Modèles linéaires généralisés à effets mixtes 2 - Solutions

Données

Le jeu de données aiv_ducks.csv contient une partie des données de l'étude de Papp et al. (2017) sur la présence de la grippe aviaire (AIV pour *avian influenza virus*) dans les populations de différentes espèces de canards dans l'est du Canada.

> Papp, Z., Clark, R.G., Parmley, E.J., Leighton, F.A., Waldner, C., Soos, C. (2017) The ecology of avian influenza viruses in wild dabbling ducks (Anas spp.) in Canada. PLoS ONE 12: e0176297. https://doi.org/10.1371/journal.pone.0176297.

```
aiv <- read.csv("../donnees/aiv_ducks.csv")
str(aiv)
```

```
## 'data.frame':    8967 obs. of  10 variables:
##  $ Species           : chr  "MALL" "MALL" "MALL" "MALL" ...
##  $ Age               : chr  "HY" "HY" "HY" "AHY" ...
##  $ Sex               : chr  "M" "F" "F" "M" ...
##  $ AIV               : int  1 0 1 1 1 0 1 1 1 0 ...
##  $ Site              : chr  "Amherst Point" "White Birch" "White Birch" "Tower Goose" ...
##  $ Latitude          : num  45.8 46 46 46 46 ...
##  $ Longitude         : num  -64.2 -64.3 -64.3 -64.3 -64.3 ...
##  $ Year              : int  2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...
##  $ Temperature       : num  18.6 17.6 17.6 17.6 17.6 ...
##  $ Population_Density: num  1.2 1.16 1.16 1.16 1.16 ...
```

Voici la description des champs de données:

- *Species*: Code d'espèce (ABDU = canard noir, AGWT = sarcelle à ailes vertes, AMWI = canard d'Amérique, BWTE = sarcelle à ailes bleues, MALL = canard mallard, MBDH = hybride canard noir / mallard, NOPI = canard pilet)
- *Age*: Âge (HY = année d'éclosion, AHY = après l'année d'éclosion)
- *Sex*: Sexe (F/M)
- *AIV*: Présence (1) ou absence (0) du virus de la grippe aviaire
- *Site*: Site d'échantillonnage
- *Latitude* et *Longitude*: Coordonnées géographiques du site
- *Year*: Année d'échantillonnage
- *Temperature*: Température moyenne dans les 2 semaines précédant l'échantillonnage
- *Population_Density*: Densité de population de canards (toutes espèces confondues) estimée pour le site et l'année

1. Ajustement du modèle

a) Estimez les paramètres d'un modèle complet visant à prédire la présence/absence de l'AIV, incluant: les effets fixes de l'âge et du sexe des canards, de la température et de la densité de population du

site; ainsi que les effets aléatoires de l'espèce, du site, de l'année et de l'interaction site x année (cette dernière est notée `(1 | Site:Year)` dans le modèle). Faut-il vérifier s'il y a surdispersion pour ce modèle?

Solution

```
library(lme4)
mod_comp <- glmer(AIV ~ Age + Sex + Temperature + Population_Density + (1 | Species) +
                    (1|Site) + (1|Year) + (1|Site:Year), data = aiv, family = binomial)
summary(mod_comp)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: AIV ~ Age + Sex + Temperature + Population_Density + (1 | Species) +
##     (1 | Site) + (1 | Year) + (1 | Site:Year)
##    Data: aiv
##
##      AIC      BIC   logLik deviance df.resid
##   8718.5   8782.4  -4350.3   8700.5     8958
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9768 -0.5406 -0.2763  0.6780  6.1936
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  Site:Year (Intercept) 1.21845  1.1038
##  Site      (Intercept) 0.48947  0.6996
##  Year      (Intercept) 1.11851  1.0576
##  Species   (Intercept) 0.05698  0.2387
## Number of obs: 8967, groups:  Site:Year, 211; Site, 72; Year, 7; Species, 7
##
## Fixed effects:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.52680    0.63830  -0.825   0.4092
## AgeHY               0.75327    0.10774   6.991 2.72e-12 ***
## SexM                0.12222    0.05531   2.210   0.0271 *
## Temperature        -0.12411    0.02194  -5.657 1.54e-08 ***
## Population_Density  0.16045    0.21692   0.740   0.4595
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) AgeHY  SexM   Tmprtr
## AgeHY       -0.118
## SexM        -0.053  0.010
## Temperature -0.525 -0.025 -0.009
## Ppltn_Dnsty -0.430 -0.018  0.017 -0.089
```

Il ne peut pas y avoir de surdispersion pour des données binaires.

