Generalized linear mixed-effects models

Nora Wickelmaier

Last modified: May 20, 2025

GLMMs

- Like linear models, mixed-effects models can be extended so that they allow for response variables that have arbitrary distributions
- A GLM(M) is a specific combination of a response distribution, a link function, and a linear predictor
- We can choose (almost) all the link functions that work with `glm()` for `glmer()`

```
## Family name         Link functions
   binomial            logit, probit, log, cloglog
   gaussian            identity, log, inverse
   Gamma               identity, inverse, log
   inverse.gaussian    1/mu^2, identity, inverse, log
   poisson             log, identity, sqrt
```

Example: 1989 Bangladesh Fertility Survey[1]

Subset containing variables about

- District where women live
- Age (mean centered)
- Living children (1 = no children, 2 = one child, 3 = two children, 4 = three or more children)
- Use of artificial contraception (yes or no)
- Area where women live is rural or urban

[1] Reanalysis of https://repsychling.github.io/SMLP2024/glmm.html

Exercise

- Read the data set `contra.dat` into R
- Create a plot
  - Showing the probability to use artificial contraception depending on age
  - Draw separate lines for number of living children
  - Make two panels: one for rural and one for urban
  - Use `ggplot2` or `lattice`

Hint: Use `geom_smooth()` for `ggplot2` and `type = "smooth"` for `lattice::xyplot()`

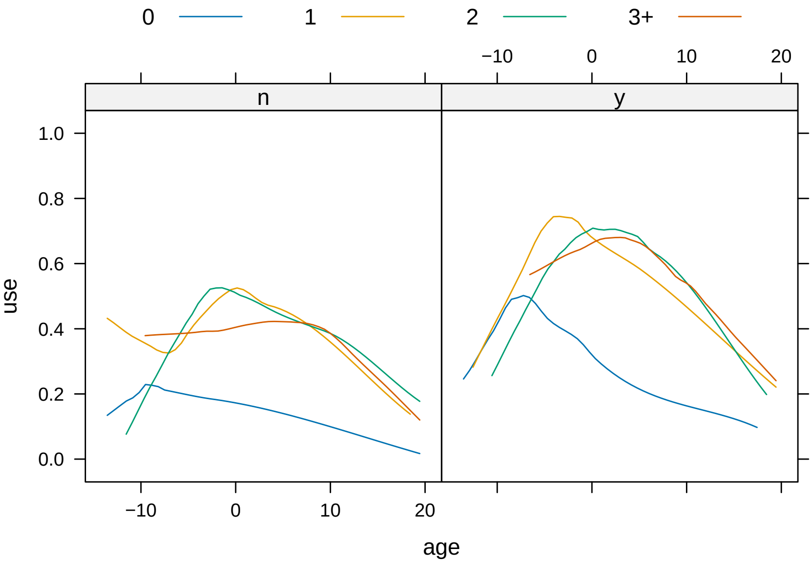
Data set contra

```
dat <- read.table("data/contra.dat", header = TRUE)

dat$districtID <- factor(dat$districtID)
dat$childCode <- factor(dat$childCode,
                        levels = 1:4,
                        labels = c("0", "1", "2", "3+"))
dat$isUrban <- factor(dat$isUrban,
                      levels = 0:1,
                      labels = c("n", "y"))

# Simplify names
names(dat) <- c("dist", "use", "livch", "age", "urban")
```

Visualization



GLMM

- We fit the following model to the data

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 urban + \beta_4 livch + \upsilon_0$$

  with $\upsilon_0 \sim N(0, \sigma^2_{\upsilon_0})$
- We are assuming that there are differences between the districts, so we add a random intercept for district
- What are your conclusions about differences between women with and without children?

Dichotomize `livch`

- We create a new factor `children` that indicates if a woman has children or not (independently of how many)
- Next, we add an interaction term for `children` and age to our model

  $$log(\frac{p}{1-p}) = \beta_0 + \beta_1 age + \beta_2 children + \beta_3 (age \times children) + \beta_4 age^2 + \beta_5 urban + \upsilon_0$$

  with $\upsilon_0 \sim N(0, \sigma^2_{\upsilon_0})$
- How can we compare which model fits the data better?

Nested random effect

- It turns out that districts are very big and can incorporate rural as well as urban areas
- The districts can be very different depending on this
- Add a random intercept to the model taking this into account

```
xtabs( ~ urban + dist, dat)
xtabs( ~ urban + factor(urban:dist), dat)
xtabs( ~ dist + factor(dist:urban), dat) |> print(zero = ".")
```

- Compare the models – what is your conclusion?
- What exactly is the difference between these models?

Model predictions

```
# Compare models
data.frame(models = c("gm1", "gm2", "gm3"),
           df = AIC(gm1, gm2, gm3)[, 1],
           AIC = AIC(gm1, gm2, gm3)[, 2],
           BIC = BIC(gm1, gm2, gm3)[, 2],
           deviance = c(deviance(gm1), deviance(gm2), deviance(gm3)))

# Model predictions
newdat <- data.frame(children = factor(rep(c("true", "false"),
                                           each = 2)),
                     urban = factor(rep(c("y", "n"), times = 2)),
                     age = 0)
newdat$pre <- predict(gm3, type = "response", newdata = newdat,
                      re.form = NA)
```

Exercise

- Create a new data frame with variables `children` ("true", "false"), `urban` ("yes", "no"), and `age` (ranging from -14 to 20)
- Add a prediction for each combination of the three variables
- Draw a plot of your predictions
- Interpret the results

Summary

- From the data plot we can see a quadratic trend in the probability by age.
- The patterns for women with children are similar and we do not need to distinguish between 1, 2, and 3+ children.
- We do distinguish between those women who do not have children and those with children. This shows up in a significant *age* $\times$ *children* interaction term.