# The use of Generalized Linear Mixed Models for the study of dynamical systems

Andrés Garchitorena

Researcher, Institut de Recherche pour le Développement

Research Advisor, PIVOT Madagascar

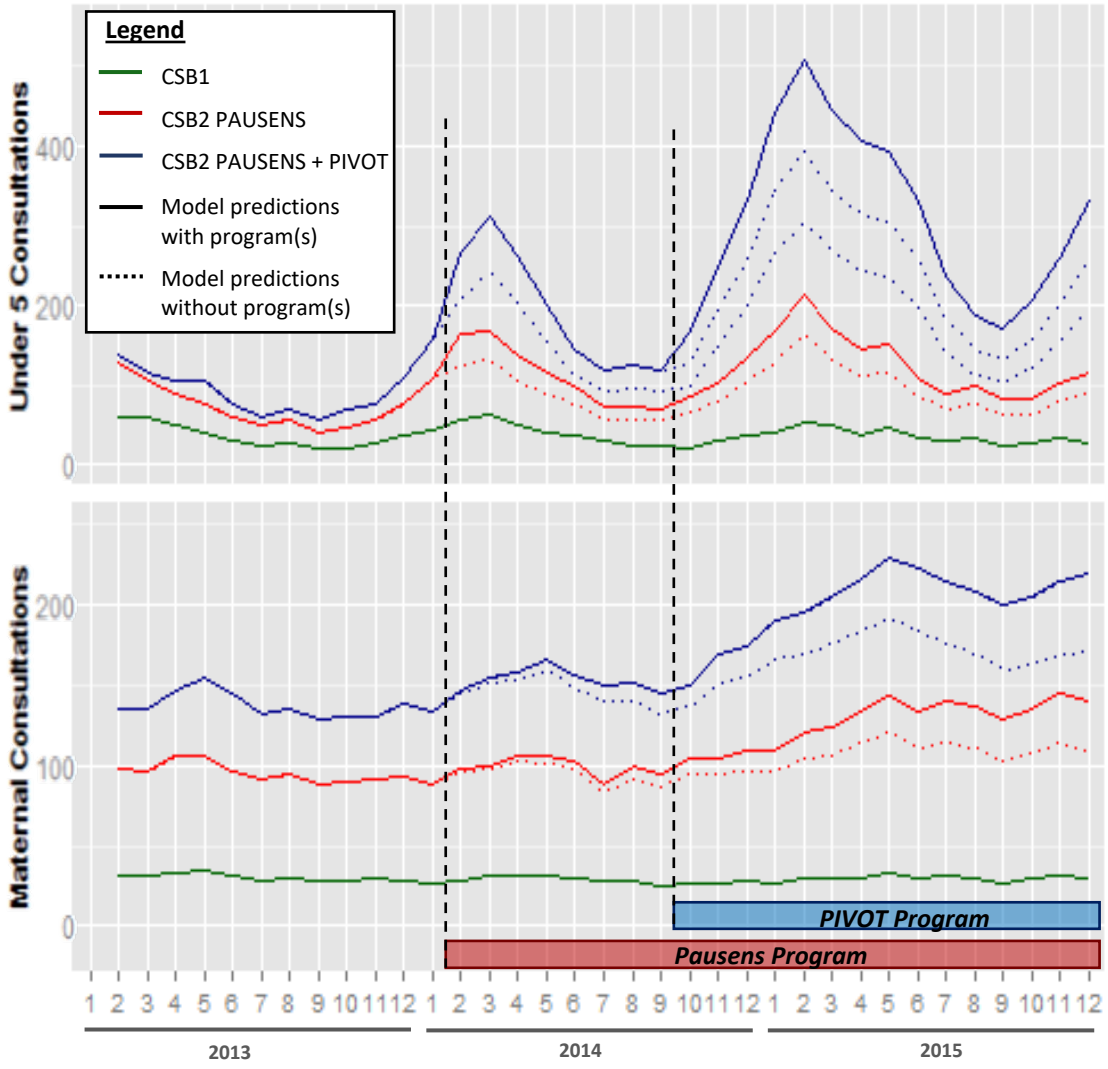
*$E^2M^2$ Workshop*
*Ranomafana, December 2022*

# Objectives of the lecture

- Understand alternatives to the use of mathematical models for the study of dynamical systems
- Remind some basic principles of linear regression and statistical models
- Introduce the use of generalized linear mixed models for the study of dynamical systems
- Provide an overview of the steps involved in developing a generalized linear mixed model (tutorial)

Why statistical models if my system is dynamic?