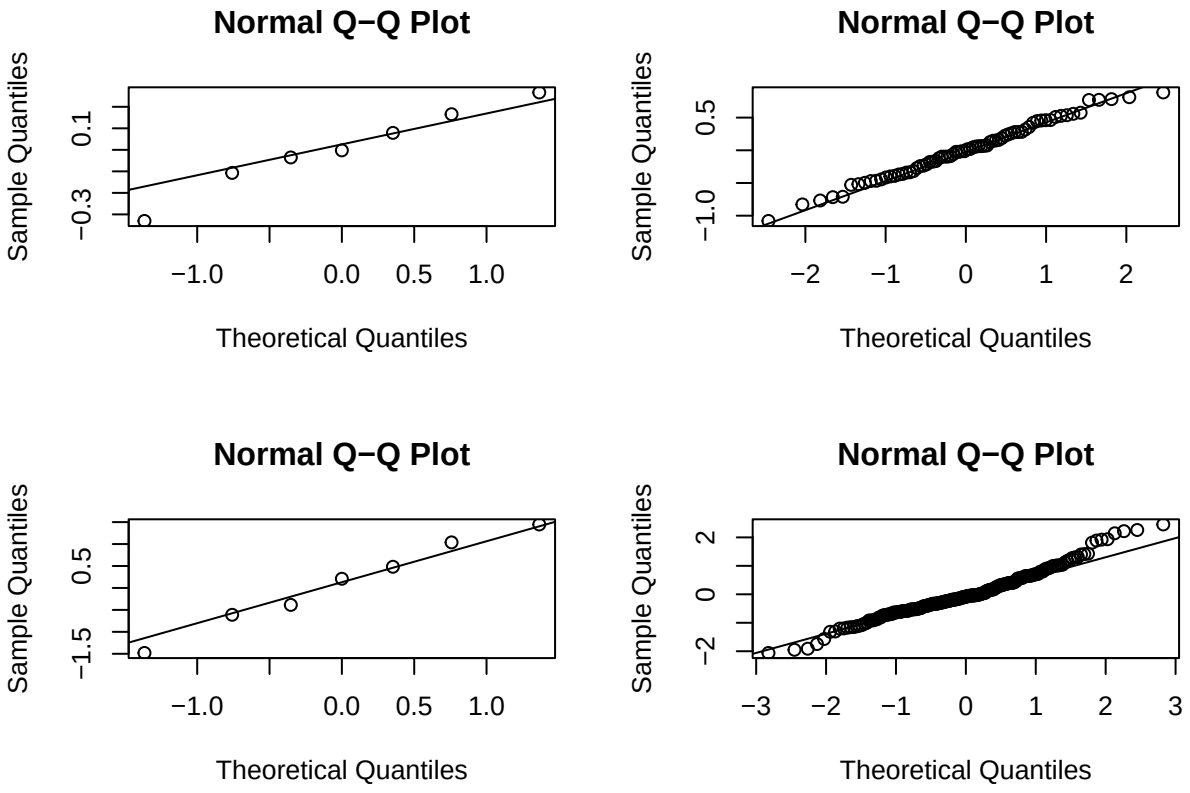
b) Quelle est la raison d'inclure chacun des effets aléatoires du modèle en (a)? Vérifiez si ces effets aléatoires suivent une distribution normale.

Solution

- Espèce: Le virus peut être plus présent chez certaines espèces, donc les observations d'une même espèce sont corrélées.
- Site: Le virus peut être plus présent à certains sites indépendamment de l'année, donc les observations d'un même site sont corrélées.
- Année: Le virus peut être plus présent globalement certaines années plutôt que d'autres, donc les observations d'une même année sont corrélées.
- Site x Année: La présence du virus est corrélée pour les canards observés sur le même site lors de la même année, davantage que pour les canards mesurés au même site au cours d'années différentes ou la même année sur des sites différents.

