Generalized linear and additive models

E2M2
2024

Emily Ruhs
Postdoctoral Scientist
ecruhs@uchiago.edu

Please ask questions!

Veuillez poser des questions - en anglais, français ou malgasy !

Thumbs up

Thumbs down

Outline for today

1. *Linear models (LM) brief review*
    1. *Check-in*
2. *Introduction to generalized linear mixed models (GLMM)*
3. *Random effects*
4. *Why and when might you use a GLMM?*
    1. *Check-in*
5. *Introducing generalized additive models (GAM)*
6. *Why and when might you use a GAM?*
7. *Introducing the dataset*
8. *Break – 5 minutes*

Progression of model selection

Simple statistics	*Data normality; T-tests, ANOVA*
1. Univariate Linear models	*One predictor variable*
↓	
2. Multivariate Linear models	*More than one predictor variable*
↓	
3. Generalized Linear models	*Utilizes link function; specify distribution*
↓	
4. Generalized Linear mixed models	*Includes random and/or nested effects*

Linear models

- Linear models describe a continuous response variable as a function of one or more predictor variables. They can help you understand and predict the behavior of complex systems or analyze experimental, financial, and biological data.
- E.g., describes the relationship between two or more variables

- There are several types of linear regression:
- **Simple linear regression:** models using only one predictor
- **Multiple linear regression:** models using multiple predictors
- **Multivariate linear regression:** models for multiple response variables

Linear models

$$y = X\beta + e$$

- y = vector of observed dependent values
- X = Design matrix: observations of the variables in the assumed linear model
- β = *vector of unknown parameters to estimate*
- *e = vector of residuals (deviation from model fit),*
  $e = y - X\beta$

Linear models

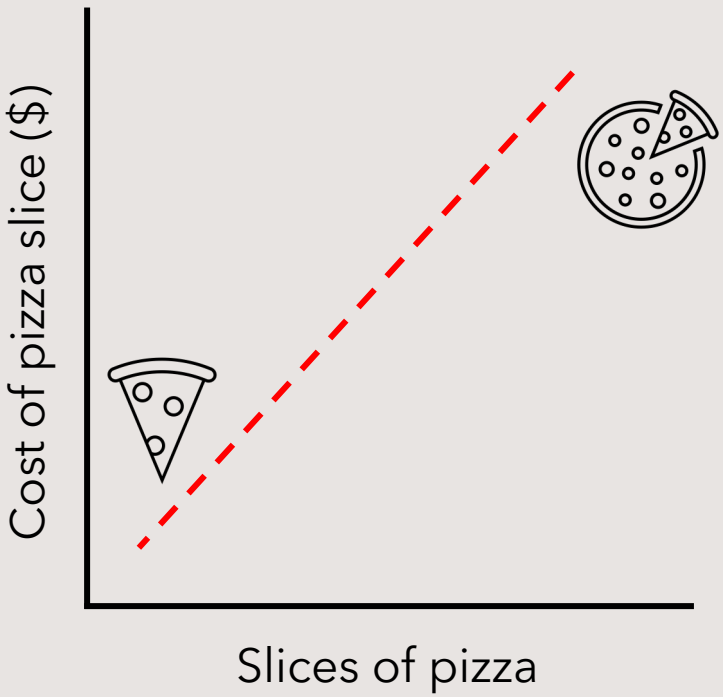
Fixed vs. Random effects

- Goals: Estimate the values of model parameters and estimate any appropriate variances

- Example: in a regression model, $y=\alpha+\beta x+e$, we estimate the values for α and β and the variance of e.
- Note: α/β are fixed constants, we are trying to estimate the fixed factors or fixed effects, while e_i is drawn from a probability distribution (e.g. usually a normal distribution). E_i are random effects.
- In a simple linear model, no additional random effects (e.g. variable from your dataset) are included.

A simple example

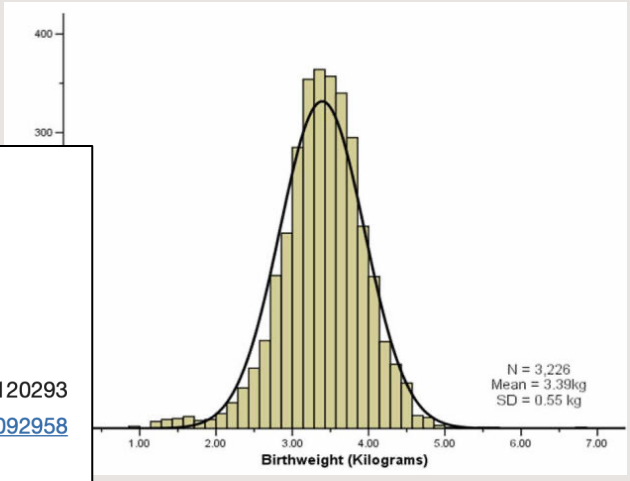
- A linear model example is a verbal scenario that can be modeled using a linear equation or vice versa. An example could be each pizza slice costs $2 and the delivery fee is $4, so the linear model would be y=2x+4, where y represents the total cost and x represents the number of pizza slices.



LM vs. GLM

Can't we just transform our data?

GLM are extensions of linear models (LM), but we can use them for data that is not normally distributed.



Journal List > Shanghai Arch Psychiatry > v.26(2); 2014 Apr > PMC4120293

SHANGHAI ARCHIVES OF PSYCHIATRY
About us | Editorial board | Free e-subscription | Info for authors

Shanghai Arch Psychiatry. 2014 Apr; 26(2): 105–109.
doi: 10.3969/j.issn.1002-0829.2014.02.009

PMCID: PMC4120293
PMID: 25092958

Language: English | Chinese

Log-transformation and its implications for data analysis

Changyong FENG,[1,*,*] Hongyue WANG,[1] Naiji LU,[1] Tian CHEN,[1] Hua HE,[1] Ying LU,[2] and Xin M. TU[1]

Author information ▸ Copyright and License information Disclaimer

Shanghai Arch Psychiatry

Revisiting fixed effects

GLM are extensions of linear models (LM)!