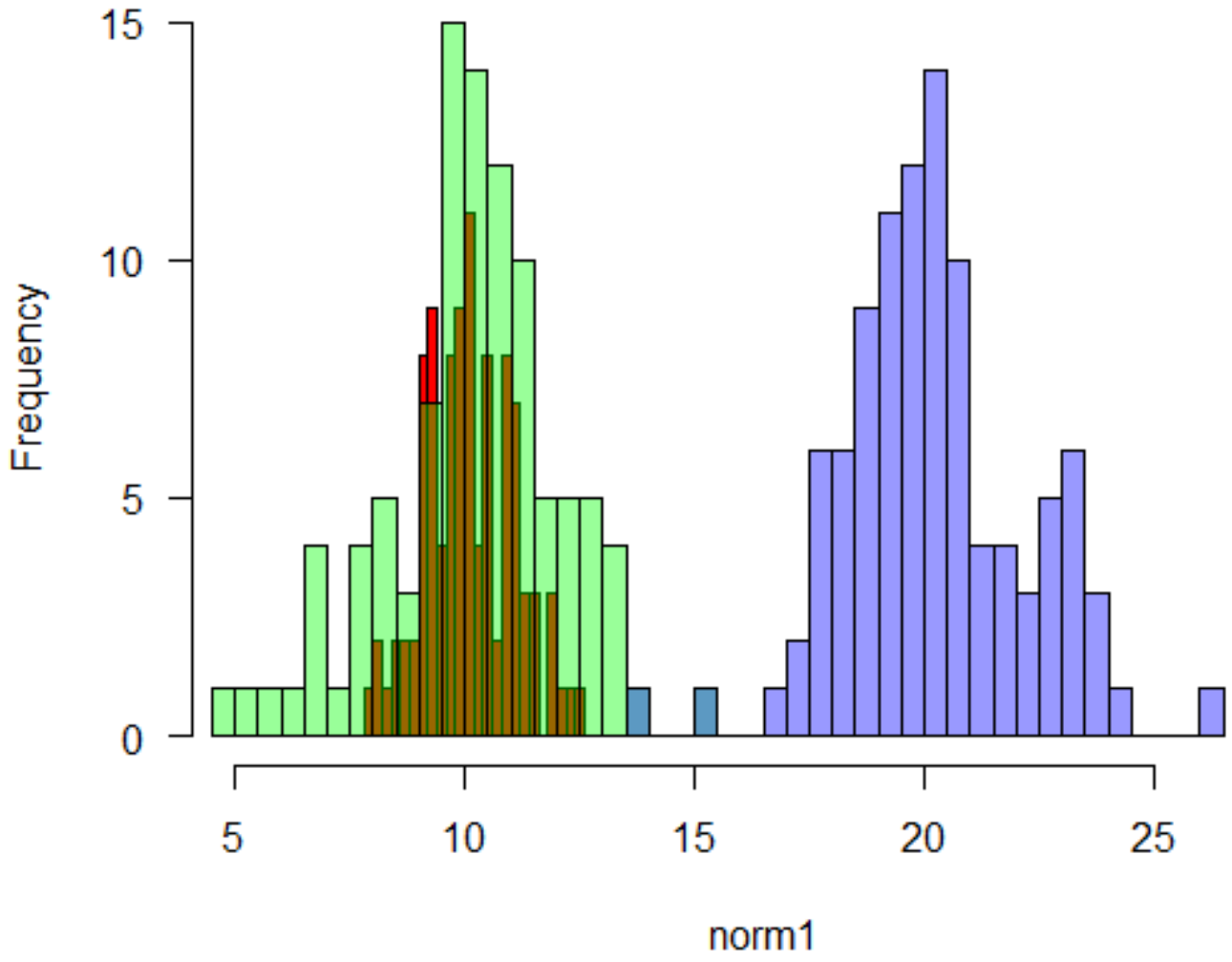
# Example of utilization trends



1. Univariate
Linear Models

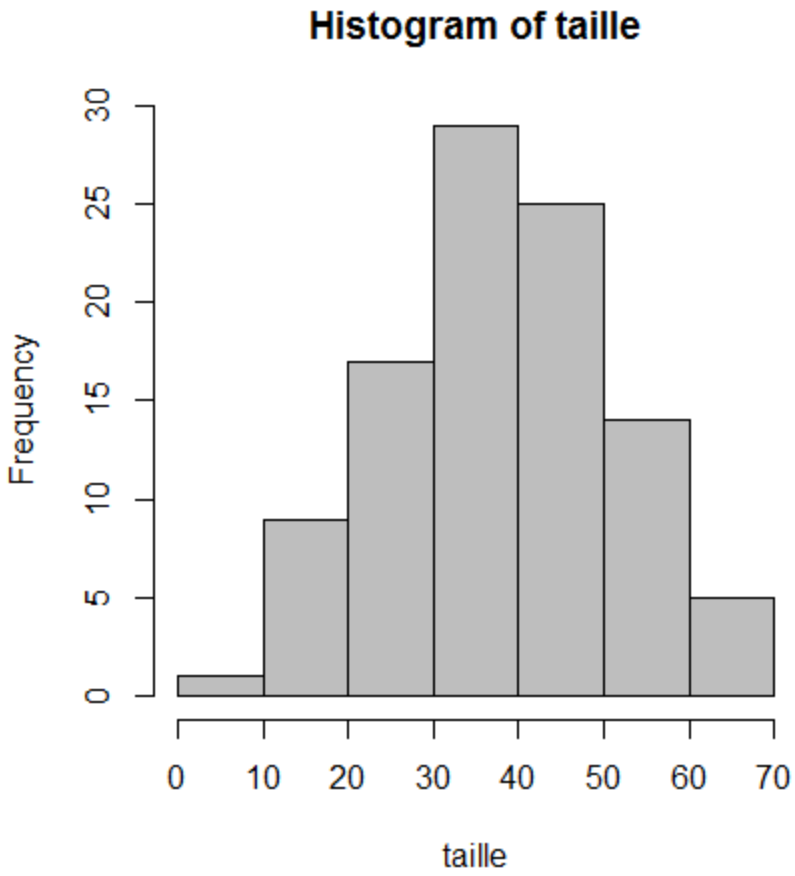
# SOME BASICS FIRST…

# Variables and distributions

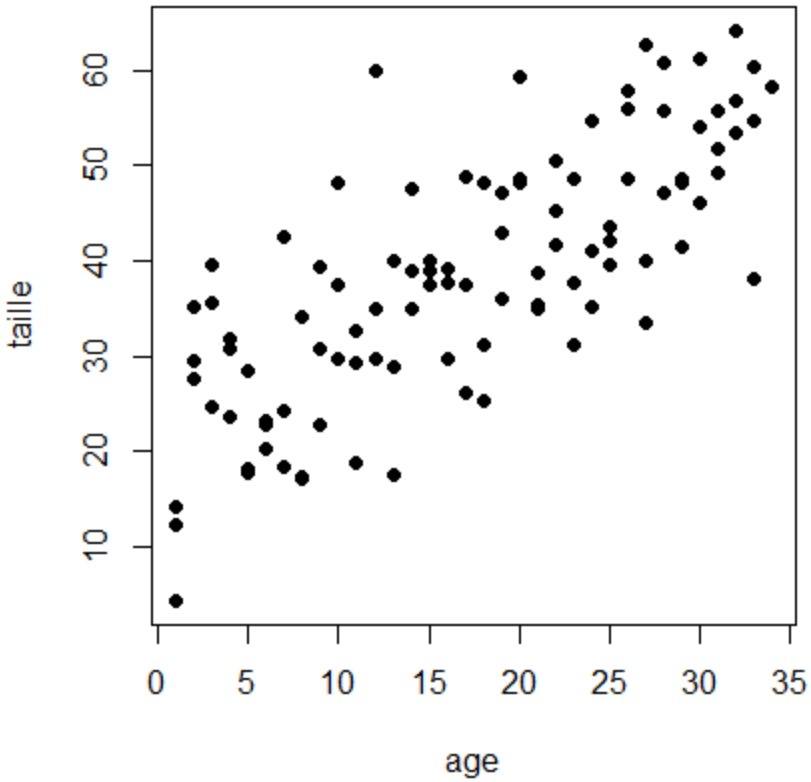


# Let’s work through a cute example

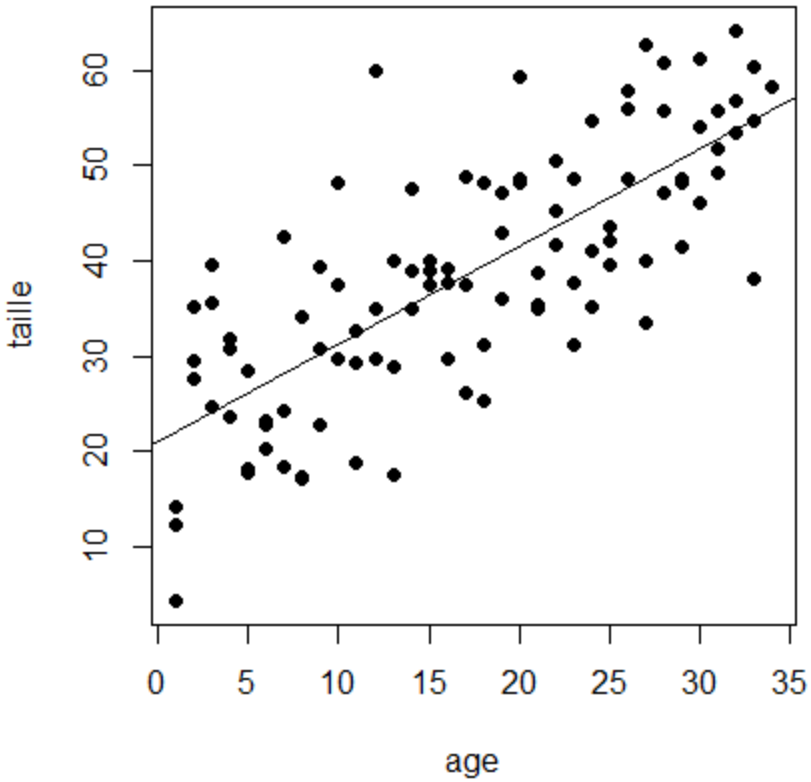
# Lemur weight and determinants



# Lemur weight and determinants

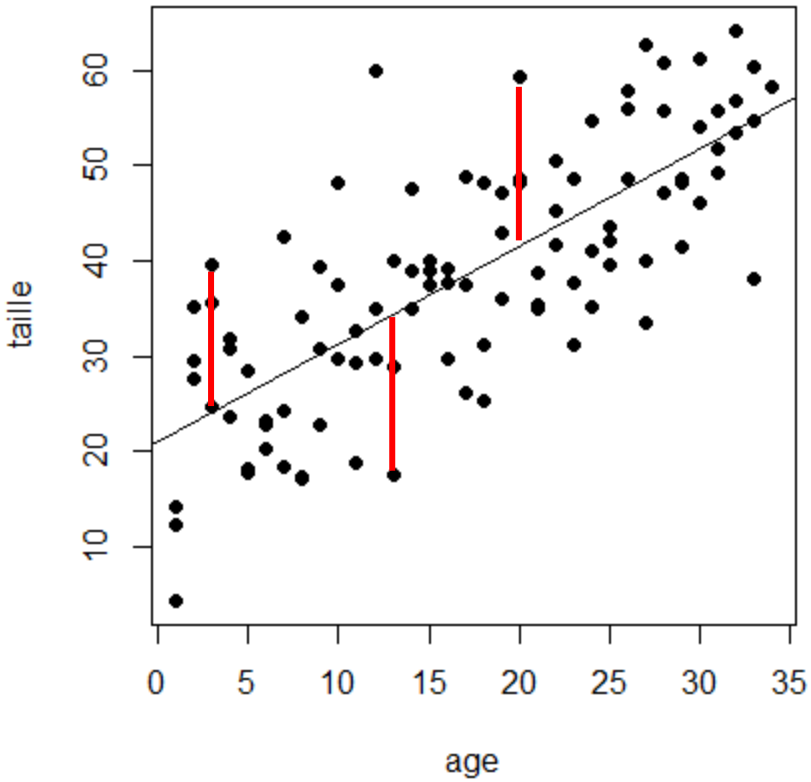


