

L'intercept s'interprète dans l'espace des log-odds... pour l'interpréter comme une probabilité, il faut appliquer la fonction de lien inverse. On peut utiliser la fonction `rethinking::logistic()` ou la fonction `plogis()`.

```
fixed_effects <- fixef(mod1.2) # effets fixes (ou constants)
rethinking::logistic(fixed_effects) # fonction de lien inverse
```

Régression logistique

L'intercept s'interprète dans l'espace des log-odds... pour l'interpréter comme une probabilité, il faut appliquer la fonction de lien inverse. On peut utiliser la fonction `rethinking::logistic()` ou la fonction `plogis()`.

```
fixed_effects <- fixef(mod1.2) # effets fixes (ou constants)
rethinking::logistic(fixed_effects) # fonction de lien inverse
```

```
          Estimate Est.Error      Q2.5     Q97.5
Intercept  0.57824 0.5224536 0.5354475 0.6207319
```

Régression logistique

L'intercept s'interprète dans l'espace des log-odds… pour l'interpréter comme une probabilité, il faut appliquer la fonction de lien inverse. On peut utiliser la fonction `rethinking::logistic()` ou la fonction `plogis()`.

```
fixed_effects <- fixef(mod1.2) # effets fixes (ou constants)
rethinking::logistic(fixed_effects) # fonction de lien inverse
```

```
           Estimate Est.Error      Q2.5     Q97.5
Intercept  0.57824 0.5224536 0.5354475 0.6207319
```

```
plogis(fixed_effects) # fonction de lien inverse
```

Régression logistique

L'intercept s'interprète dans l'espace des log-odds… pour l'interpréter comme une probabilité, il faut appliquer la fonction de lien inverse. On peut utiliser la fonction `rethinking::logistic()` ou la fonction `plogis()`.

```r
fixed_effects <- fixef(mod1.2) # effets fixes (ou constants)
rethinking::logistic(fixed_effects) # fonction de lien inverse
```

```
          Estimate Est.Error      Q2.5     Q97.5
Intercept  0.57824 0.5224536 0.5354475 0.6207319
```

```r
plogis(fixed_effects) # fonction de lien inverse
```

```
          Estimate Est.Error      Q2.5     Q97.5
Intercept  0.57824 0.5224536 0.5354475 0.6207319
```

Régression logistique

L'intercept s'interprète dans l'espace des log-odds... pour l'interpréter comme une probabilité, il faut appliquer la fonction de lien inverse. On peut utiliser la fonction `rethinking::logistic()` ou la fonction `plogis()`.

```
fixed_effects <- fixef(mod1.2) # effets fixes (ou constants)
rethinking::logistic(fixed_effects) # fonction de lien inverse
```

```
          Estimate Est.Error      Q2.5     Q97.5
Intercept  0.57824 0.5224536 0.5354475 0.6207319
```

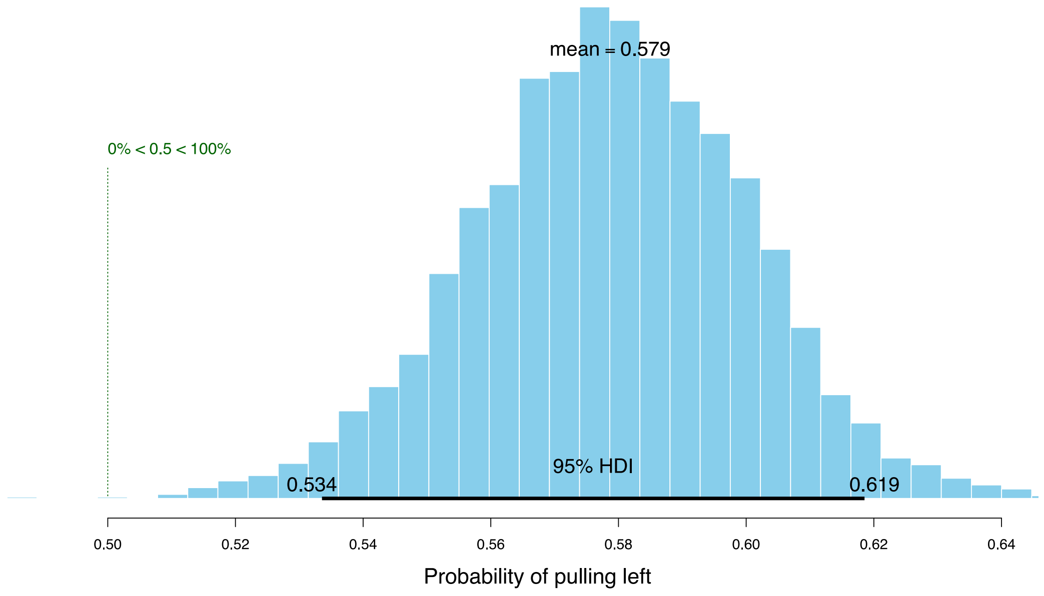
```
plogis(fixed_effects) # fonction de lien inverse
```

```
          Estimate Est.Error      Q2.5     Q97.5
Intercept  0.57824 0.5224536 0.5354475 0.6207319
```

En moyenne (sans considérer les prédicteurs), il semblerait que les singes aient plus tendance à appuyer sur le levier gauche que sur le levier droit...

Régression logistique

```
post <- posterior_samples(mod1.2)
intercept_samples <- plogis(post$b_Intercept)

library(BEST)
library(coda)

plotPost(intercept_samples, compVal = 0.5, xlab = "Probability of pulling left")
```



Régression logistique

Et si on ajoute des prédicteurs… comment choisir une valeur pour ω ?

$$
\begin{aligned}
L_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \alpha + \beta_P P_i + \beta_C C_i + \beta_{PC} P_i C_i \\
\alpha &\sim \text{Normal}(0, 1) \\
\beta_P, \beta_C, \beta_{PC} &\sim \text{Normal}(0, \omega)
\end{aligned}
$$

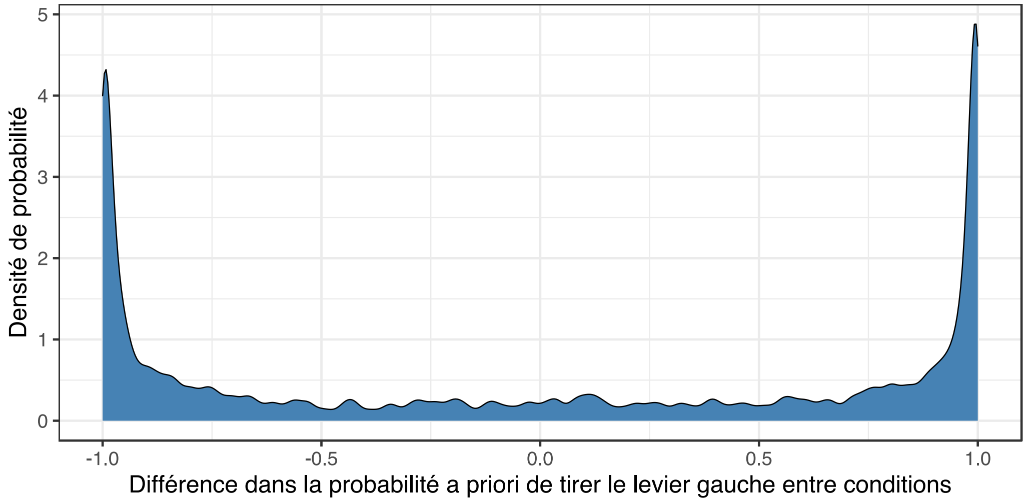
- L_i indique si le singe a poussé le levier gauche (`pulled_left`)
- P_i indique si le coté gauche correspond au coté prosocial
- C_i indique la présence d'un partenaire

Régression logistique

```
# recoding predictors
df1 <- df1 %>%
  mutate(
    prosoc_left = ifelse(prosoc_left == 1, 0.5, -0.5),
    condition = ifelse(condition == 1, 0.5, -0.5)
    )

priors <- c(
  set_prior("normal(0, 1)", class = "Intercept"),
  set_prior("normal(0, 10)", class = "b")
  )

mod2.1 <- brm(
  formula = pulled_left | trials(1) ~ 1 + prosoc_left * condition,
  family = binomial,
  prior = priors,
  data = df1,
  sample_prior = "yes"
  )
```