$Y = \alpha + \beta_i x + \varepsilon_{ij}$

Model 1 = lm(response ~ predictor, data=data)

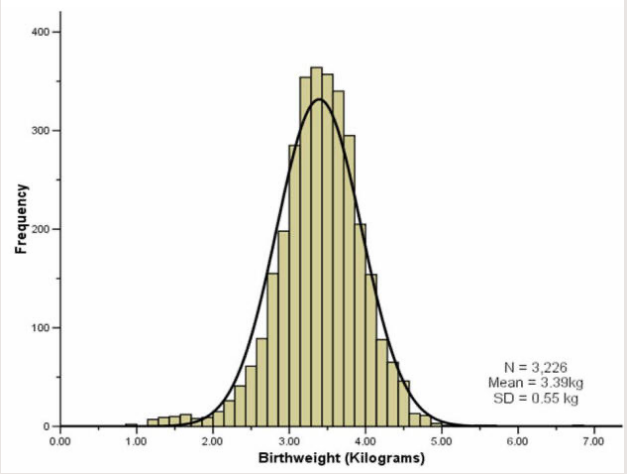
$Y_{ij} = \alpha + \beta_i x + \varepsilon_{ij}$

Model 2 = glm(response ~ predictor, family=family, data=data)

*Model 1 = lm(weight ~ height, data=data) *assumes normal*

Model 2 = glm(weight ~ height, data=data)

Model 1 results = Model 2 results



Review of GLM

How to include parameters in a model?

size ~ age + sex + id = random

According to your research questions and hypotheses, does a categorical variable represent a ***driver of primary interest****, or a* ***cohort*** *of collected data* ***representing a broader population or distribution****?*

If a ***driver,*** *consider including it in your model as a* ***fixed effect.***

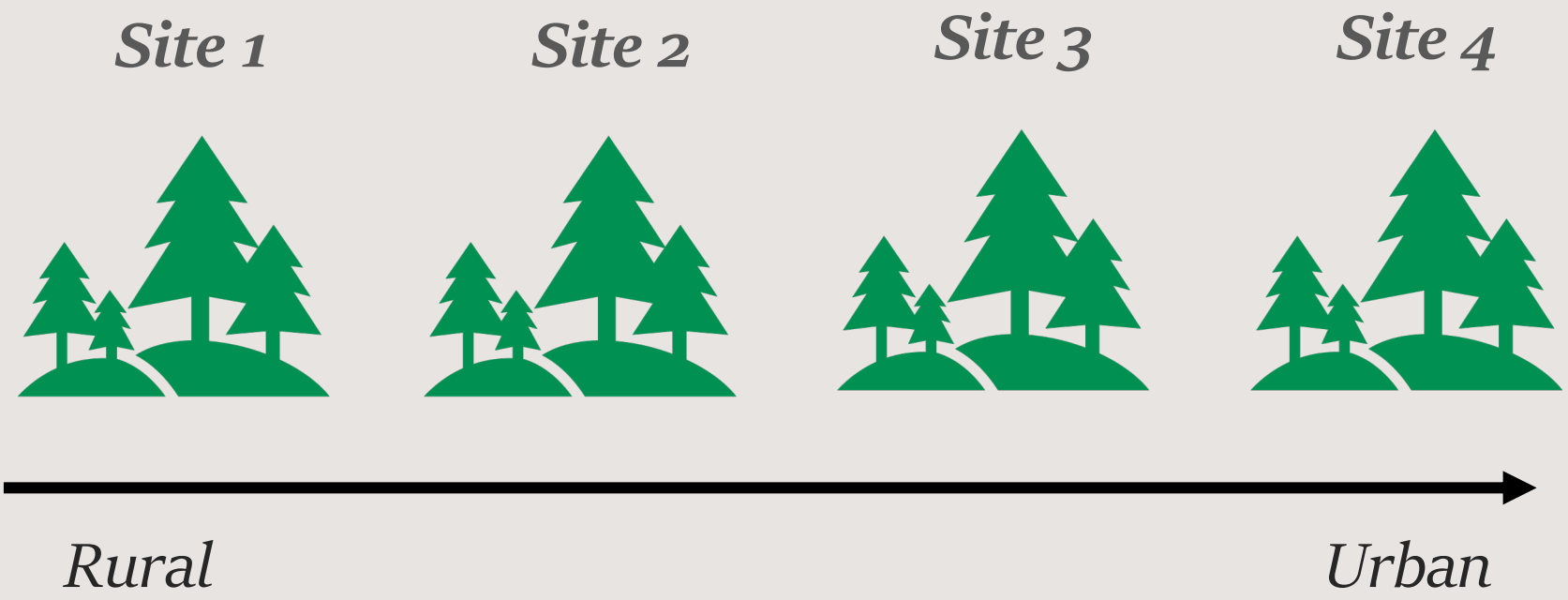
If a ***cohort drawn from a population,*** *or a* ***repeated measure*** *of the same state, consider including it in your model as a* ***random effect.***

Review of GLM

Some variables could be best represented as fixed or random, depending on your questions:

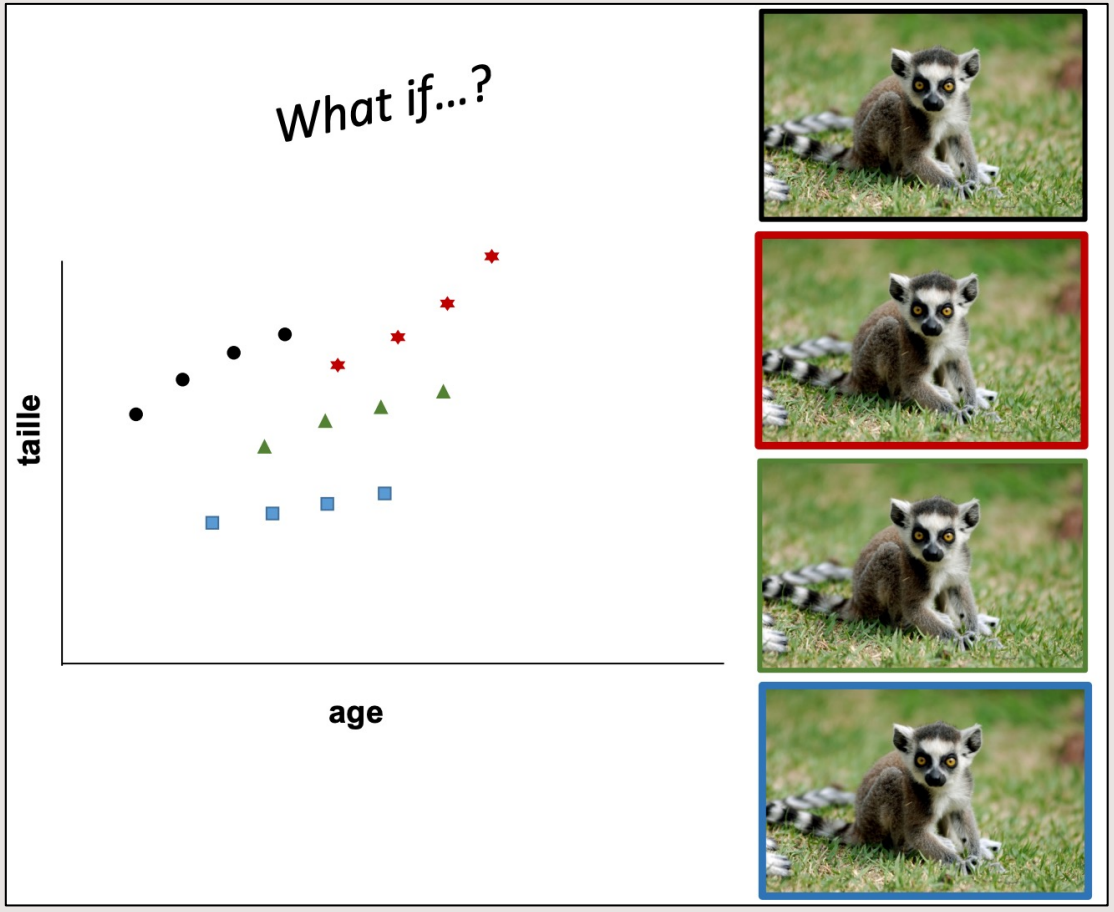
Site: repeated measures of forest habitat

Size ~ age + sex + site
Size ~ age + sex + (1|site)



Review of GLM

What if you have multiple individuals that are sampled across different ages?



Check-in time!

Veuillez poser des questions - en anglais, français ou malgasy !

Introducing generalized linear mixed models (GLMMs)

What are GLMMs?

Model that allows for non-normally distributed response variables (y) and predictor variables (x) as fixed or random effects

Lemur size ~ Age + Sex + (1| Forest)

Poisson Num Factor Random

Importantly there are three parts: (1) A systematic component (e.g. fixed effect), (2) a random component, and (3) a link function

Introducing GLMM

Assumptions that must be met:

1. *Assumes the distribution of the response is a member of the exponential family*
    1. *Gaussian, binomial, Poisson, Negative binomial, gamma, inverse gaussian, beta, Weibull, Multinomial, Dirichlet*
2. *There is a linear predictor*
3. *There is a link function that related the linear predictor to the mean of the response*
    1. *Canonical link of the response distribution*

**** Check or be aware of collinearity and singularity!*

Introducing GLMM

Goals of GLMM:

1. *Estimate the values of the model parameters*
2. *Estimate any appropriate variances*

$Y_{ij} = \alpha + \beta_i x + Z_i x + e$

The β_i are fixed effects, Z_i are random effects, e is the variance.

Introducing GLMM

GLMMs can be extensions of generalized linear models (GLM)

$Y_{ij} = \alpha + \beta_i x + \varepsilon_{ij}$

Model 1 = glm(response ~ predictor, family=family, data=data)

$Y_{ij} = \alpha + \beta_i x + \ldots \beta_n x + \varepsilon_{ij}$

Model 2 = glmer(response ~predictor + (1|variable), family=family, data=data)

This is your random effect. Could include things like sex, individual ID, nest ID, geographic region, hospital

Introducing GLMM

Benefits of including random effects:

It is often useful to treat certain effects as random, as opposed to fixed

- *Suppose we have k predictor variables. If we treat these as fixed, we lose k degrees of freedom.*
- *If we assume each of the k realizations are drawn from a normal distribution with mean zero and unknown variance, only one degree of freedom is lost – for estimating the variance*

Lemur size ~ Age + Site + error

300 samples – 3df = 297 df

Lemur size ~ Age + (1| Site) + error

300 samples – 2df = 298 df

When might we use GLMM?

Generalized linear mixed models include both fixed effects and random effects to allow for:

- *Repeated measures*
- *Temporal correlation*
- *Spatial correlation*
- *Heterogeneity*
- *Nested data*