# Lemur weight and determinants



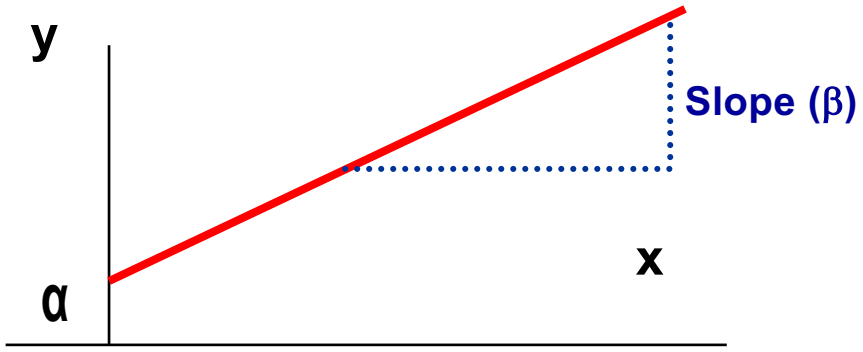
# Lemur weight and determinants

The goal is to minimize the difference between what we predict and what we observe



# Simple linear regression

- Relation between 2 continuous variables




- *Intercept ($\alpha$)*
  - Value of y when x is 0
- *Regression coefficient $\beta_1$*
  - Measures association between y and x
