Generalized linear mixed models 2 - Solutions

Data

The aiv_ducks.csv dataset contains some of the data from the study by Papp et al. (2017) on the occurrence of avian influenza (AIV) in populations of different species of ducks in eastern 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 ...
```

Here is the description of the data fields:

- *Species*: Species code (ABDU = black duck, AGWT = green-winged teal, AMWI = American wigeon, BWTE = blue-winged teal, MALL = mallard, MBDH = black duck / mallard hybrid, NOPI = northern pintail)
- *Age*: Age (HY = hatching year, AHY = after hatching year)
- *Sex*: Sex (F/M)
- *AIV*: Presence (1) or absence (0) of avian influenza virus
- *Site*: Sampling site
- *Latitude* and *Longitude*: Geographical coordinates of the site
- *Year*: Year of sampling
- *Temperature*: Mean temperature in the 2 weeks prior to sampling
- *Population_Density*: Estimated duck population density (all species) for the site and year.

1. Fitting the model

a) Estimate the parameters of a comprehensive model to predict the presence/absence of AIV, including: the fixed effects of duck age and sex, temperature, and site population density; and the random effects of species, site, year, and site x year interaction (the latter is denoted as `(1 | Site:Year)` in the model). Should we check for overdispersion in this model?

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
```

There can be no overdispersion for binary data.

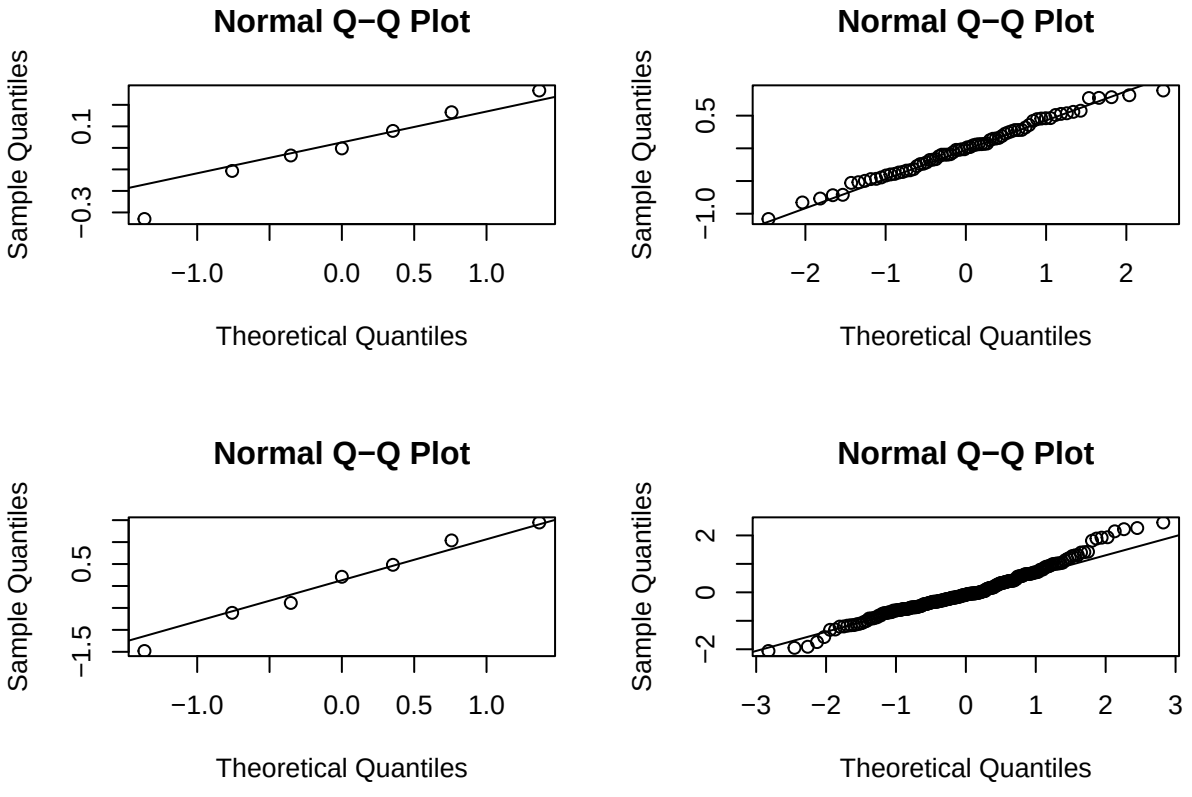
b) What is the reason for including each of the random effects of the model in (a)? Check whether these random effects follow a normal distribution.