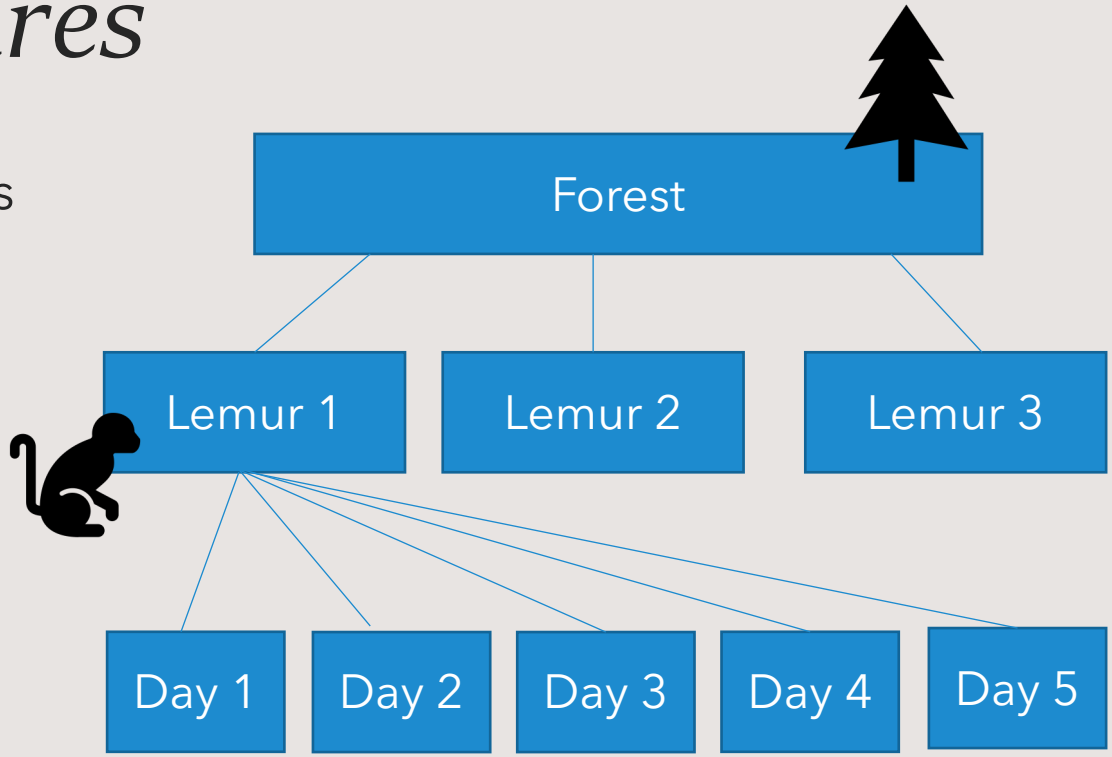
We use the model glmer() in R

Model = glmer(formula, family=family, data=data)

Example of repeated measures

Sampling the same individuals multiple times

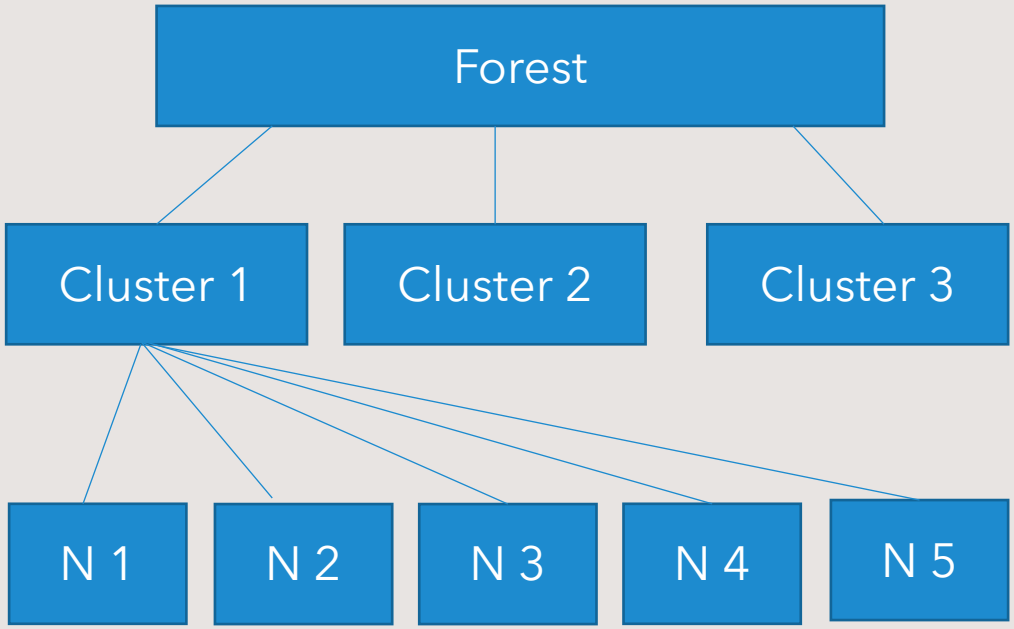
1. Tracking seasonality
2. Monitoring growth
3. Change in status (infected/not infected)
4. Longitudinal clinical data