  - Amount by which y changes on average when x changes by one unit
- *Error ($\varepsilon$)*
  - Difference between the predicted values and observed values of y

# Simple linear regression

$$y = \alpha + \beta * x + \varepsilon$$

Response variable = Systematic component + Residual component

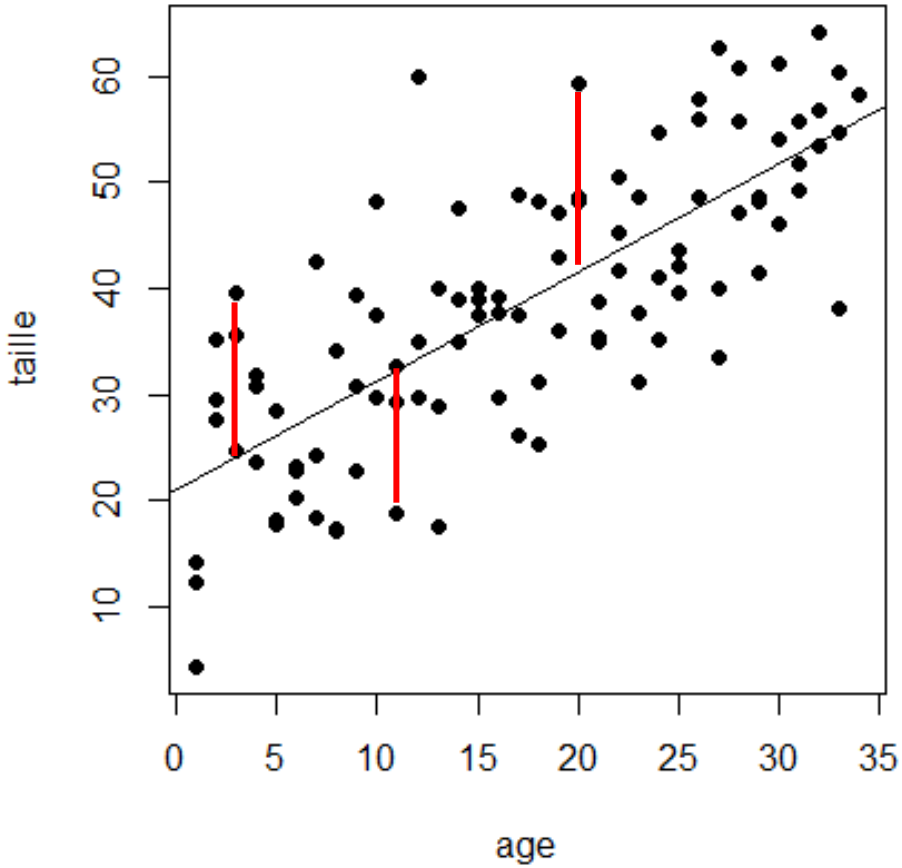
Systematic component → Intercept and explanatory variables