Prior predictive check

```
prior_samples(mod2.1) %>%
  mutate(
    condition1 = plogis(Intercept - 0.5 * b),
    condition2 = plogis(Intercept + 0.5 * b)
    ) %>%
  ggplot(aes(x = condition2 - condition1) ) +
  geom_density(fill = "steelblue", adjust = 0.1) +
  theme_bw(base_size = 20) +
  labs(
    x = "Différence dans la probabilité a priori de tirer le levier gauche entre conditions",
    y = "Densité de probabilité"
    )
```

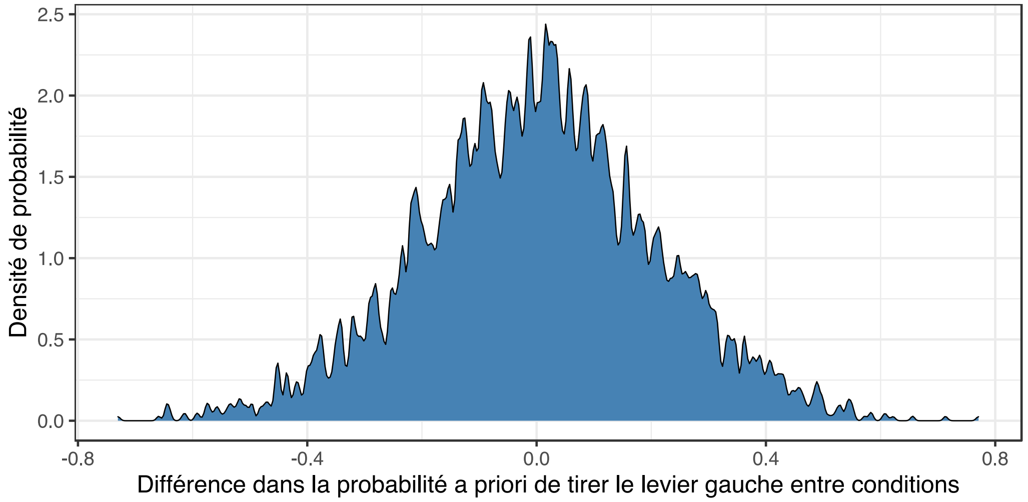


Régression logistique

```r
priors <- c(
  set_prior("normal(0, 1)", class = "Intercept"),
  set_prior("normal(0, 1)", class = "b")
  )

mod2.2 <- brm(
  formula = pulled_left | trials(1) ~ 1 + prosoc_left * condition,
  family = binomial,
  prior = priors,
  data = df1,
  sample_prior = "yes"
  )
```

Prior predictive check

```
prior_samples(mod2.2) %>%
  mutate(
    condition1 = plogis(Intercept - 0.5 * b),
    condition2 = plogis(Intercept + 0.5 * b)
    ) %>%
  ggplot(aes(x = condition2 - condition1) ) +
  geom_density(fill = "steelblue", adjust = 0.1) +
  theme_bw(base_size = 20) +
  labs(
    x = "Différence dans la probabilité a priori de tirer le levier gauche entre conditions",
    y = "Densité de probabilité"
    )
```



Régression logistique

```r
summary(mod2.2)
```

```
 Family: binomial
  Links: mu = logit
Formula: pulled_left | trials(1) ~ 1 + prosoc_left * condition
   Data: df1 (Number of observations: 504)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
                      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
Intercept                 0.33      0.09     0.15     0.50 1.00     5207
prosoc_left               0.54      0.18     0.18     0.89 1.00     4623
condition                -0.20      0.18    -0.55     0.16 1.00     5355
prosoc_left:condition     0.16      0.35    -0.54     0.84 1.00     4111
                      Tail_ESS
Intercept                 3055
prosoc_left               2990
condition                 3396
prosoc_left:condition     3224

Samples were drawn 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).
```

Effet relatif vs. Effet absolu

Le modèle linéaire ne prédit pas directement la probabilité mais le log-odds de la probabilité :

Effet relatif vs. Effet absolu

Le modèle linéaire ne prédit pas directement la probabilité mais le log-odds de la probabilité :

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta x_i$$

Effet relatif vs. Effet absolu

Le modèle linéaire ne prédit pas directement la probabilité mais le log-odds de la probabilité :

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta x_i$$

On peut donc distinguer et interpréter deux types d'effets.

Effet relatif vs. Effet absolu

Le modèle linéaire ne prédit pas directement la probabilité mais le log-odds de la probabilité :

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i$$

On peut donc distinguer et interpréter deux types d'effets.

Effet relatif : L'effet relatif porte sur le logarithme du rapport des probabilités. Il indique une *proportion* de changement induit par le prédicteur sur *les chances* de succès (ou plutôt, sur la cote). Cela ne nous dit rien de la probabilité de l'événement, dans l'absolu.

Effet relatif vs. Effet absolu

Le modèle linéaire ne prédit pas directement la probabilité mais le log-odds de la probabilité :

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i$$

On peut donc distinguer et interpréter deux types d'effets.

Effet relatif : L'effet relatif porte sur le logarithme du rapport des probabilités. Il indique une *proportion* de changement induit par le prédicteur sur *les chances* de succès (ou plutôt, sur la cote). Cela ne nous dit rien de la probabilité de l'événement, dans l'absolu.

Effet absolu : Effet qui porte directement sur la probabilité d'un événement. Il dépend de tous les paramètres du modèle et nous donne l'impact effectif d'un changement d'une unité d'un prédicteur (dans l'espace des probabilités).

Effet relatif

Il s'agit d'une **proportion** de changement induit par le prédicteur sur le rapport des chances ou “cote” (*odds*). Illustration avec un modèle sans interaction.

Effet relatif

Il s'agit d'une **proportion** de changement induit par le prédicteur sur le rapport des chances ou “cote” (*odds*). Illustration avec un modèle sans interaction.

$$
\begin{aligned}
\log\left(\frac{p_i}{1-p_i}\right) &= \alpha + \beta x_i \\
\frac{p_i}{1-p_i} &= \exp(\alpha + \beta x_i)
\end{aligned}
$$

Effet relatif

Il s'agit d'une **proportion** de changement induit par le prédicteur sur le rapport des chances ou "cote" (*odds*). Illustration avec un modèle sans interaction.

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) &= \alpha + \beta x_i \\ \frac{p_i}{1-p_i} &= \exp(\alpha + \beta x_i)\end{aligned}$$

La cote proportionnelle q d'un événement est le nombre par lequel la cote est multipliée lorsque x_i augmente d'une unité.

Effet relatif

Il s'agit d'une **proportion** de changement induit par le prédicteur sur le rapport des chances ou "cote" (*odds*). Illustration avec un modèle sans interaction.

$$\begin{aligned} \log\left(\frac{p_i}{1-p_i}\right) &= \alpha + \beta x_i \\ \frac{p_i}{1-p_i} &= \exp(\alpha + \beta x_i) \end{aligned}$$

La cote proportionnelle q d'un événement est le nombre par lequel la cote est multipliée lorsque x_i augmente d'une unité.

$$q = \frac{\exp(\alpha + \beta(x_i + 1))}{\exp(\alpha + \beta x_i)} = \frac{\exp(\alpha)\exp(\beta x_i)\exp(\beta)}{\exp(\alpha)\exp(\beta x_i)} = \exp(\beta)$$

Effet relatif

Il s'agit d'une **proportion** de changement induit par le prédicteur sur le rapport des chances ou “cote” (*odds*). Illustration avec un modèle sans interaction.

$$\begin{aligned} \log\left(\frac{p_i}{1-p_i}\right) &= \alpha + \beta x_i \\ \frac{p_i}{1-p_i} &= \exp(\alpha + \beta x_i) \end{aligned}$$

La cote proportionnelle q d'un événement est le nombre par lequel la cote est multipliée lorsque x_i augmente d'une unité.

$$q = \frac{\exp(\alpha + \beta(x_i + 1))}{\exp(\alpha + \beta x_i)} = \frac{\exp(\alpha)\exp(\beta x_i)\exp(\beta)}{\exp(\alpha)\exp(\beta x_i)} = \exp(\beta)$$

Lorsque $q = 2$ (par exemple), une augmentation de x_i d'une unité génère un doublement de la cote.