Solution

- Species: The virus may be more present in some species, so observations of the same species are correlated.
- Site: The virus may be more present at some sites regardless of the year, so observations from the same site are correlated.
- Year: The virus may be more present globally in some years than others, so observations from the same year are correlated.
- Site x Year: The presence of the virus is correlated for ducks observed at the same site in the same year, more so than for ducks measured at the same site in different years or in the same year at different sites.

According to the quantile-quantile plots below, the random effects are close to normal.

```
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)`)
```

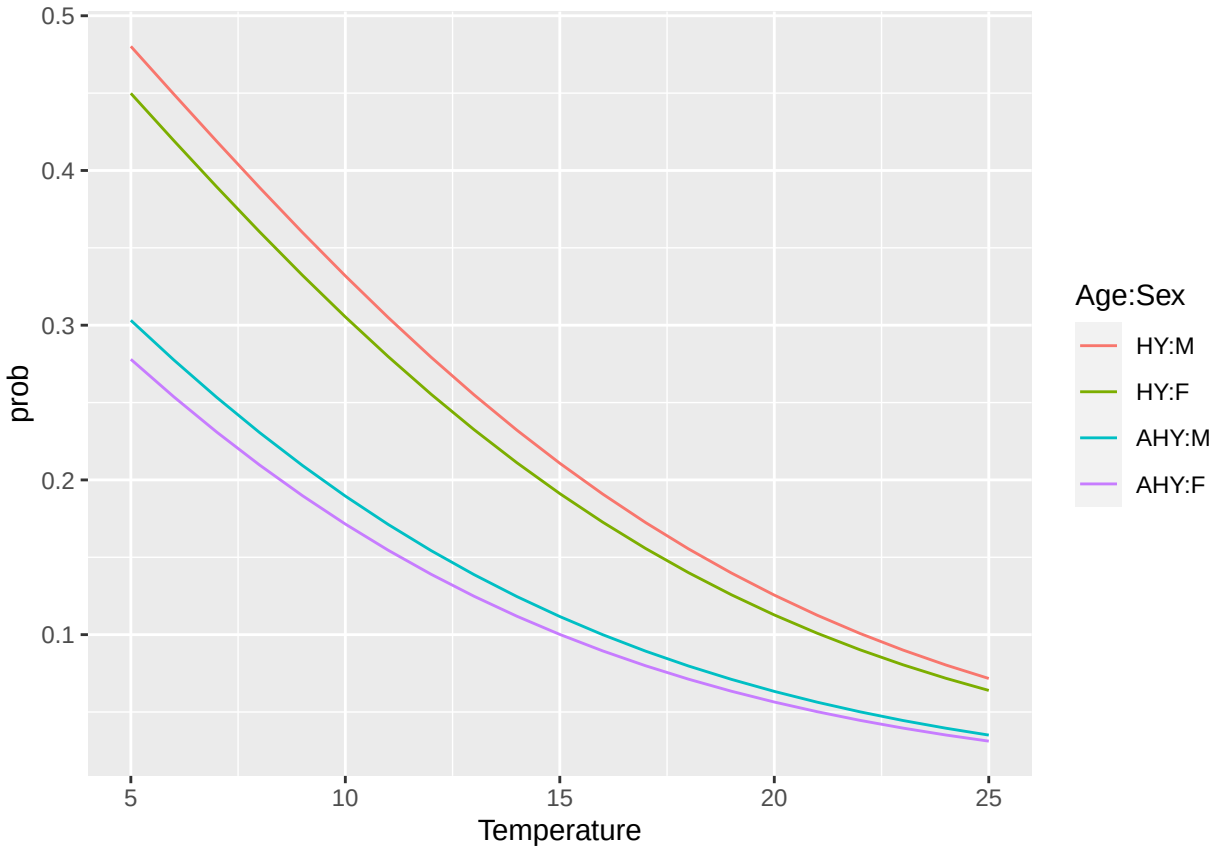


c) Produce a graph of the model-predicted probability of occurrence of AIV in (a) as a function of temperature for each of the four age and sex categories (HY/F, HY/M, AHY/F, AHY/M). The population density will not appear in the graph, but you can set it to its mean value for the predictions.