Residual component →
- Null mean
- Independence
- Fixed variance
- Normality

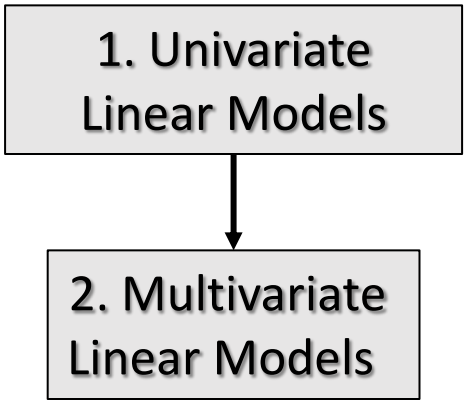
The R function to fit a linear model is lm() which uses the form
**fitted.model <- lm(formula, data=data.frame)**

# Simple linear regression

*Taille = 20 + 1.15 x Age (months) + Error*

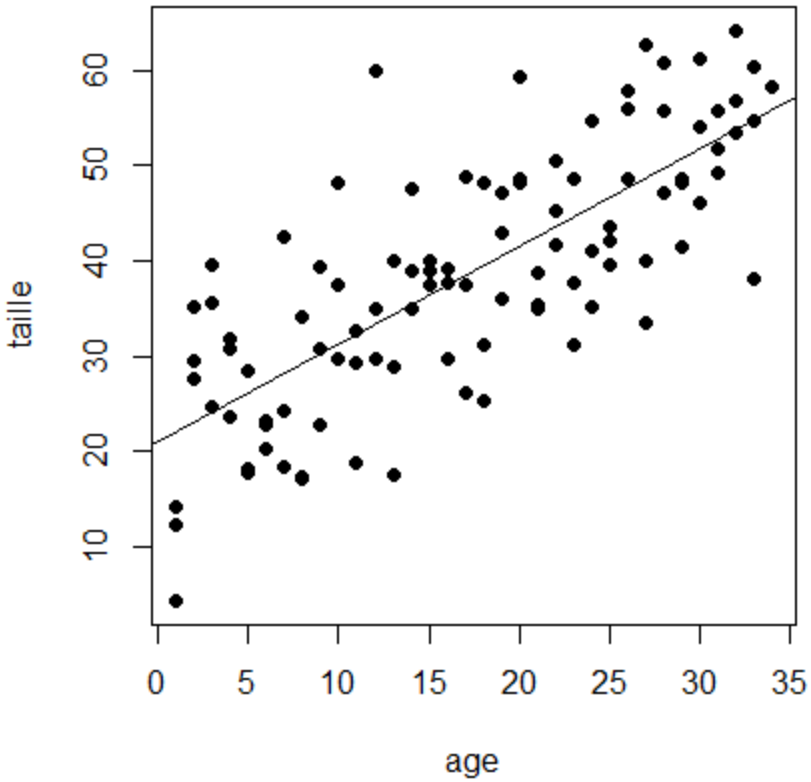


A process is generally the result of several others…



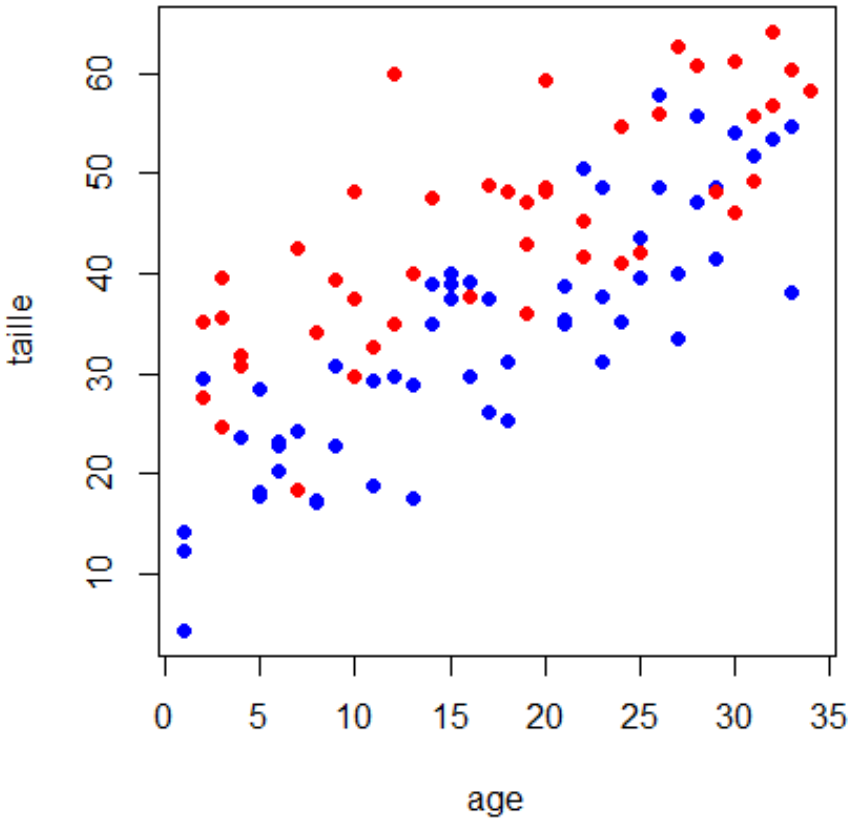

# INTRODUCING MULTIVARIATE LINEAR MODELS

# Lemur weight and determinants



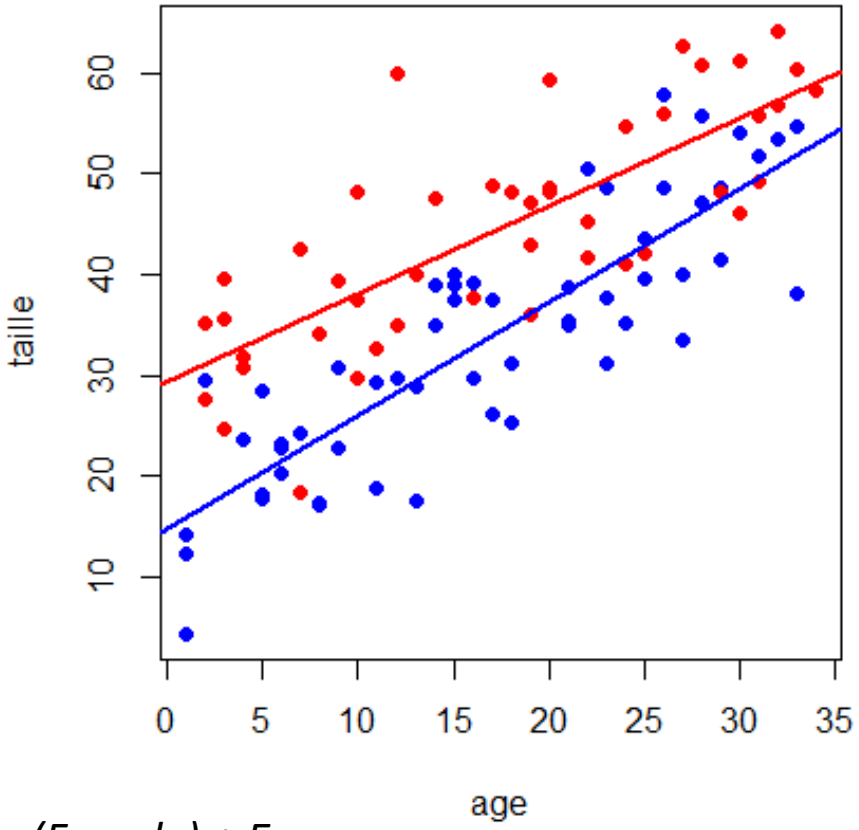