Interprétation de l'effet relatif

L'effet relatif d'un paramètre **dépend également des autres paramètres**. Dans le modèle précédent, le prédicteur `prosoc_left` augmente le log de la cote d'environ 0.54, ce qui se traduit par une augmentation de la cote de $\exp(0.54) \approx 1.72$ soit une augmentation d'environ 72% de la cote.

Interprétation de l'effet relatif

L'effet relatif d'un paramètre **dépend également des autres paramètres**. Dans le modèle précédent, le prédicteur `prosoc_left` augmente le log de la cote d'environ 0.54, ce qui se traduit par une augmentation de la cote de $\exp(0.54) \approx 1.72$ soit une augmentation d'environ 72% de la cote.

Supposons que l'intercept $\alpha = 4$.

Interprétation de l'effet relatif

L'effet relatif d'un paramètre **dépend également des autres paramètres**. Dans le modèle précédent, le prédicteur `prosoc_left` augmente le log de la cote d'environ 0.54, ce qui se traduit par une augmentation de la cote de $\exp(0.54) \approx 1.72$ soit une augmentation d'environ 72% de la cote.

Supposons que l'intercept $\alpha = 4$.

- La probabilité de pousser le levier sans autre considération est de $\operatorname{logit}^{-1}(4) = 0.98$.

Interprétation de l'effet relatif

L'effet relatif d'un paramètre **dépend également des autres paramètres**. Dans le modèle précédent, le prédicteur `prosoc_left` augmente le log de la cote d'environ 0.54, ce qui se traduit par une augmentation de la cote de $\exp(0.54) \approx 1.72$ soit une augmentation d'environ 72% de la cote.

Supposons que l'intercept $\alpha = 4$.

- La probabilité de pousser le levier sans autre considération est de $\mathrm{logit}^{-1}(4) = 0.98$.
- En considérant l'effet de `prosoc_left`, on obtient $\mathrm{logit}^{-1}(4 + 0.54) \approx 0.99$.

Interprétation de l’effet relatif

L'effet relatif d'un paramètre **dépend également des autres paramètres**. Dans le modèle précédent, le prédicteur `prosoc_left` augmente le log de la cote d'environ 0.54, ce qui se traduit par une augmentation de la cote de $\exp(0.54) \approx 1.72$ soit une augmentation d'environ 72% de la cote.

Supposons que l'intercept $\alpha = 4$.

- La probabilité de pousser le levier sans autre considération est de $\text{logit}^{-1}(4) = 0.98$.
- En considérant l'effet de `prosoc_left`, on obtient $\text{logit}^{-1}(4 + 0.54) \approx 0.99$.

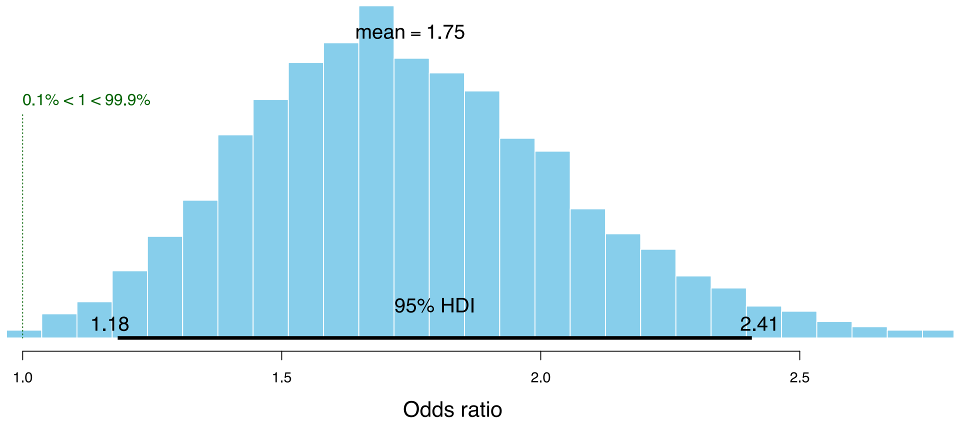
Une augmentation de 72% sur le log-odds se traduit par une augmentation de seulement 1% sur la probabilité effective... Les effets relatifs peuvent conduire à de mauvaises interprétations lorsqu'on ne considère pas l'échelle de la variable mesurée.

Interprétation de l'effet relatif

```r
fixef(mod2.2)
```

```
                          Estimate  Est.Error       Q2.5     Q97.5
Intercept                0.3256619 0.08927425  0.1547740 0.5008530
prosoc_left              0.5420179 0.18147340  0.1801776 0.8925005
condition               -0.1965239 0.18160463 -0.5515497 0.1577763
prosoc_left:condition    0.1567847 0.34932415 -0.5408243 0.8383583
```

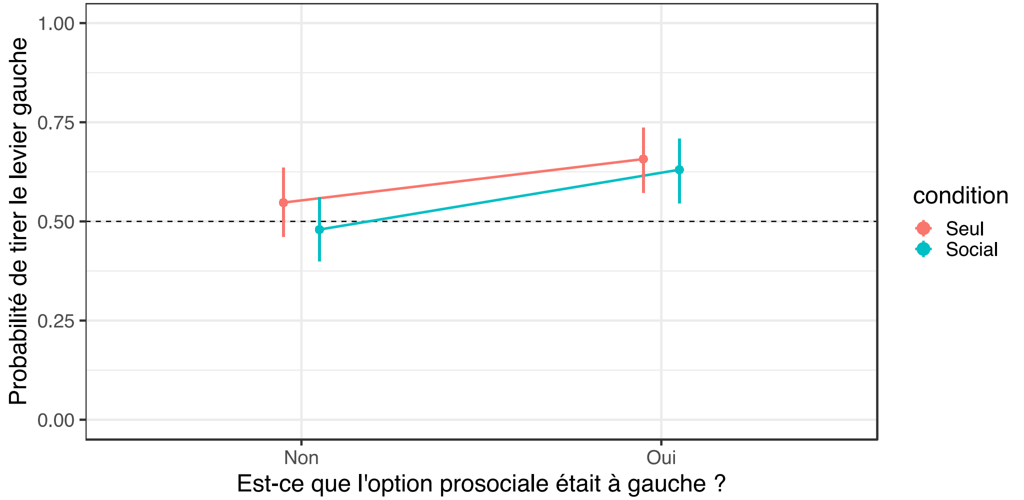
```r
post <- posterior_samples(mod2.2)
plotPost(exp(post$b_prosoc_left), compVal = 1, xlab = "Odds ratio")
```



Effet absolu

L'effet absolu dépend de tous les paramètres du modèle et nous donne l'impact effectif d'un changement d'une unité d'un prédicteur (dans l'espace des probabilités).

```
model_predictions <- fitted(mod2.2) %>%
  data.frame() %>%
  bind_cols(df1) %>%
  mutate(
    condition = factor(condition),
    prosoc_left = factor(prosoc_left)
    )
```



Régression binomiale agrégée

Ces données représentent le nombre de candidatures à l'université de Berkeley par sexe et par département. Chaque candidature est acceptée ou rejetée et les résultats sont agrégés par département et par sexe.

```r
library(rethinking)
data(UCBadmit)
(df2 <- UCBadmit)
```

```
   dept applicant.gender admit reject applications
1     A             male   512    313          825
2     A           female    89     19          108
3     B             male   353    207          560
4     B           female    17      8           25
5     C             male   120    205          325
6     C           female   202    391          593
7     D             male   138    279          417
8     D           female   131    244          375
9     E             male    53    138          191
10    E           female    94    299          393
11    F             male    22    351          373
12    F           female    24    317          341
```

Existe-t-il un biais de recrutement lié au sexe ?