Example of spatial correlation

Sampling closely connected areas

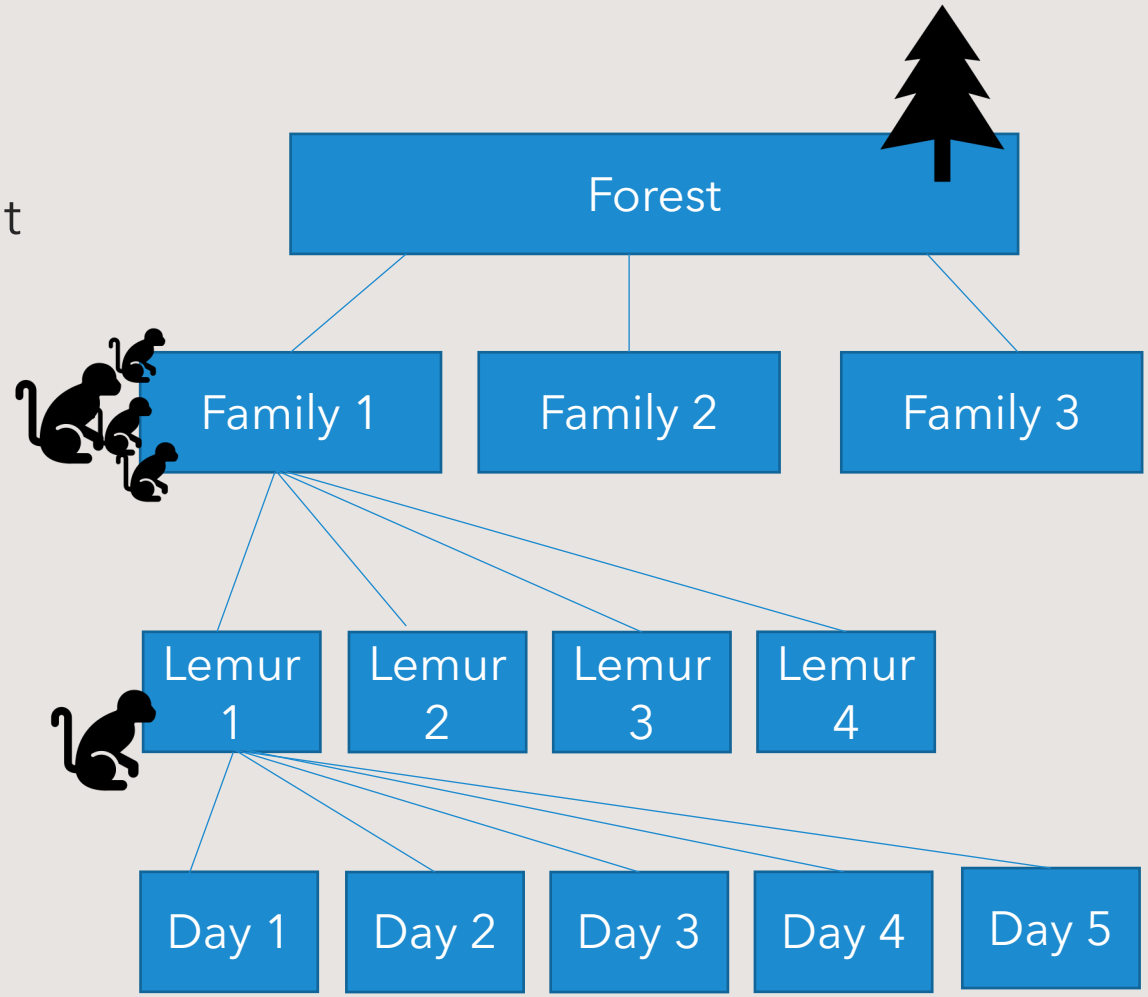
1. Geographic regions



Example of nested data

Individuals within a treatment or area

- Might be repeatedly sampled

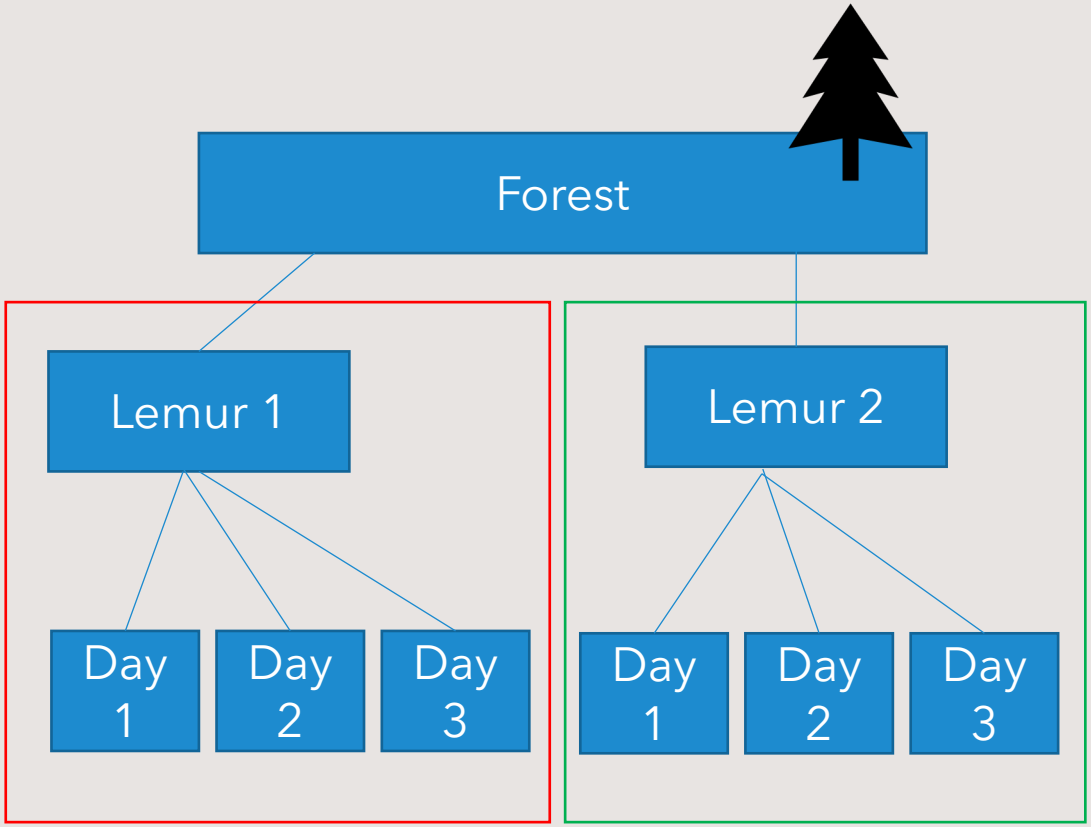


Example of random effects

Model 2 = glmer(response ~predictor + (1| random), family=family, data=data)

$Y_{ij} = \alpha + \beta_i x + Z_i x + \varepsilon_{ij}$

When we include random effects, we assume observations within the same Y_{i1} are correlated, and observations between Y's are independent (e.g. Y_{i1} and Y_{i2})



Explaining a GLMM results table

```
> summary(m6)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: Negative Binomial(1.3844)  ( log )
Formula: GIparasites ~ treatment + (1 | year)
   Data: lemur.data

     AIC      BIC   logLik deviance df.resid
  3363.3   3379.3  -1677.7   3355.3      396

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.1448 -0.6781 -0.3015  0.5365  3.3872

Random effects:
 Groups Name        Variance Std.Dev.
 year   (Intercept) 0.01263  0.1124
Number of obs: 400, groups:  year, 4

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.2623     0.0743  43.906  < 2e-16 ***
treatment1   -0.5841     0.1640  -3.562 0.000368 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr)
treatment1 -0.283
```