# Lemur weight and determinants

The effect of gender



# Lemur weight and determinants

The effect of gender

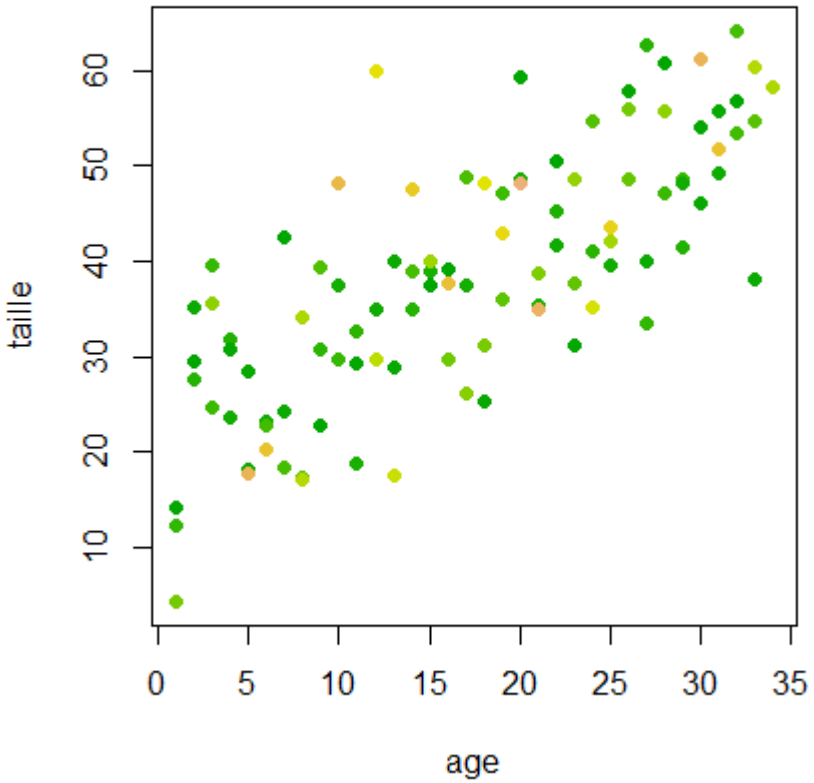



*Taille = 15 + 1.15 x Age (months) + 15 x Sexe (Female) + Error*

# Lemur weight and determinants

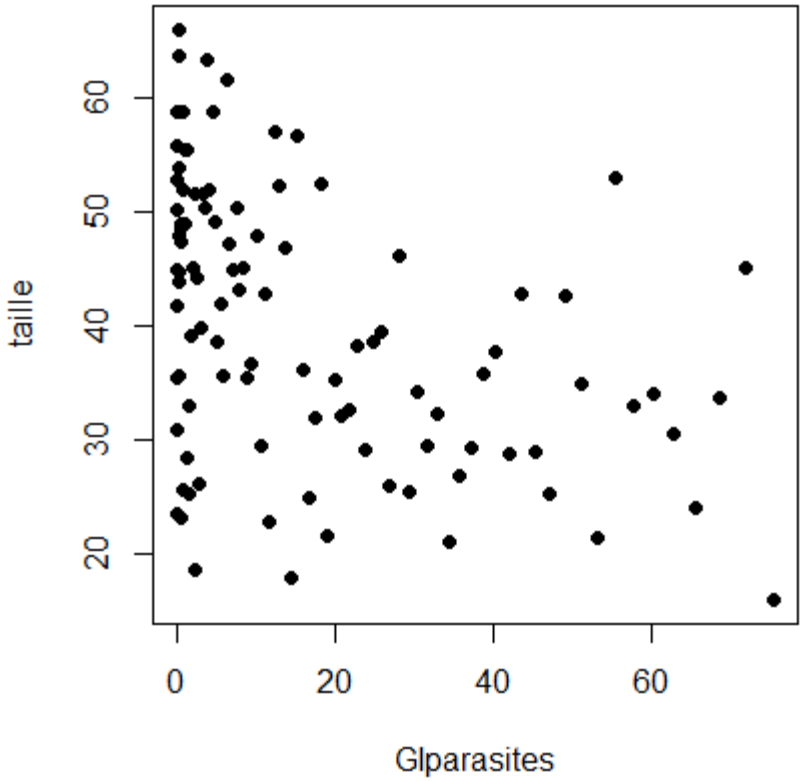
## The effect of parasites

Green: low GI parasite burden
Yellow: high GI parasite burden



# Lemur weight and determinants

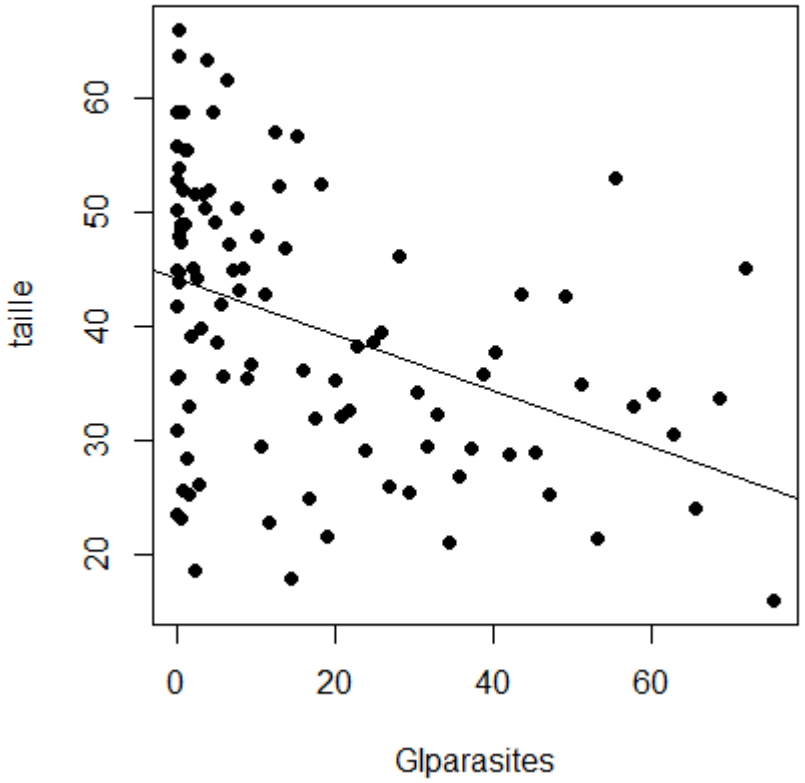
The effect of parasites



# Lemur weight and determinants

The effect of parasites