Régression binomiale agrégée

On va construire un modèle de la décision d'admission en prenant comme prédicteur le sexe du candidat.

$$
\begin{aligned}
\text{admit}_i &\sim \text{Binomial}(n_i, p_i) \\
\text{logit}(p_i) &= \alpha + \beta_m m_i \\
\alpha &\sim \text{Normal}(0, 1) \\
\beta_m &\sim \text{Normal}(0, 1)
\end{aligned}
$$

Les variables :

- admit_i : Le nombre de candidatures acceptées (`admit`)
- n_i : Le nombre total de candidatures (`applications`)
- m_i : Le sexe du candidat (`1 = male`)

Régression binomiale agrégée

```
priors <- c(set_prior("normal(0, 1)", class = "Intercept") )

mod3 <- brm(
  formula = admit | trials(applications) ~ 1,
  family = binomial(link = "logit"),
  prior = priors,
  data = df2,
  sample_prior = "yes"
  )
```

Régression binomiale agrégée

```r
priors <- c(
  set_prior("normal(0, 1)", class = "Intercept"),
  set_prior("normal(0, 1)", class = "b")
  )

# dummy-coding
df2$male <- ifelse(df2$applicant.gender == "male", 1, 0)

mod4 <- brm(
  formula = admit | trials(applications) ~ 1 + male,
  family = binomial(link = "logit"),
  prior = priors,
  data = df2,
  sample_prior = "yes"
  )
```

Régression binomiale agrégée

```r
summary(mod4)
```

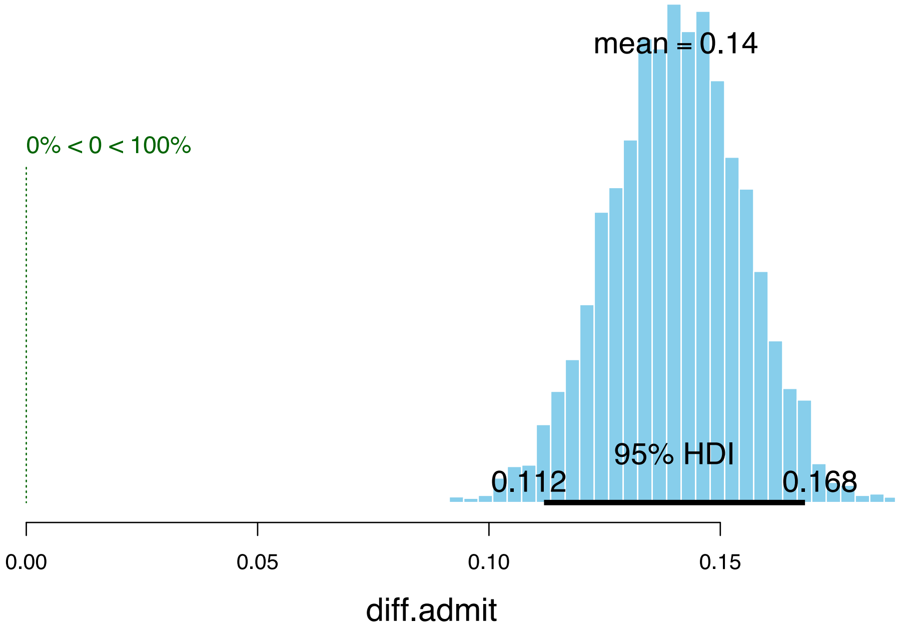
```
 Family: binomial
  Links: mu = logit
Formula: admit | trials(applications) ~ 1 + male
   Data: df2 (Number of observations: 12)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.83      0.05    -0.93    -0.72 1.00     2175     2208
male          0.61      0.06     0.48     0.73 1.00     2563     2147

Samples were drawn 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).
```

Être un homme semble être un avantage… ! Le rapport des cotes est de $\exp(0.61) \approx 1.84$.

Régression binomiale agrégée

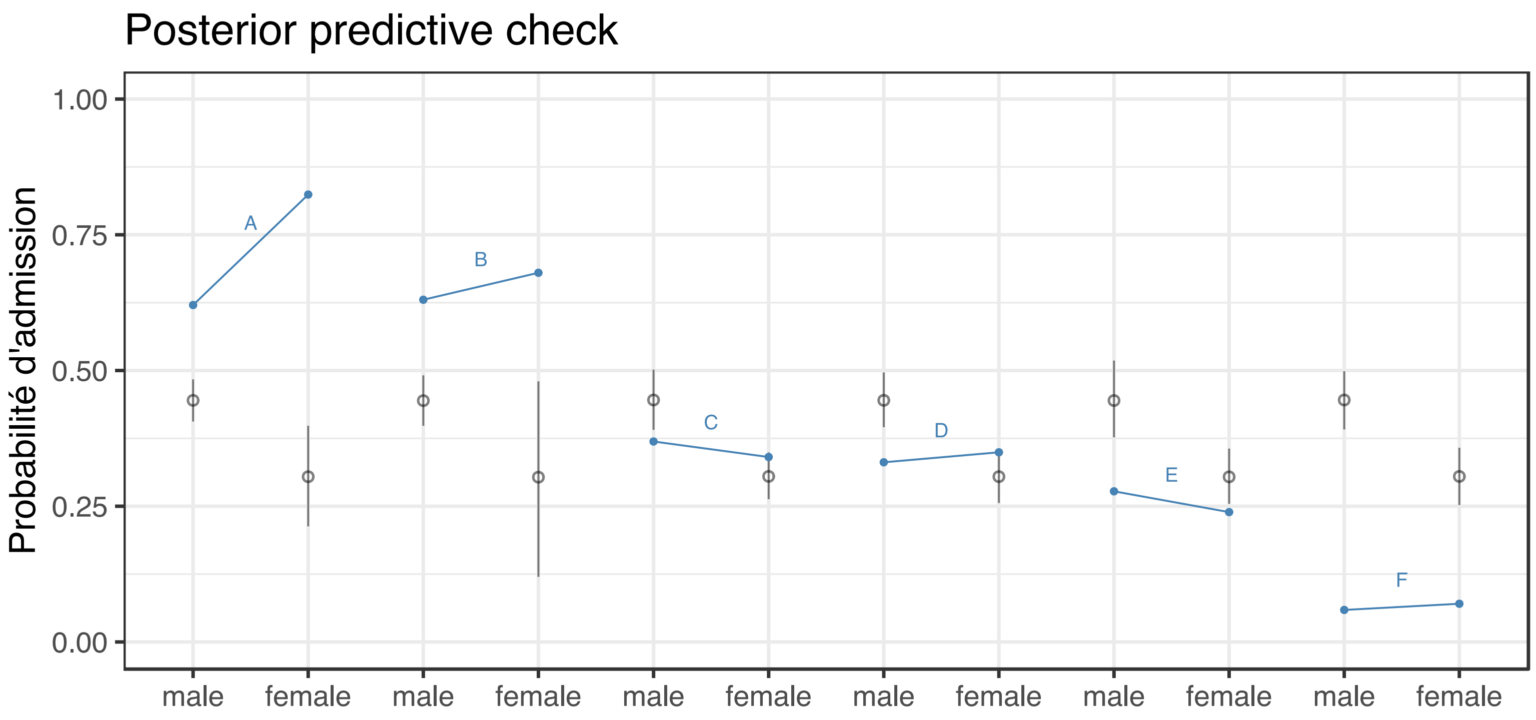
Calculons la différence de probabilité d'admission entre hommes et femmes.

```
post <- posterior_samples(mod4)
p.admit.male <- plogis(post$b_Intercept + post$b_male)
p.admit.female <- plogis(post$b_Intercept)
diff.admit <- p.admit.male - p.admit.female
plotPost(diff.admit, compVal = 0)
```



Visualiser les prédictions du modèle

On examine les prédictions du modèle par département.



Régression binomiale agrégée

Les prédictions du modèle sont très mauvaises… Il n'y a que deux départements pour lesquels les femmes ont de moins bonnes prédictions que les hommes (C et E) alors que le modèle prédit une probabilité d'admission plus basse pour tous les départements…

Régression binomiale agrégée

Les prédictions du modèle sont très mauvaises… Il n'y a que deux départements pour lesquels les femmes ont de moins bonnes prédictions que les hommes (C et E) alors que le modèle prédit une probabilité d'admission plus basse pour tous les départements…

Le problème est double :

Régression binomiale agrégée

Les prédictions du modèle sont très mauvaises... Il n'y a que deux départements pour lesquels les femmes ont de moins bonnes prédictions que les hommes (C et E) alors que le modèle prédit une probabilité d'admission plus basse pour tous les départements...