Selon les graphiques quantile-quantile ci-dessous, les effets aléatoires sont proches de la normalité.

```
re <- ranef(mod_comp)
par(mfrow = c(2, 2))
qqnorm(re$Species$`(Intercept)`)
qqline(re$Species$`(Intercept)`)
qqnorm(re$Site$`(Intercept)`)
qqline(re$Site$`(Intercept)`)
qqnorm(re$Year$`(Intercept)`)
qqline(re$Year$`(Intercept)`)
qqnorm(re$`Site:Year`$`(Intercept)`)
qqline(re$`Site:Year`$`(Intercept)`)
```

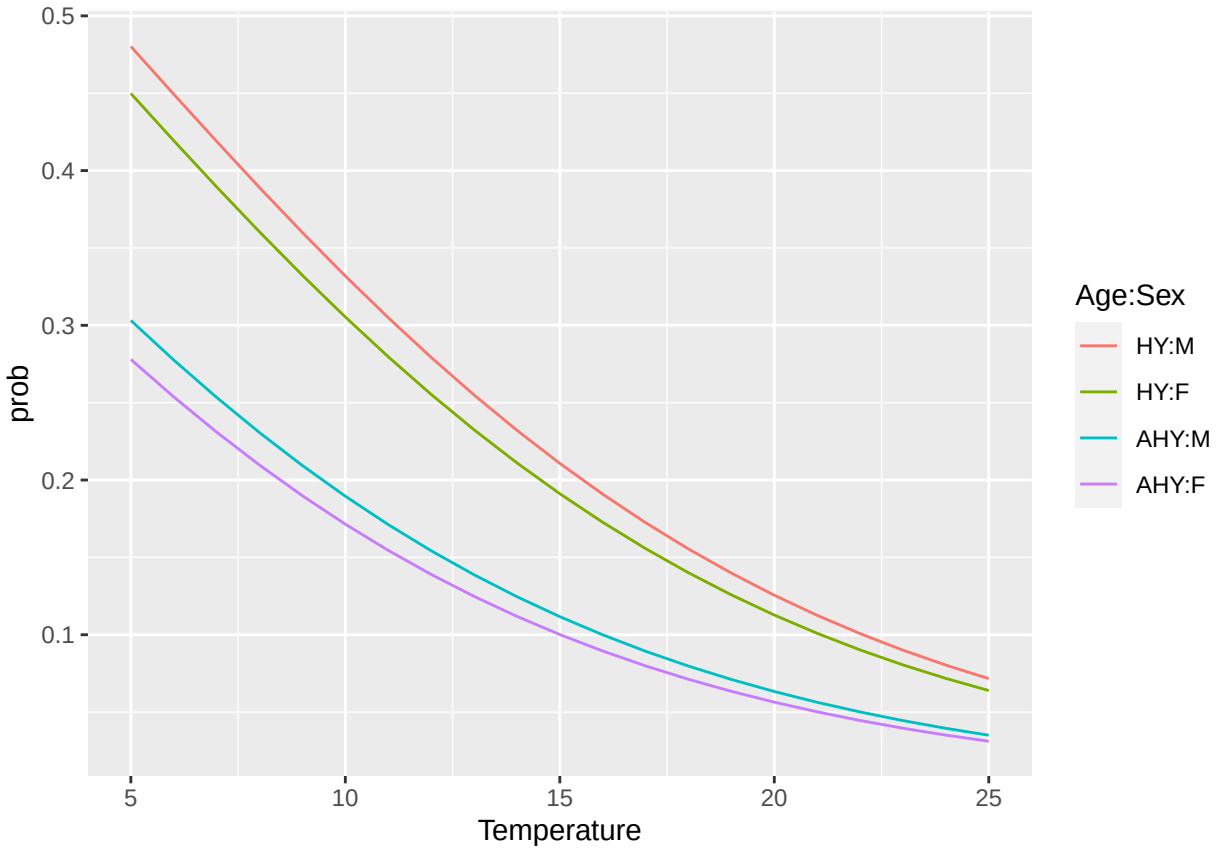


c) Produisez un graphique de la probabilité de présence de l'AIV prédite par le modèle en (a) en fonction de la température pour chacune des quatre catégories d'âge et de sexe (HY/F, HY/M, AHY/F, AHY/M). La densité de population n'apparaîtra pas dans le graphique, mais vous pouvez la fixer à sa valeur moyenne pour les prédictions.