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) Starting from the full model, use the AIC to determine whether or not to include each of the following effects: temperature, population density, and the random effect for site x year interaction.

Solution

First let's compare with or without the site x year interaction.

```
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
```

Then let's compare the fixed effects with or without temperature, and with or without population density.

```
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
```

The model with temperature but without population density has the lowest AICc.

e) The authors of the original study determined a significant effect of population density by fitting a model with random site and year effects, but without their interaction. Why might the conclusions of your model differ from this result?

Solution

There is one measure of population density per site per year, so if the random effect of site x year is large, meaning that measurements taken at the same site in the same year are not independent, this decreases the significance of the population density effect. In other words, this effect is confounded by other factors that change between sites from one year to the next.

2. Model predictions

a) Add columns to the original dataset representing the prediction of the probability of occurrence of AIV (1) based only on the fixed effects of the model; (2) based on both fixed and random effects. Use the best model identified in the previous section.

Solution

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

b) For each type of prediction obtained (fixed effects; fixed and random effects), determine the predicted mean probability of occurrence of AIV for observations with AIV = 1 and the predicted mean probability of occurrence for observations with AIV = 0. Based on your results, do the model's fixed effects provide a good distinction between presence and absence? What about random effects?

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
```

With fixed effects, the probability of predicted infection is only slightly higher for infected vs. uninfected individuals (17.3% vs. 16.6%). With random effects, the predicted probability of infection for infected individuals is about double that of uninfected individuals (48.6% vs. 21.8%). Thus, inter-annual and spatial variation explains more the presence and absence cases than the fixed effects.

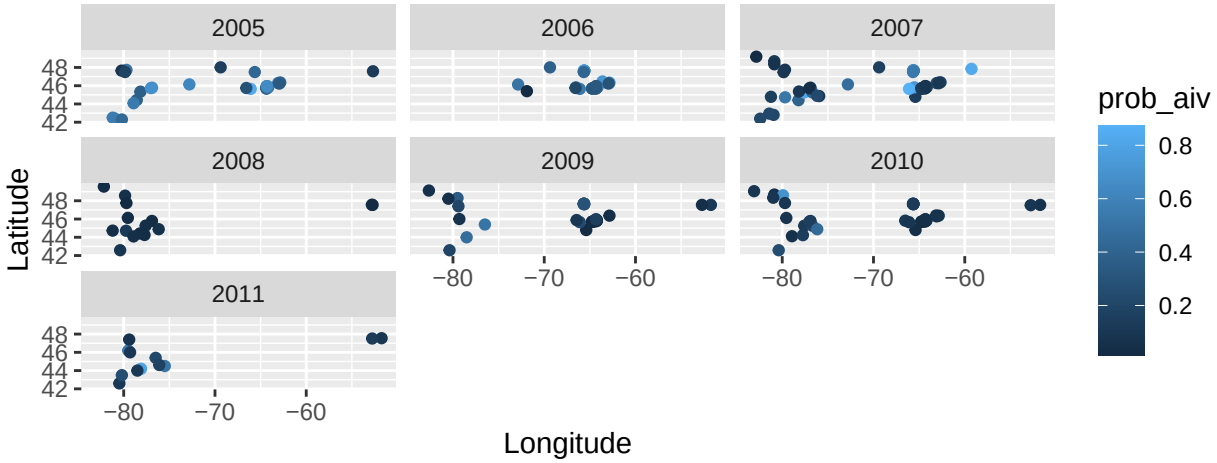
c) Group the dataset by site and year, then calculate the mean longitude, latitude, and probability of AIV predicted by the full model for each site-year combination. Using these variables, produce a map of the sites with their predicted probability of AIV for each year. (You can use the facets in *ggplot2* to separate the graph into panels for each year).

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() # optional, makes sure x and y axes are on same scale
```


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