Software for Analyzing Data

Tutorial – Day 10

January 6, 2022

John-Oliver Engler
Rieke Baier & Chân Lê
Quantitative Methods of Sustainability Science Group

Today

- Generalized linear models (GLMs): Conclusions
- GLM excercise: Titanic survival probabilities
- Linear mixed models (LMMs) and generalized linear mixed models (GLMMs)
- R demonstration

`glm()` in practice

- GLMs are not estimated via OLS, but via maximum likelihood.
- What you already know: `mymodel <- lm(y~x)`
- GLMs can be coded in a similar way, e.g.:

```
mymodel2 <- glm(y ~ x, poisson)
mymodel3 <- glm(y ~ x, Gamma(link = "log"))
Mymodel4 <- glm(y ~ x, family = "binomial")
```

→ You don‘t usually need to tamper with the link function (The Gamma distribution being an exception).

→ Built-in canonical links are usually a good starting point.

`glm()` in practice: convergence issues

- When fitting GLMs, a frequent issue is non-convergence of models due to flatness of the likelihood surface in maximum-likelihood estimation.
- Possible reasons:
  - Mental model of the process is wrong / uninformative explanatory variables
  - Process is inherently non-linear
  - Insufficient data available to fit the model
  - Wrong error distribution or/and link function
- Parameter values from non-converged models might be nonsense

Model diagnostics for GLMs

You should ask the same questions as for simple linear modeling, in addition you need to check the following:

1. `vif()`
   - measures multicollinearity of predictors
   - rule of thumb: model ok, if `vif(predictor)` < 4 for all predictors
2. `c_hat()`:
   - measures overdispersion for binomial or Poisson models
   - rule of thumb: model ok if `c_hat(predictor)` ≈ 1 for all predictors
3. `RsqGLM(), NagelkerkeR2(), r2` and others
   - returns pseudo-R^2 values for GLMs

Excercise: Titanic survival (15‘)

Let us practice using glms in the following.

a.) Load and inspect the dataset „titanic.txt“ that you will find on mystudy.

b.) Given this dataset, what models would you find interesting and why?

c.) Set up a model for `Survival` using `Pclass`, `Sex` and `Age` as predictor variables. What is a sensible random component / error structure for this model? How would you check model fit?

d.) Interpret the output using the `predict` function. According to this model, what is the survival probability of a 40-year old female travelling in First Class? How about a 40-year old male from Third Class?

(Generalized) Linear mixed models

LMMs | GLMMs

Why mixed models?

Some violations of the assumptions of the linear model can be dealt with using GLMs instead of LMs, e.g.:

- The expected value of errors is always zero.
- Residual variance is constant.
- Residuals follow a normal distribution.

However, a (G)LM cannot deal with violations of the assumption that

- Data are random sample from the population, i.e. the errors are statistically independent of one another

➢ Can be tackled with (generalized) linear mixed models, (G)LMMs!

Why mixed models? (2)

- Independence assumption often violated in practice
- Potential source: repeated measurements taken from
  - the same individual (temporal pseudoreplication)
  - the same place (spatial pseudoreplication)
- Peculiarities from one individual or place will be common to all samples from that individual or place

Pseudoreplication: Examples

- Bacterial infection of patients allocated at random to different treatments (placebo, drug, drug+supplement): 11-week trial, patients (ID) assessed at different times

- Animal numbers in different districts in southern Mongolia, recorded for each year from 1981–2015