Solution

```
pred_df <- expand.grid(Temperature = seq(5, 25, 1),
                       Age = unique(aiv$Age), Sex = unique(aiv$Sex),
                       Population_Density = mean(aiv$Population_Density))
pred_df$prob <- predict(mod_comp, newdata = pred_df, type = "response", re.form = ~0)

library(ggplot2)
ggplot(pred_df, aes(x = Temperature, y = prob, color = Age:Sex)) +
    geom_line()
```



d) À partir du modèle complet, utilisez l'AIC pour déterminer s'il est préférable d'inclure ou non chacun des effets suivants: température, densité de population, ainsi que l'effet aléatoire pour l'interaction site x année.

Solution

D'abord comparons avec ou sans l'interaction site x année.

```
library(AICcmodavg)
mod_sans_inter <- glmer(AIV ~ Age + Sex + Temperature + Population_Density + (1 | Species) +
                        (1|Site) + (1|Year), data = aiv, family = binomial)
aictab(list(mod_comp = mod_comp, mod_sans_inter = mod_sans_inter))
```

```
##
## Model selection based on AICc:
##
##                K    AICc Delta_AICc AICcWt Cum.Wt       LL
## mod_comp       9 8718.52       0.00      1      1 -4350.25
## mod_sans_inter 8 8926.12     207.59      0      1 -4455.05
```

Ensuite, comparons les effets fixes avec ou sans température, avec ou sans densité de population.

```
mod_temp <- glmer(AIV ~ Age + Sex + Temperature + (1 | Species) +
                        (1|Site) + (1|Year) + (1|Site:Year), data = aiv, family = binomial)
mod_pop_dens <- glmer(AIV ~ Age + Sex + Population_Density + (1 | Species) +
                            (1|Site) + (1|Year) + (1|Site:Year), data = aiv, family = binomial)
mod_aucun <- glmer(AIV ~ Age + Sex + (1 |Species) + (1|Site) + (1|Year) +
                        (1|Site:Year), data = aiv, family = binomial)
aictab(list(mod_comp = mod_comp, mod_temp = mod_temp,
            mod_pop_dens = mod_pop_dens, mod_aucun = mod_aucun))
```

```
##
## Model selection based on AICc:
##
##              K    AICc Delta_AICc AICcWt Cum.Wt       LL
## mod_temp     8 8717.05       0.00   0.68   0.68 -4350.52
## mod_comp     9 8718.52       1.48   0.32   1.00 -4350.25
## mod_aucun    7 8746.52      29.47   0.00   1.00 -4366.25
## mod_pop_dens 8 8748.47      31.43   0.00   1.00 -4366.23
```

Le modèle avec température, mais sans densité de population a l'AICc le plus faible.

e) Les auteurs de l'étude originale ont déterminé un effet significatif de la densité de population en ajustant un modèle avec effets aléatoires du site et de l'année, mais sans leur interaction. Pour quelle raison les conclusions de votre modèle pourraient-elles différer de ce résultat?

Solution

Il y a une mesure de densité de population par site et par année, donc si l'effet aléatoire site x année est important, ce qui signifie que les mesures prises sur le même site la même année ne sont pas indépendantes, cela diminue la significativité de l'effet de la densité de population. Autrement dit, cet effet est confondu avec d'autres facteurs qui changent entre les sites d'une année à l'autre.

2. Prédictions du modèle

a) Ajoutez au jeu de données original des colonnes représentant la prédiction de la probabilité de présence de l'AIV (1) en fonction seulement des effets fixes du modèle; (2) en fonction des effets fixes et aléatoires. Utilisez le meilleur modèle identifié dans la partie précédente.