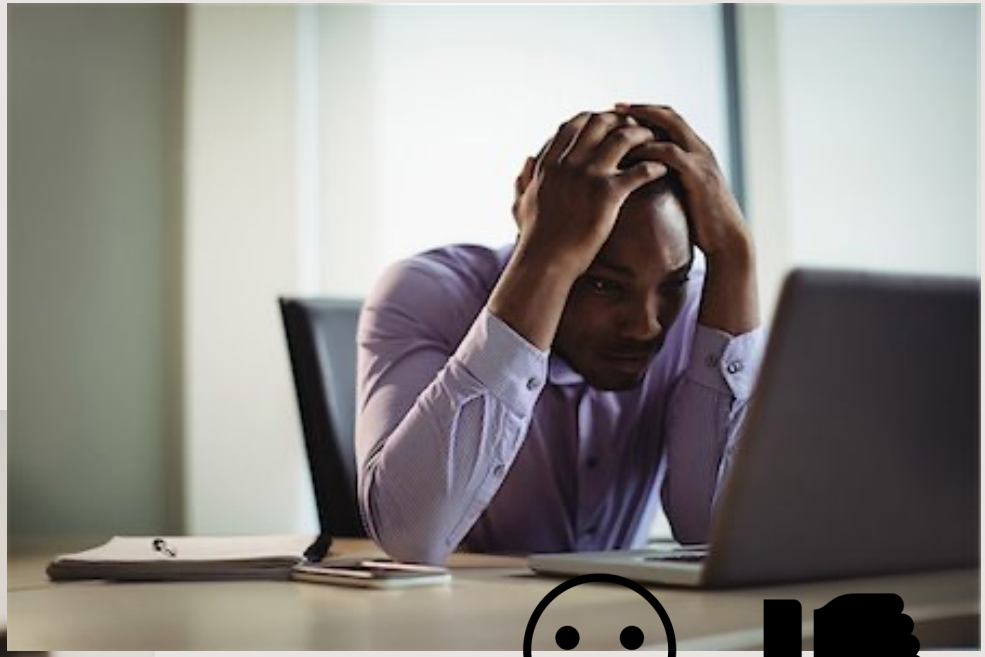
Form of your model

AIC for model selection

Random effect variance

Fixed effects

Check-in time!



Veuillez poser des questions - en anglais, français ou malgasy !

Introducing generalized additive models (GAMs)

A generalized additive model is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates.

Allows for:

- *Flexible specification of the dependent variable*

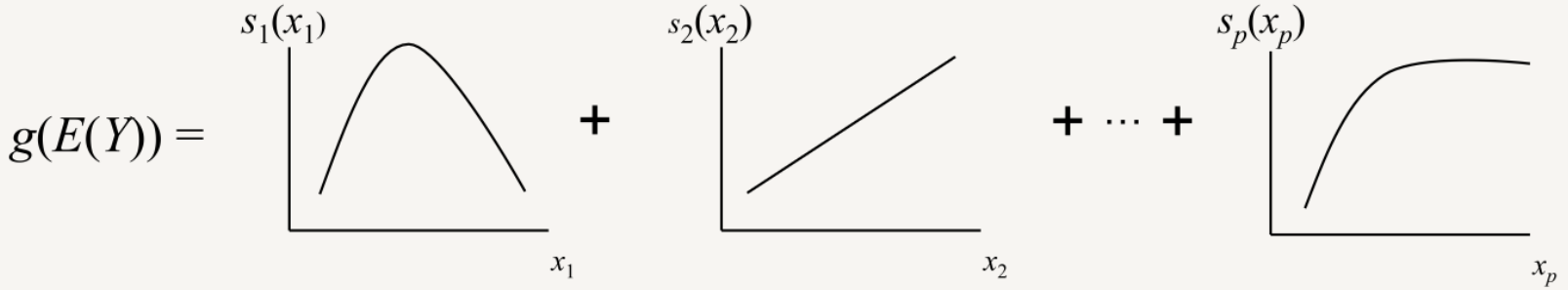
But you must:

- *Specify the "smooth function"*
- *Choose how "smooth" to make it.*
- *R can also estimate this for you….*

WHAT DOES THIS MEAN?

Introducing GAMs

An additive model where the impact of the predictive or independent variables is captured through smoothing functions:



We can write the GAM structure as:

$$g(E(Y)) = \alpha + s_1(x_1) + \cdots + s_p(x_p),$$

***Importantly – you can control how smooth the predictor functions are!*

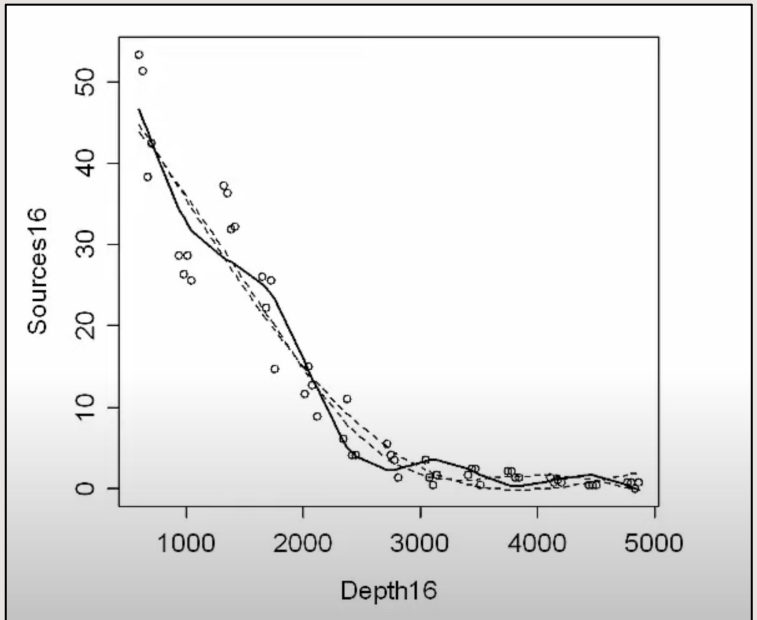
Introducing GAMs

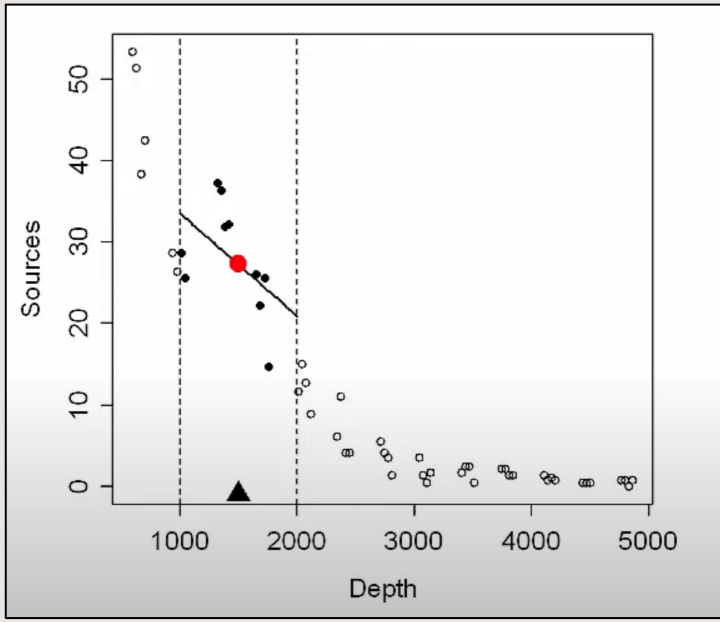
A spline curve is a is piecewise polynomial curve, i.e., it joins two or more polynomial curves. The locations of the joins are known as "knots".