city	camel	horse	cow	sheep	goat	total	total1	total2	ndvi	ndvi1	ndvi2	presum	presum1	presum2	year	wool_price	wool_price1	wool_price2
Bayandalai	4564	5501	3393	18952	50410	82820	78316	82324	0.0870			111.70	107.00		1981	737.60	721.02	
Bayan-Ovoo	12757	6681	1344	14449	18095	53326	51250	53917	0.0864			351.40	49.70		1981	737.60	721.02	
Bulgan	5531	6550	1745	22801	26385	63012	59449	68131	0.0914			126.50	151.80		1981	737.60	721.02	
Dalanzadgad	8208	11976	3287	47964	47295	118730	109657	118518	0.0891			121.90	148.20		1981	737.60	721.02	
Gurvantes	11222	1947.00	611	10396	63743	87919	86169	85178	0.0677			119.80	104.10		1981	737.60	721.02	
Khanbogd	20716	3002	2383	23594	7994	57689	61388	64751	0.1002			129.80	71.30		1981	737.60	721.02	
Khurmen	11425	5342	1104	21224	27778	66873	64697		0.0779			96.20	114.30		1981	737.60	721.02	
Mandal-Ovoo	16476	5011	1886	18623	17275	59271	57996	70643	0.0728			30.70	76.60		1981	737.60	721.02	
Manlai	9674	8125	970	24177	22629	65575	61969	81361	0.0864						1981	737.60	721.02	
Nomgon	13346	8481	535	17560	51406	91328	90937	90747	0.0779			16.80	9.10		1981	737.60	721.02	
Noyon	9082	1007	492	9713	25296	45590	42152	42457	0.0758			152.40	61.90		1981	737.60	721.02	
Sevrey	7355	3314	1444	18147	49378	79638	75866	82750	0.0806			96.20			1981	737.60	721.02	
Tsogt-Ovoo	8988	6357	540	34267	17464	67616	65866	64751	0.0801			64.90	56.20		1981	737.60	721.02	
Tsogttsetsii	4079	7610	1090	21575	21752	56106	53928	62614	0.0873			82.30	25.70		1981	737.60	721.02	
Bayandalai	4677	5511	3563	19809	52306	85866	82820	78316	0.0822	0.0870		30.80	111.70	107.00	1982	689.51	737.60	721.02
Bayan-Ovoo	12763	6693	1419	14947	19199	55021	53326	51250	0.0801	0.0864		4.20	351.40	49.70	1982	689.51	737.60	721.02
Bulgan	5612	6739	1856	23813	25911	63931	63012	59449	0.0933	0.0914		86.50	126.50	151.80	1982	689.51	737.60	721.02
Dalanzadgad	11544	12112	3555	47319	47710	122240	118730	109657	0.0722	0.0891		76.60	121.90	148.20	1982	689.51	737.60	721.02
Gurvantes	11292	2144.00	632	9975	65996	90039	87919	86169	0.0723	0.0677		56.20	119.80	104.10	1982	689.51	737.60	721.02
Khanbogd	20247	2993	2571	24227	8836	58874	57689	61388	0.0952	0.1002		38.10	129.80	71.30	1982	689.51	737.60	721.02
Khurmen	8274	5746	1224	21673	28376	65293	66873	64697	0.0767	0.0779		47.30	96.20	114.30	1982	689.51	737.60	721.02
Mandal-Ovoo	16333	4819	1824	19520	16879	59375	59271	57996	0.0860	0.0728		2.00	30.70	76.60	1982	689.51	737.60	721.02
Manlai	9708	7597	1023	22427	23321	64076	65575	61969	0.0882	0.0864					1982	689.51	737.60	721.02

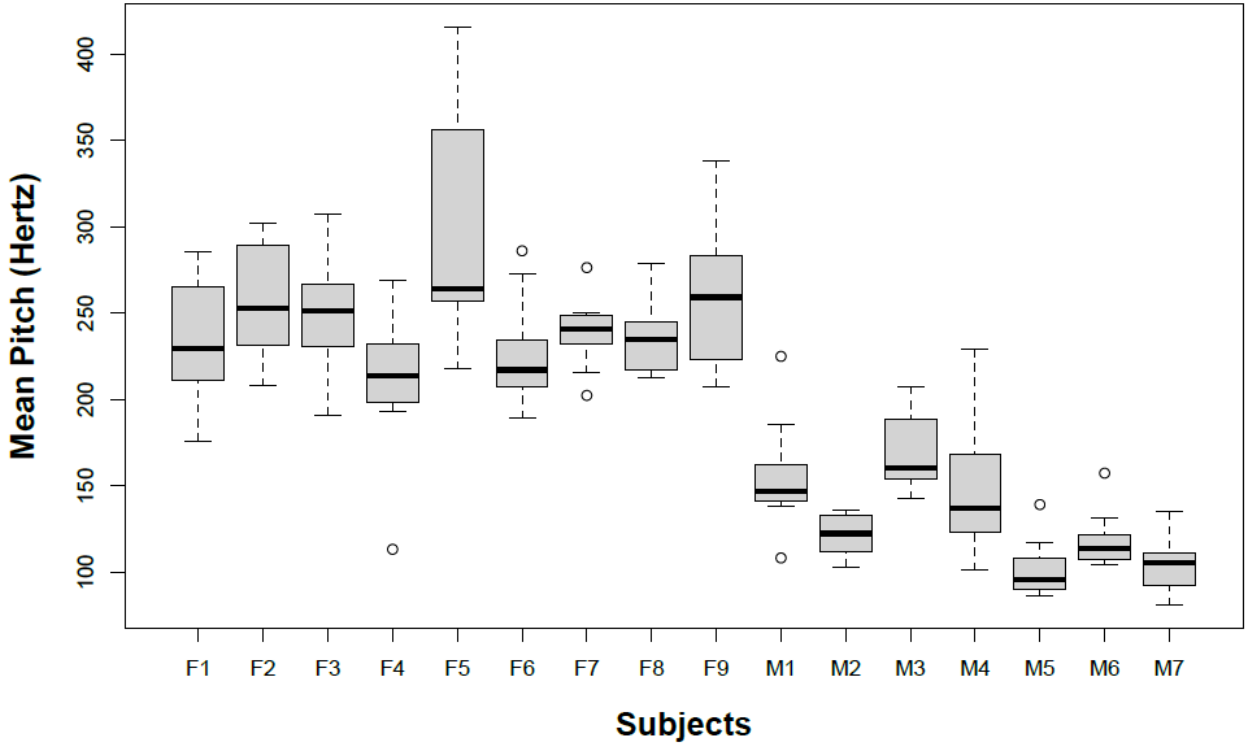
What are mixed models?