Le problème est double :

- Les hommes et les femmes ne postulent pas aux mêmes départements

Régression binomiale agrégée

Les prédictions du modèle sont très mauvaises… Il n'y a que deux départements pour lesquels les femmes ont de moins bonnes prédictions que les hommes (C et E) alors que le modèle prédit une probabilité d'admission plus basse pour tous les départements…

Le problème est double :

- Les hommes et les femmes ne postulent pas aux mêmes départements
- Les départements n'ont pas tous les mêmes effectifs

Régression binomiale agrégée

Les prédictions du modèle sont très mauvaises… Il n'y a que deux départements pour lesquels les femmes ont de moins bonnes prédictions que les hommes (C et E) alors que le modèle prédit une probabilité d'admission plus basse pour tous les départements…

Le problème est double :

- Les hommes et les femmes ne postulent pas aux mêmes départements
- Les départements n'ont pas tous les mêmes effectifs

C'est le paradoxe de Simpson… remarques :

Régression binomiale agrégée

Les prédictions du modèle sont très mauvaises… Il n'y a que deux départements pour lesquels les femmes ont de moins bonnes prédictions que les hommes (C et E) alors que le modèle prédit une probabilité d'admission plus basse pour tous les départements…

Le problème est double :

- Les hommes et les femmes ne postulent pas aux mêmes départements
- Les départements n'ont pas tous les mêmes effectifs

C'est le paradoxe de Simpson… remarques :

- La distribution postérieure seule n'aurait pas permis de détecter ce problème

Régression binomiale agrégée

Les prédictions du modèle sont très mauvaises… Il n'y a que deux départements pour lesquels les femmes ont de moins bonnes prédictions que les hommes (C et E) alors que le modèle prédit une probabilité d'admission plus basse pour tous les départements…

Le problème est double :

- Les hommes et les femmes ne postulent pas aux mêmes départements
- Les départements n'ont pas tous les mêmes effectifs

C'est le paradoxe de Simpson… remarques :

- La distribution postérieure seule n'aurait pas permis de détecter ce problème
- C'est l'étude des prédictions du modèle qui nous a permis de mettre le doigt sur le problème…

Régression binomiale agrégée

On construit donc un modèle de la décision d'admission en fonction du genre, *au sein de chaque département*.

$$
\begin{aligned}
\text{admit}_i &\sim \text{Binomial}(n_i, p_i) \\
\text{logit}(p_i) &= \alpha_{\text{dept}[i]} + \beta_m m_i \\
\alpha_{\text{dept}[i]} &\sim \text{Normal}(0, 1) \\
\beta_m &\sim \text{Normal}(0, 1)
\end{aligned}
$$

Régression binomiale aggrégée

```r
# modèle sans prédicteur
mod5 <- brm(
  admit | trials(applications) ~ 0 + dept,
  family = binomial(link = "logit"),
  prior = set_prior("normal(0, 1)", class = "b"),
  data = df2
  )

# modèle avec prédicteur
mod6 <- brm(
  admit | trials(applications) ~ 0 + dept + male,
  family = binomial(link = "logit"),
  prior = set_prior("normal(0, 1)", class = "b"),
  data = df2
  )
```

Régression binomiale aggrégée

```
summary(mod6)
```

```
 Family: binomial
  Links: mu = logit
Formula: admit | trials(applications) ~ 0 + dept + male
   Data: df2 (Number of observations: 12)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
deptA     0.69      0.10     0.50     0.88 1.00     1940     2881
deptB     0.64      0.11     0.43     0.88 1.00     2317     2754
deptC    -0.57      0.08    -0.72    -0.43 1.00     3262     3136
deptD    -0.61      0.08    -0.77    -0.44 1.00     2751     3087
deptE    -1.05      0.09    -1.23    -0.87 1.00     4073     2585
deptF    -2.57      0.15    -2.87    -2.29 1.00     3984     3143
male     -0.11      0.08    -0.26     0.04 1.00     1572     2479

Samples were drawn 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).
```

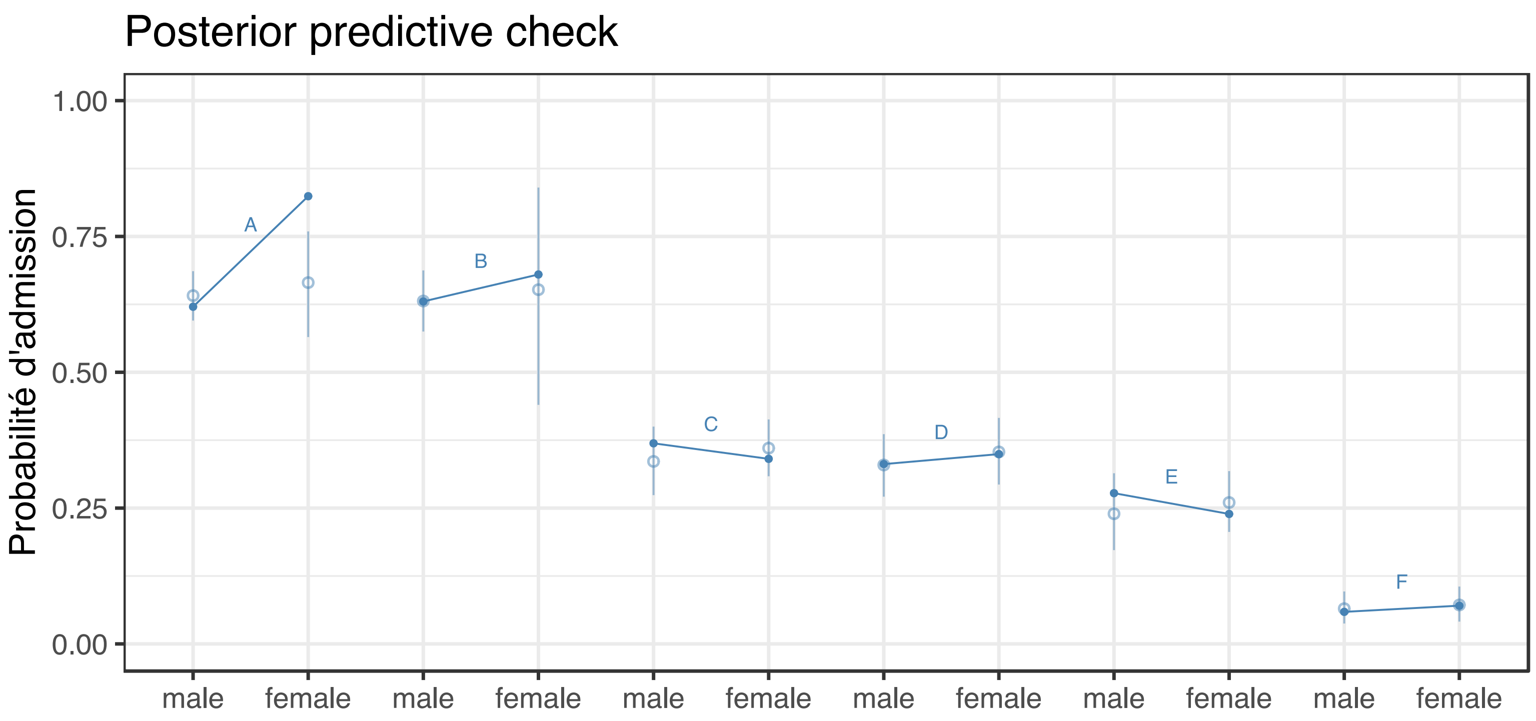
Régression binomiale agrégée

```
fixef(mod6)
```

```
        Estimate  Est.Error       Q2.5       Q97.5
deptA  0.6883208 0.09681919  0.4963408  0.87929763
deptB  0.6439304 0.11263513  0.4264837  0.87517762
deptC -0.5741505 0.07513734 -0.7222238 -0.43027696
deptD -0.6061606 0.08424317 -0.7697554 -0.44358589
deptE -1.0467390 0.09327729 -1.2303495 -0.87008308
deptF -2.5714851 0.14964839 -2.8655062 -2.28795971
male  -0.1074853 0.07808494 -0.2612262  0.04122383
```

Maintenant, la prédiction pour β_m va dans l'autre sens... La rapport des cotes (odds ratio) est de $\exp(-0.1)$ soit 90% de la cote des femmes.

Régression binomiale agrégée



Conclusions

Les hommes et les femmes ne postulent pas aux mêmes départements et les départements varient par leur probabilité d'admission. En l'occurrence, les femmes ont plus postulé aux départements E et F (avec une probabilité d'admission plus faible) et ont moins postulé aux départements A ou B, avec une probabilité d'admission plus haute.