Three classes of smoothers:

- *Local regression (loess)*
- *Smoothing splines*
- *Regression splines (B-splines, P-splines, thin plate splines)*

gam <- gam(response ~ predictor+ s(smoother), k=#, bs="X"), data=data, family="family")





Helpful link!

Introducing GAMs

Most common options for smoothers

bs= "?"

Thin-plate = "tp"

Cubic regression = "cr" or "cs"

P-spline = "ps" or "cp" (cyclic version)

Random effects = "re"

And more….

Broadly speaking the default penalized thin plate regression splines tend to give the best MSE performance… - Simon Wood (mgcv author)

https://stat.ethz.ch/R-manual/R-devel/library/mgcv/html/smooth.terms.html
https://www.mainard.co.uk/post/why-mgcv-is-awesome/

When might we use a GAM?

Why are GAMs useful?

- *Easy to interpret*
- Flexible predictor functions can uncover hidden patterns in the data
- Regularization of predictor functions helps avoid overfitting

When do you use GAMs?

1. There is a clear contribution of the independent variable to the prediction
2. But the relationship might not be known or not expected to be linear or non-linear

Do you love GAMs now and you want to learn more?

noamross.github.io

Generalized Additive Models in R · A Free Interactive Course

GAMs IN R by NOAM ROSS

A Free, Interactive Course using mgcv

Hello! Welcome to Generalized Additive Models in R. This short course will teach you how to use these flexible, powerful tools to model data and solve data science problems. GAMs offer offer a middle ground between simple linear models and complex machine-learning techniques, allowing you to model and understand complex systems.

To take this course, you need basic R programming skills and experience with linear regression. Now let's get started!

1 - Introduction to Generalized Additive Models

In this chapter, you will learn how Generalized additive models work and how to use flexible, nonlinear functions to model

https://noamross.github.io/gams-in-r-course/

Introducing the dataset

Builds off the dataset that Sophia used for lm lecture

100 lemurs tracked over time and space.

There are:

- *Female/male*
- *Family groupings*
- *Three locations across Madagascar*

Missing one year (2020) – COVID

Lemurs were treated in 2022 for GI parasites

Lemur dataset

AutoSave OFF — Lemurdata_glm_2

Home Insert Draw Page Layout Formulas Data Review View Tell me — Share Comments

Paste Font Alignment Number Conditional Formatting Format as Table Cell Styles Cells Editing Analyze Data Sensitivity

I13 fx 2022

	location	ID	family	age	sex	tail	Glparasites	sickGlparasit	year	treatment	predictions
2	Toliara	1	1	0	Male	27.140514	32	infected	2018	Non	37.904865
3	Toliara	1	1	1	Male	30.309658	25	infected	2019	Non	33.190076
4	Toliara	1	1	2	Male	30.309658	10	not infected	2021	Non	25.446898
5	Toliara	1	1	3	Male	30.309658	10	not infected	2022	Oui	22.28169
6	Toliara	2	1	0	Female	22.939998	26	infected	2018	Non	25.60474
7	Toliara	2	1	1	Female	25.416729	23	infected	2019	Non	22.4199
8	Toliara	2	1	2	Female	25.416729	25	infected	2021	Non	17.189382
9	Toliara	2	1	3	Female	25.416729	5	infected	2022	Oui	15.051284
10	Antsiranana	3	2	0	Male	26.044464	58	infected	2018	Non	37.904865
11	Antsiranana	3	2	1	Male	28.868004	49	infected	2019	Non	33.190076
12	Antsiranana	3	2	2	Male	28.868004	43	infected	2021	Non	25.446898
13	Antsiranana	3	2	3	Male	28.868004	20	infected	2022	Oui	22.28169
14	Antsiranana	4	2	1	Male	34.688244	42	infected	2018	Non	37.904865
15	Antsiranana	4	2	2	Male	38.071702	35	infected	2019	Non	33.190076
16	Antsiranana	4	2	3	Male	38.071702	23	infected	2021	Non	25.446898
17	Antsiranana	4	2	4	Male	38.071702	12	infected	2022	Oui	22.28169
18	Antsiranana	5	2	1	Female	16.668908	34	infected	2018	Non	25.60474
19	Antsiranana	5	2	2	Female	20.663607	31	infected	2019	Non	22.4199
20	Antsiranana	5	2	3	Female	20.663607	25	infected	2021	Non	17.189382
21	Antsiranana	5	2	4	Female	20.663607	25	infected	2022	Oui	15.051284
22	Toliara	6	3	1	Male	21.351961	63	infected	2018	Non	37.904865
23	Toliara	6	3	2	Male	26.270834	55	infected	2019	Non	33.190076
24	Toliara	6	3	3	Male	26.270834	15	infected	2021	Non	25.446898
25	Toliara	6	3	4	Male	26.270834	17	infected	2022	Oui	22.28169
26	Toliara	7	3	2	Male	27.48683	54	infected	2018	Non	37.904865
27	Toliara	7	3	3	Male	32.795534	47	infected	2019	Non	33.190076
28	Toliara	7	3	4	Male	32.795534	51	infected	2021	Non	25.446898
29	Toliara	7	3	5	Male	32.795534	20	infected	2022	Oui	22.28169
30	Antsiranana	8	5	2	Female	18.034205	16	infected	2018	Non	25.60474
31	Antsiranana	8	5	3	Female	20.783593	13	infected	2019	Non	22.4199
32	Antsiranana	8	5	4	Female	20.783593	9	infected	2021	Non	17.189382

Lemurdata_glm_2 +

Ready 110%

Break – 5 minutes

Please download the tutorial if you have not already!