- Repeated measurements on the same subject will feature the particularities of that subject

- Such particularities generally
  - have an unpredictable, non-systematic or ‚random' influence on data
  - cannot be controled by the experimenter/scientist

➢ Categorical explanatory variables with these characteristics are called random effects.

By-subject variation: Example



Source: Winter (2013)

From linear model to linear mixed model

- Linear regression model:

$$y = a + bx + \epsilon$$

➢ Only the fixed effect (of *x* on *y*) included

- Linear mixed model with random intercept:

$$y = a + bx + (1|\text{subject}) + \epsilon$$

➢ Includes by-subject variation as a random factor

➢ Model assumes different random intercepts for each subject

From linear model to linear mixed model (2)

- Fixed effects influence only the mean of *y*
- Random effects influence only the variance of *y*

- The mixed model divides the world into
  - things that are systematic (or that we somehow understand)
  - things that we cannot control (or that we don‘t understand)
- distinction useful for more accurate modeling of systematic patterns in data

- Fixed-effect estimates will be more accurate as compared to the simple linear model approach

Random intercept

- Random intercept model: assumes each subject has different intercept (but same slope)
- In R: `y ~ x + z + (1|subject)`

```
$subject
   (Intercept) attitudepol   genderM
F1    243.3684   -19.72207 -108.5173
F2    266.9443   -19.72207 -108.5173
F3    260.2276   -19.72207 -108.5173
M3    284.3536   -19.72207 -108.5173
M4    262.0575   -19.72207 -108.5173
M7    224.1292   -19.72207 -108.5173
```

Random intercept and random slope

- Random intercept and slope model: additionally assumes by-subject variation in reaction to treatment
- In R: `y ~ x + z + (1 + x|subject)`

```
$subject
   (Intercept) attitudepol   genderM
F1    243.8053   -20.68245 -110.8021
F2    266.7321   -19.17028 -110.8021
F3    260.1484   -19.60452 -110.8021
M3    285.6958   -17.91950 -110.8021
M4    264.1982   -19.33741 -110.8021
M7    227.3551   -21.76744 -110.8021
```

Useful R packages, functions and reading

- Linear mixed models:
  - `lme4::lmer()`

- Generalized linear mixed models:
  - `lme4::glmer()`

Bates, D., M. Mächler, B.M. Bolker and S.C. Walker (2015), Fitting linear mixed-effects models using lme4, *Journal of Statistical Software* 67(1), 1–48

Ben Bolker's GLMM FAQ (2021): https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html

Bolker, B.M. et al. (2008), GLMMs: a practical guide for ecology and evolution, Trends in Ecology and Evolution 24(3), 127–135

Winter, B. (2013), Linear models and linear mixed-effects models in R with linguistic applications, arXiv:1308.5499

http://mfviz.com/hierarchical-models/