Conclusions

Les hommes et les femmes ne postulent pas aux mêmes départements et les départements varient par leur probabilité d'admission. En l'occurrence, les femmes ont plus postulé aux départements E et F (avec une probabilité d'admission plus faible) et ont moins postulé aux départements A ou B, avec une probabilité d'admission plus haute.

Pour évaluer l'effet du sexe sur la probabilité d'admission, il faut donc se poser la question suivante : “Quelle est la différence de probabilité d'admission entre hommes et femmes *au sein de chaque département* ?” (plutôt que de manière générale).

Conclusions

Les hommes et les femmes ne postulent pas aux mêmes départements et les départements varient par leur probabilité d'admission. En l'occurrence, les femmes ont plus postulé aux départements E et F (avec une probabilité d'admission plus faible) et ont moins postulé aux départements A ou B, avec une probabilité d'admission plus haute.

Pour évaluer l'effet du sexe sur la probabilité d'admission, il faut donc se poser la question suivante : “Quelle est la différence de probabilité d'admission entre hommes et femmes *au sein de chaque département* ?” (plutôt que de manière générale).

Retenir que le modèle de régression peut être généralisé à différents modèles de génération des données (i.e., différentes distributions de probabilité, comme la distribution Normale, Binomiale, Poisson, etc) et que l'espace des paramètres peut-être “connecté” à l'espace des prédicteurs (variables mesurées) grâce à des fonctions de lien (e.g., la fonction logarithme, exponentielle, logit, etc).

Conclusions

Les hommes et les femmes ne postulent pas aux mêmes départements et les départements varient par leur probabilité d'admission. En l'occurrence, les femmes ont plus postulé aux départements E et F (avec une probabilité d'admission plus faible) et ont moins postulé aux départements A ou B, avec une probabilité d'admission plus haute.

Pour évaluer l'effet du sexe sur la probabilité d'admission, il faut donc se poser la question suivante : "Quelle est la différence de probabilité d'admission entre hommes et femmes *au sein de chaque département* ?" (plutôt que de manière générale).

Retenir que le modèle de régression peut être généralisé à différents modèles de génération des données (i.e., différentes distributions de probabilité, comme la distribution Normale, Binomiale, Poisson, etc) et que l'espace des paramètres peut-être "connecté" à l'espace des prédicteurs (variables mesurées) grâce à des fonctions de lien (e.g., la fonction logarithme, exponentielle, logit, etc).

Retenir la distinction entre *effet relatif* (e.g., un changement de cote) et *effet absolu* (e.g., une différence de probabilité).

Travaux pratiques - Absentéisme expérimental

Travailler avec des sujets humains implique un minimum de coopération réciproque. Mais ce n'est pas toujours le cas. Une partie non-négligeable des étudiants qui s'inscrivent pour passer des expériences de Psychologie ne se présentent pas le jour prévu... On a voulu estimer la **probabilité de présence d'un étudiant inscrit** en fonction de l'envoi (ou non) d'un mail de rappel (cet exemple est présenté en détails dans deux blogposts, accessibles ici, et ici).

Travaux pratiques - Absentéisme expérimental

Travailler avec des sujets humains implique un minimum de coopération réciproque. Mais ce n'est pas toujours le cas. Une partie non-négligeable des étudiants qui s'inscrivent pour passer des expériences de Psychologie ne se présentent pas le jour prévu... On a voulu estimer la **probabilité de présence d'un étudiant inscrit** en fonction de l'envoi (ou non) d'un mail de rappel (cet exemple est présenté en détails dans deux blogposts, accessibles ici, et ici).

```
df3 <- read.csv("data/absence.csv")
df3 %>% sample_frac %>% head(10)
```

Travaux pratiques - Absentéisme expérimental

Travailler avec des sujets humains implique un minimum de coopération réciproque. Mais ce n'est pas toujours le cas. Une partie non-négligeable des étudiants qui s'inscrivent pour passer des expériences de Psychologie ne se présentent pas le jour prévu... On a voulu estimer la **probabilité de présence d'un étudiant inscrit** en fonction de l'envoi (ou non) d'un mail de rappel (cet exemple est présenté en détails dans deux blogposts, accessibles ici, et ici).

```
df3 <- read.csv("data/absence.csv")
df3 %>% sample_frac %>% head(10)
```

```
         day inscription reminder absence presence total
1     Friday      doodle       no       7       11    18
2     Monday      doodle      yes       2        6     8
3  Wednesday      doodle       no       6       11    17
4     Monday      doodle       no       5        4     9
5    Tuesday      doodle      yes       1        7     8
6    Tuesday      doodle       no       4       10    14
7     Monday       panel      yes       6       12    18
8     Friday       panel      yes       0       10    10
9  Wednesday       panel      yes       0       14    14
10    Friday      doodle      yes       0        2     2
```

Travaux pratiques

1. **Quelle est la probabilité qu'un participant, qui s'est inscrit de son propre chef, vienne effectivement passer l'expérience ?**
2. Quel est l'effet du rappel ?
3. Quel est l'effet du mode d'inscription ?
4. Quel est l'effet conjoint de ces deux prédicteurs ?

Travaux pratiques

Écrire le modèle qui prédit la présence d'un participant sans prédicteur.

$$
\begin{aligned}
y_i &\sim \mathrm{Binomial}(n_i, p_i) \\
\mathrm{logit}(p_i) &= \alpha \\
\alpha &\sim \mathrm{Normal}(0, 1)
\end{aligned}
$$

Travaux pratiques

```r
mod7 <- brm(
    presence | trials(total) ~ 1,
    family = binomial(link = "logit"),
    prior = set_prior("normal(0, 1)", class = "Intercept"),
    data = df3,
    cores = parallel::detectCores()
    )
```

```r
fixef(mod7) # effet relatif (log de la cote)
```

```
          Estimate Est.Error      Q2.5    Q97.5
Intercept 1.154494 0.1897284 0.7897511 1.534535
```

```r
fixef(mod7) %>% plogis # effet absolu (probabilité de présence)
```

```
           Estimate Est.Error      Q2.5     Q97.5
Intercept 0.7603308 0.5472903 0.6877779 0.8226688
```

Travaux pratiques

1. Quelle est la probabilité qu'un participant, qui s'est inscrit de son propre chef, vienne effectivement passer l'expérience ?
2. **Quel est l'effet du rappel ?**
3. Quel est l'effet du mode d'inscription ?
4. Quel est l'effet conjoint de ces deux prédicteurs ?

Travaux pratiques

On commence par re-coder en dummy variables le *reminder* et l'*inscription*.

```
df3 <-
  df3 %>%
  mutate(
    reminder = ifelse(reminder == "no", 0, 1),
    inscription = ifelse(inscription == "panel", 0, 1)
    )

head(df3, n = 10)
```

```
         day inscription reminder absence presence total
1     Friday           1        0       7       11    18
2     Friday           1        1       0        2     2
3     Friday           0        1       0       10    10
4     Monday           1        0       5        4     9
5     Monday           1        1       2        6     8
6     Monday           0        1       6       12    18
7   Thursday           1        0       3       11    14
8    Tuesday           1        0       4       10    14
9    Tuesday           1        1       1        7     8
10   Tuesday           0        1       0        9     9
```

Travaux pratiques

Écrire le modèle qui prédit la présence en fonction du rappel.

$$
\begin{aligned}
y_i &\sim \mathrm{Binomial}(n_i, p_i) \\
\mathrm{logit}(p_i) &= \alpha + \beta \times \mathrm{reminder}_i \\
\alpha &\sim \mathrm{Normal}(0, 1) \\
\beta &\sim \mathrm{Normal}(0, 1)
\end{aligned}
$$

Travaux pratiques

Écrire le modèle qui prédit la présence en fonction du rappel.

```
priors <- c(
  set_prior("normal(0, 1)", class = "Intercept"),
  set_prior("normal(0, 1)", class = "b")
  )

mod8 <- brm(
    presence | trials(total) ~ 1 + reminder,
    family = binomial(link = "logit"),
    prior = priors,
    data = df3,
    cores = parallel::detectCores()
    )
```