Week 2: Intro to simple statistics

- Monday, June 27: Introduction to simple statistics (**Kacie Ring**)
  - Simple statistics tutorial
  - Simple statistics Rmarkdown download
- Wednesday, June 29: Introduction to linear regression (**Sophia Horigan**)
  - Linear regression lecture
  - Linear regression tutorial
- *Mentor/Mentee Goal: Explore individual datasets + Refine research questions*
  - Please fill our this contract as a mentor/mentee pair by Friday, July 8, 2022!

Week 3: Intro to mixed modeling

- Monday, July 11: Intro to mixed modeling lecture (**Emily Ruhs / Dave Klinges**)
  - Mixed Models (zip)
- Wednesday, July 13: Intro to mixed modeling lecture tutorial (**Emily Ruhs / Dave Klinges**)
- *Mentor/Mentee Goal: Outline plan for research analyses on independent work*

Week 3 – here!

Week 4: Community biodiversity analyses

- Monday, July 25: Community biodiversity analysis lecture (**Katie Young**)
- Wednesday, July 27: Community biodiversity analysis tutorial (**Katie Young**)
- *Mentor/Mentee Goal: Outline 3-5 figures and accompany analyses for final paper*

Week 5: Building and fitting compartmental models in ecology

- Monday, August 8: Building and fitting compartmental models in ecology lecture (**Katie Gostic**)
- Wednesday, August 10: Building and fitting compartmental models in ecology tutorial (**Katie Gostic**)
- *Mentor/Mentee Goal: Begin work on figures for final paper*

Week 6: Model evaluation and comparison

An example from Madagascar

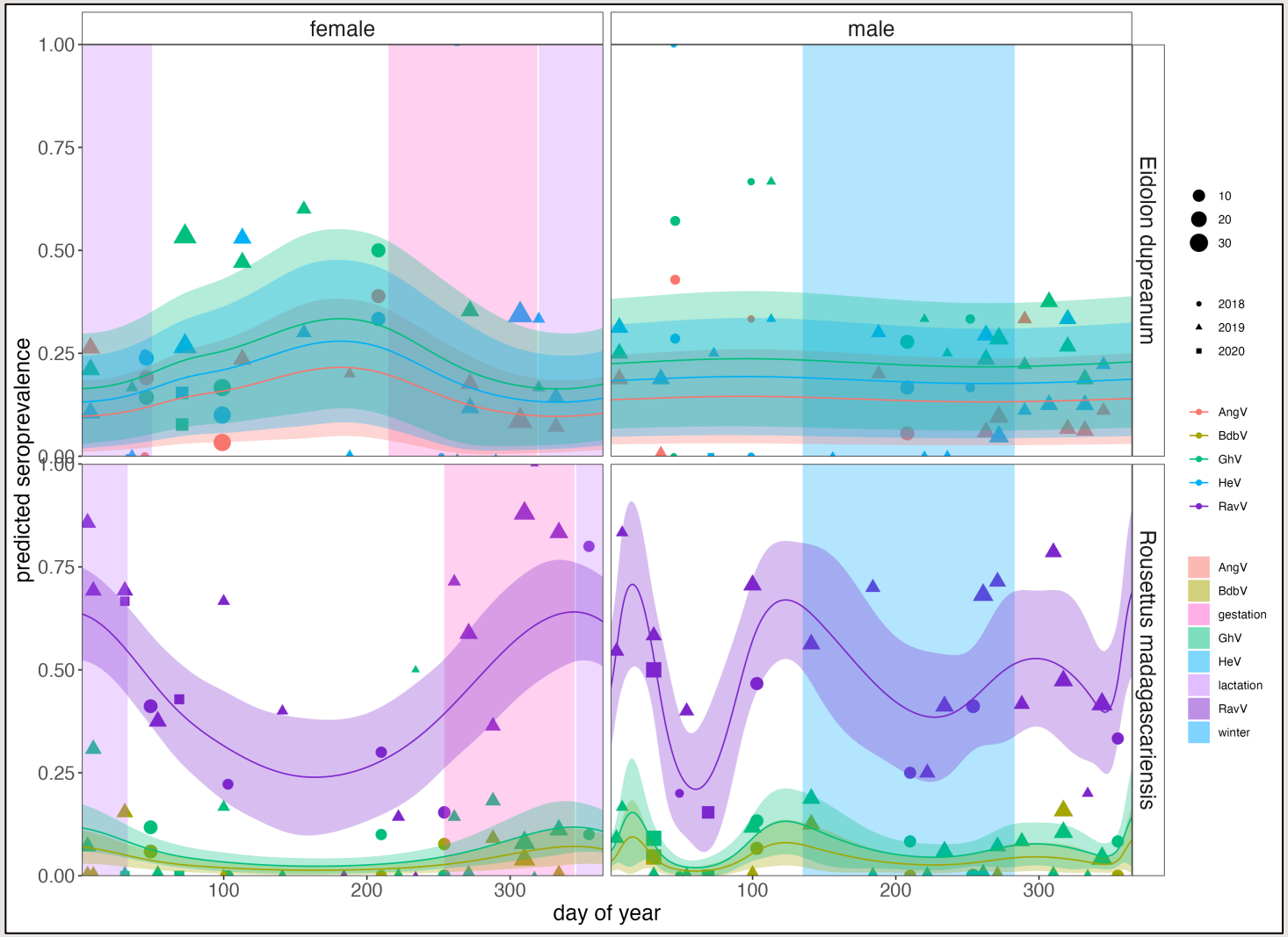
In the Brook lab – Collect data from three species of fruit bats throughout Madagascar.

What are the ecological drivers (or reasons) for increases in viral antibodies or infection throughout the year?

Reproduction? Food availability? Stress hormones?

An example from Madagascar

What about seasonality – food / reproduction?



An example from Madagascar

What best explains the patterns in seroprevalence?