Solution

```
aiv$pred_fix <- predict(mod_temp, re.form = ~0, type = "response")
aiv$pred_alea <- predict(mod_temp, type = "response")
```

b) Pour chaque type de prédictions obtenues (effets fixes; effets fixes et aléatoires), déterminez la probabilité moyenne prédite de présence de l'AIV pour les observations avec AIV = 1, ainsi que la probabilité moyenne prédite de présence pour les observations avec AIV = 0. D'après vos résultats, les effets fixes du modèle permettent-ils de bien départager les cas de présence et d'absence? Qu'en est-il des effets aléatoires?

Solution

```
library(dplyr)
group_by(aiv, AIV) %>%
    summarize(mean(pred_fix), mean(pred_alea))
```

```
## # A tibble: 2 x 3
##     AIV `mean(pred_fix)` `mean(pred_alea)`
##   <int>            <dbl>             <dbl>
## 1     0            0.166             0.218
## 2     1            0.173             0.486
```

Avec les effets fixes, la probabilité d'infection prédite est à peine plus grande pour les individus infectés vs. non-infectés (17.3% vs. 16.6%). Avec les effets aléatoires, la probabilité d'infection prédite pour les individus infectés est environ le double de celle des non-infectés (48.6% vs 21.8%). Donc la variation interannuelle et spatiale explique davantage les cas de présence et d'absence que les effets fixes.

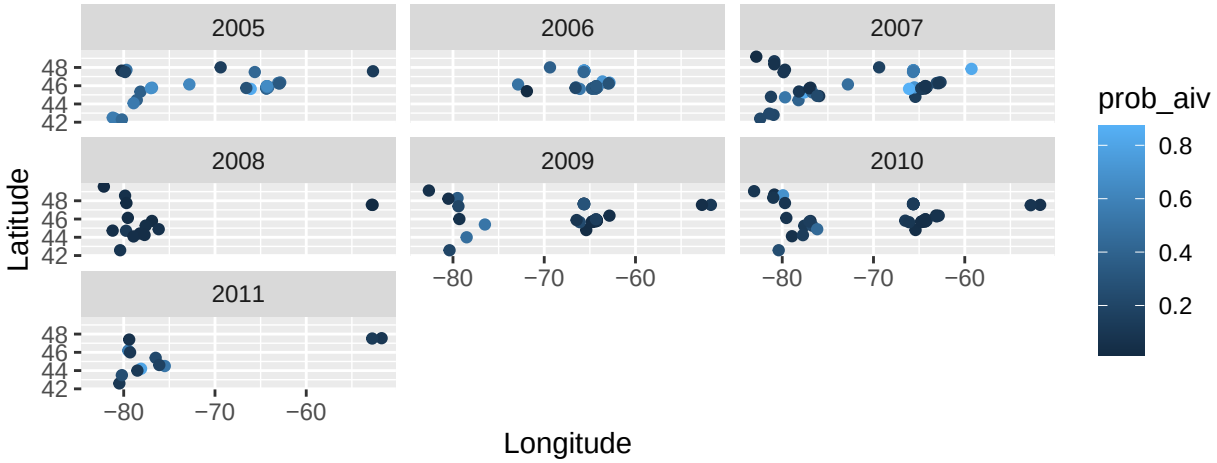
c) Groupez le jeu de données par site et année, puis calculez la moyenne de la longitude, de la latitude et de la probabilité d'AIV prédite par le modèle complet pour chaque combinaison site-année. En utilisant ces variables, produisez une carte des sites avec leur probabilité d'AIV prédite pour chaque année. (Vous pouvez utiliser les facettes dans *ggplot2* pour séparer le graphique en panneaux pour chaque année.)

Solution

```
aiv$pred_comp <- predict(mod_comp, type = "response")
aiv_sites <- group_by(aiv, Site, Year) %>%
    summarize(Latitude = mean(Latitude), Longitude = mean(Longitude),
              prob_aiv = mean(pred_comp))
```

```
## `summarise()` has grouped output by 'Site'. You can override using the
## `.groups` argument.
```

```
ggplot(aiv_sites, aes(x = Longitude, y = Latitude, color = prob_aiv)) +
    geom_point() +
    facet_wrap(~ Year) +
    coord_fixed() # optionnel, assure qu'un degré en x et en y ait la même longueur
```


2005
2006
2007
2008
2009
2010
2011
Latitude
48
46
44
42
−80
−70
−60
Longitude
prob_aiv
0.8
0.6
0.4
0.2