Travaux pratiques

Quel est l'effet **relatif** du mail de rappel ?

```
exp(fixef(mod8)[2]) # odds ratio between no-reminder and reminder
```

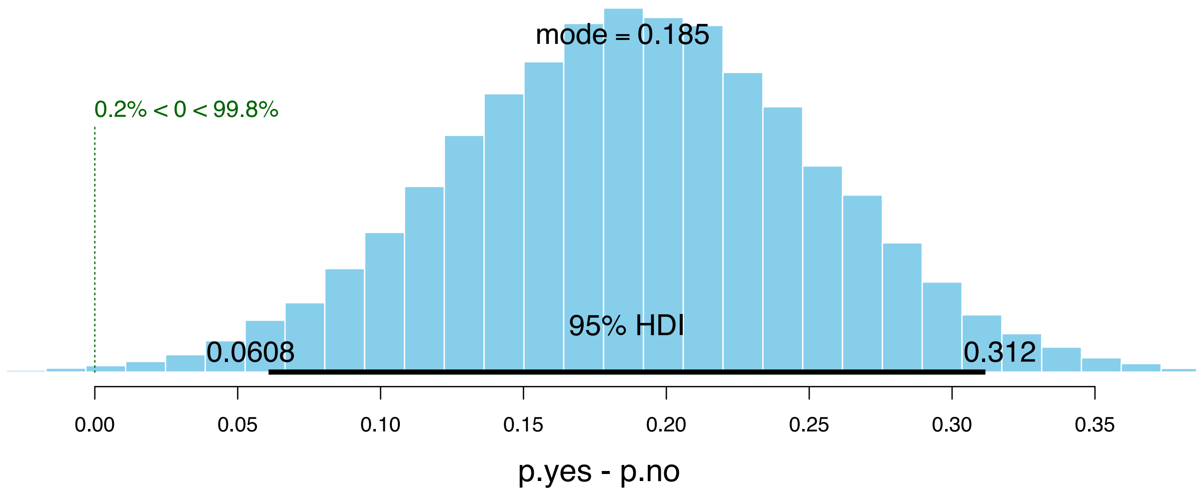
```
[1] 2.978572
```

Envoyer un rappel augmente proportionnellement les chances de présence (i.e., augmente la cote) par environ 2.98.

Travaux pratiques

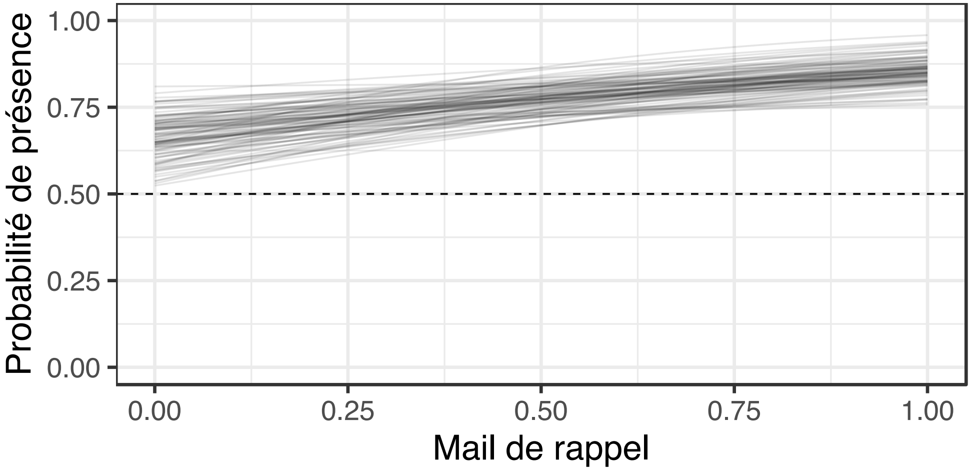
Quel est l'effet **absolu** du mail de rappel ?

```
post <- posterior_samples(mod8) # extracting posterior samples
p.no <- plogis(post$b_Intercept) # mean probability of presence when no reminder
p.yes <- plogis(post$b_Intercept + post$b_reminder) # mean probability of presence when reminder
plotPost(p.yes - p.no, compVal = 0, showMode = TRUE) # plotting it
```



Travaux pratiques

```
library(tidybayes)
library(modelr)

df3 %>%
  group_by(total) %>%
  data_grid(reminder = seq_range(reminder, n = 1e2) ) %>%
  add_fitted_draws(mod8, newdata = ., n = 100, scale = "linear") %>%
  mutate(estimate = plogis(.value) ) %>%
  group_by(reminder, .draw) %>%
  summarise(estimate = mean(estimate) ) %>%
  ggplot(aes(x = reminder, y = estimate, group = .draw) ) +
  geom_hline(yintercept = 0.5, lty = 2) +
  geom_line(aes(y = estimate, group = .draw), size = 0.5, alpha = 0.1) +
  ylim(0, 1) + theme_bw(base_size = 20) +
  labs(x = "Mail de rappel", y = "Probabilité de présence")
```



Travaux pratiques

1. Quelle est la probabilité qu'un participant, qui s'est inscrit de son propre chef, vienne effectivement passer l'expérience ?
2. Quel est l'effet du rappel ?
3. **Quel est l'effet du mode d'inscription ?**
4. Quel est l'effet conjoint de ces deux prédicteurs ?

Travaux pratiques

Écrire le modèle qui prédit la présence en fonction du mode d'inscription.

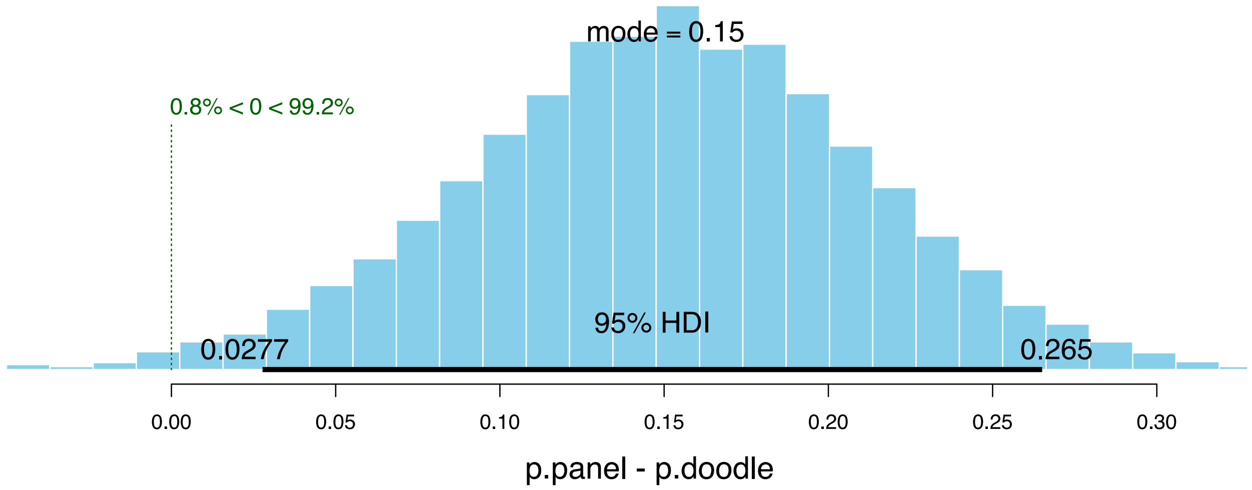
$$
\begin{aligned}
y_i &\sim \mathrm{Binomial}(n_i, p_i) \\
\mathrm{logit}(p_i) &= \alpha + \beta \times \mathrm{inscription}_i \\
\alpha &\sim \mathrm{Normal}(0, 1) \\
\beta &\sim \mathrm{Normal}(0, 1)
\end{aligned}
$$

Travaux pratiques

```r
priors <- c(
  set_prior("normal(0, 1)", class = "Intercept"),
  set_prior("normal(0, 1)", class = "b")
  )

mod9 <- brm(
    presence | trials(total) ~ 1 + inscription,
    family = binomial(link = "logit"),
    prior = priors,
    data = df3,
    cores = parallel::detectCores()
    )
```

Travaux pratiques

```r
post <- posterior_samples(mod9)
p.panel <- plogis(post$b_Intercept) # mean probability of presence for panel
p.doodle <- plogis(post$b_Intercept + post$b_inscription) # mean probability of presence for
doodle
plotPost(p.panel - p.doodle, compVal = 0, showMode = TRUE) # plotting it
```



La probabilité de présence est augmentée de 0.15 lorsque l'on s'inscrit sur un panel comparativement à une inscription sur un doodle (effet légèrement plus faible que pour le rappel).