Taille = 45 - 0.3 x Nb Parasites + Error



# Multiple linear regression

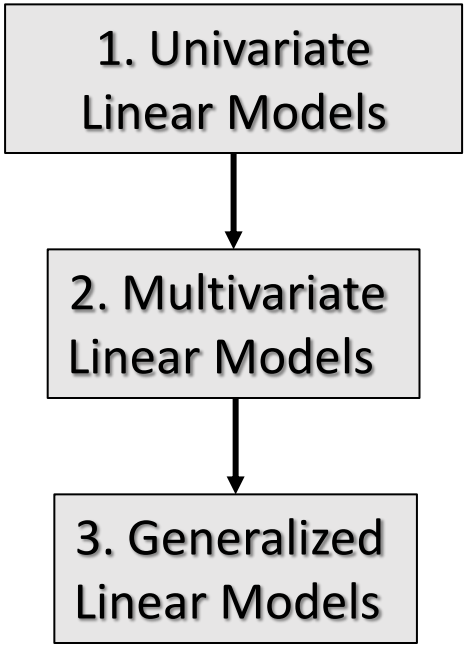
- Generalization of simple regression

- To describe the relationship between
  - The response variable, y
  - The explanatory variables, $x = (x_1, x_2, \ldots, x_n)$

- The model: $y = \alpha + \beta_1 * x_1 + \ldots + \beta_n * x_n + \varepsilon$
  with $\varepsilon \sim N(0, \sigma^2)$

- We generally select the model that best fits the data (best explains observed patterns) with the smallest number of variables