Travaux pratiques

1. Quelle est la probabilité qu'un participant, qui s'est inscrit de son propre chef, vienne effectivement passer l'expérience ?
2. Quel est l'effet du rappel ?
3. Quel est l'effet du mode d'inscription ?
4. **Quel est l'effet conjoint de ces deux prédicteurs ?**

Travaux pratiques

Écrire le modèle complet.

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n_i, p_i) \\
\text{logit}(p_i) &= \alpha + \beta_1 \times \text{reminder}_i + \beta_2 \times \text{inscription}_i \\
\alpha &\sim \text{Normal}(0, 1) \\
\beta_1, \beta_2 &\sim \text{Normal}(0, 1)
\end{aligned}
$$

Travaux pratiques

```r
priors <- c(
  set_prior("normal(0, 1)", class = "Intercept"),
  set_prior("normal(0, 1)", class = "b")
  )

mod10 <- brm(
    presence | trials(total) ~ 1 + reminder + inscription,
    family = binomial(link = "logit"),
    prior = priors,
    data = df3,
    cores = parallel::detectCores()
    )
```

Travaux pratiques

```r
summary(mod10)
```

```
 Family: binomial
  Links: mu = logit
Formula: presence | trials(total) ~ 1 + reminder * inscription
   Data: df3 (Number of observations: 13)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
Intercept                1.12      0.67    -0.16     2.44 1.00     1796
reminder                 0.77      0.63    -0.46     2.00 1.00     1740
inscription             -0.46      0.65    -1.78     0.81 1.00     1753
reminder:inscription     0.26      0.68    -1.09     1.58 1.00     2020
                     Tail_ESS
Intercept                2494
reminder                 2284
inscription              2417
reminder:inscription     2752

Samples were drawn 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).
```

Travaux pratiques

Le mode d'inscription et le mail de rappel semblent avoir moins d'effet dans le modèle complet que dans les modèles simples… pourquoi ?

```
fixef(mod8) %>% exp
```

```
          Estimate Est.Error     Q2.5    Q97.5
Intercept 1.964382  1.284145 1.221371 3.238707
reminder  2.979146  1.479422 1.384286 6.437879
```

```
fixef(mod9) %>% exp
```

```
             Estimate Est.Error      Q2.5      Q97.5
Intercept   6.1683665  1.463513 3.1212567 13.8555004
inscription 0.3883346  1.518847 0.1615534  0.8462066
```

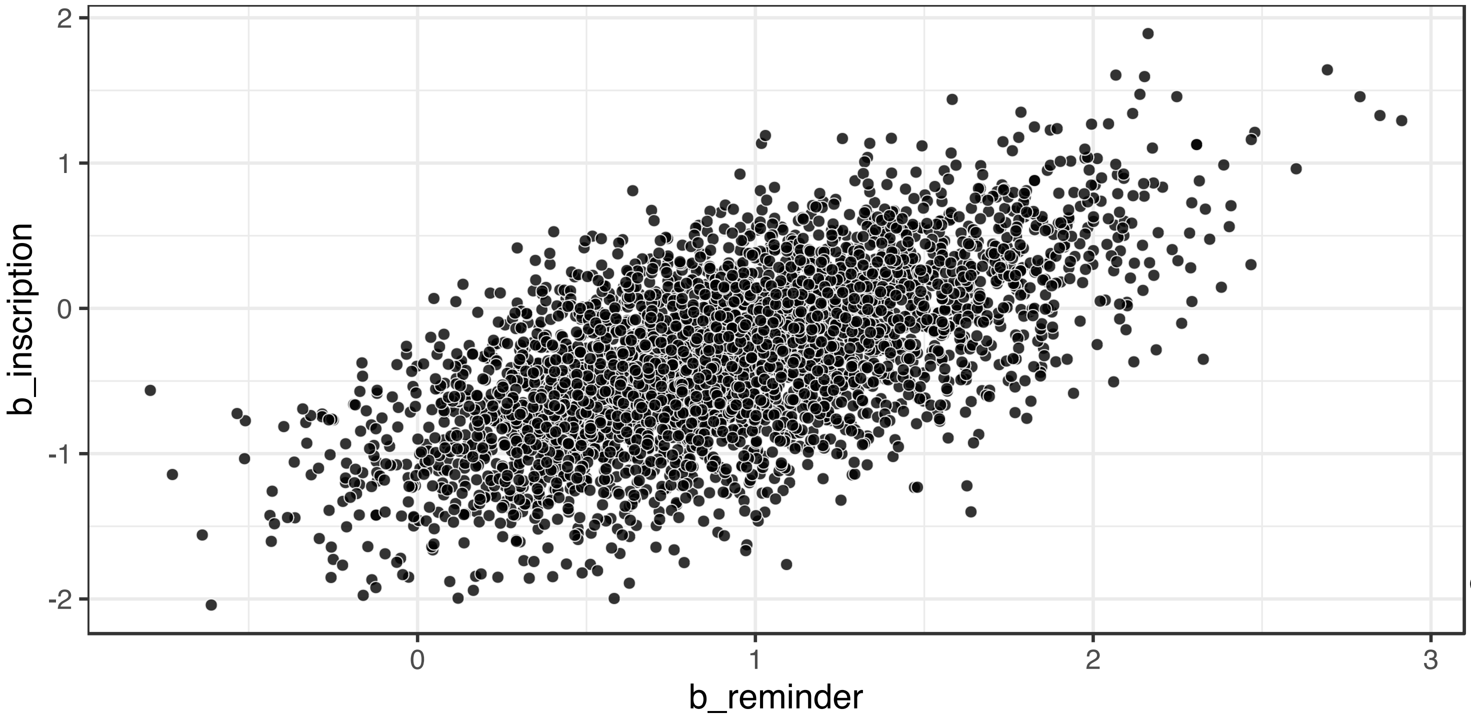
```
fixef(mod10) %>% exp
```

```
             Estimate Est.Error      Q2.5    Q97.5
Intercept   2.7173514  1.782696 0.8727010 8.390388
reminder    2.5107471  1.647695 0.9880202 6.980495
inscription 0.7153916  1.735721 0.2418714 2.135521
```

Travaux pratiques

On a déjà rencontré ce cas de figure (cf. Cours n°04). Lorsque deux prédicteurs contiennent une part d'information commune, l'estimation des pentes est corrélée...

```
posterior_samples(mod10) %>%
    ggplot(aes(b_reminder, b_inscription) ) +
    geom_point(size = 3, pch = 21, alpha = 0.8, color = "white", fill = "black") +
    theme_bw(base_size = 20)
```



Travaux pratiques

En effet, les données ont été collectées par deux expérimentateurs. L'un d'entre eux a recruté tous ses participants via doodle, et n'envoyait pas souvent de mail de rappel. Le deuxième expérimentateur a recruté tous ses participants via un panneau physique présent dans le laboratoire et envoyait systématiquement un mail de rappel. Autrement dit, ces deux variables sont presque parfaitement confondues.

```
read.csv("data/absence.csv") %>%
  sample_frac %>%
  group_by(inscription, reminder) %>%
  summarise(n = sum(total) ) %>%
  spread(key = reminder, value = n) %>%
  data.frame
```

```
  inscription no yes
1      doodle 72  22
2       panel NA  51
```

Travaux pratiques

En effet, les données ont été collectées par deux expérimentateurs. L'un d'entre eux a recruté tous ses participants via doodle, et n'envoyait pas souvent de mail de rappel. Le deuxième expérimentateur a recruté tous ses participants via un panneau physique présent dans le laboratoire et envoyait systématiquement un mail de rappel. Autrement dit, ces deux variables sont presque parfaitement confondues.

```
read.csv("data/absence.csv") %>%
  sample_frac %>%
  group_by(inscription, reminder) %>%
  summarise(n = sum(total) ) %>%
  spread(key = reminder, value = n) %>%
  data.frame
```

```
  inscription no yes
1      doodle 72  22
2       panel NA  51
```