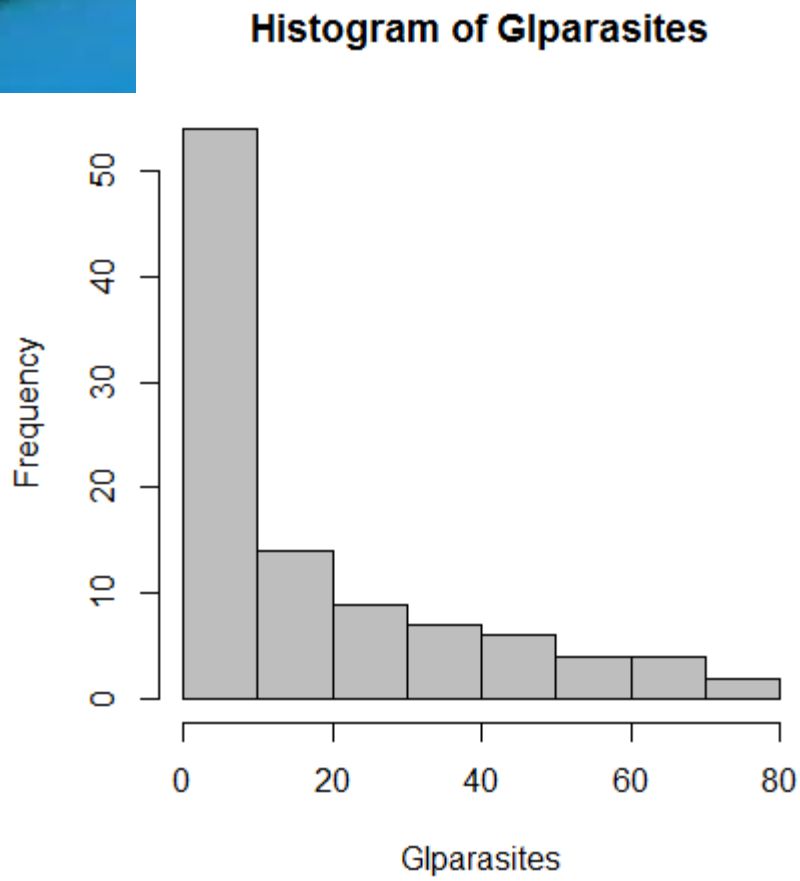
Unfortunately, not all things in life are normal…




# INTRODUCING GENERALIZED LINEAR MODELS

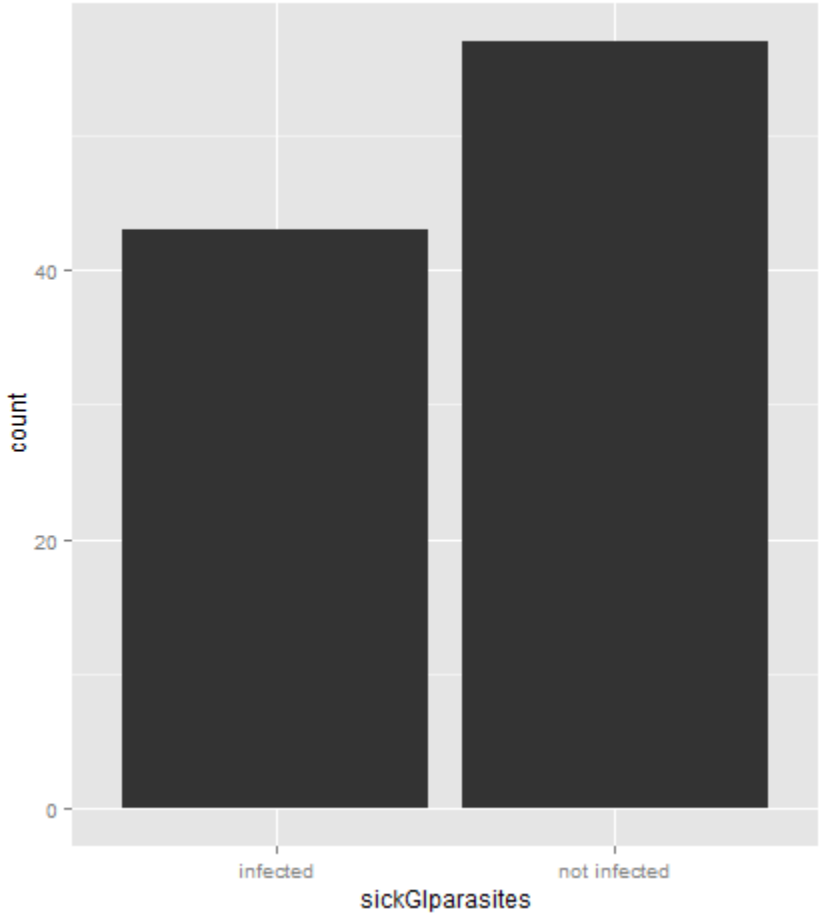
# Count data

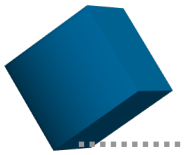
- Cannot be negative
- Discrete values
- The lower the values, the « less normal » they generally are.
- Examples:
  - Number of individuals of a species X
  - Number of people with a disease X



# Binary data (events)

- Values either 1 or 0 (either happened or not happened)
- The outcome variable is the number of successes /failures
- Examples:
  - Presence of a species X
  - Presence of a disease X





# Limitations of linear models

- In this type of situations, general linear models are not appropriate because:
  - The range of Y is restricted (e.g. binary, count)
  - The variance of Y depends on the mean

- **Generalized linear models** extend the linear model framework to address both of these issues by using a linear predictor and a link function

The R function to fit a general linear model is glm() which uses the form
**fitted.model <- glm(formula, family="model family", data=data.frame)**

# Generalized linear modeling

- One generalization of multiple linear regression. Response, y, predictor variables $x_1$, $x_2$, …. The distribution of Y depends on the X's through a single linear function, the "linear predictor"

$$\nu = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

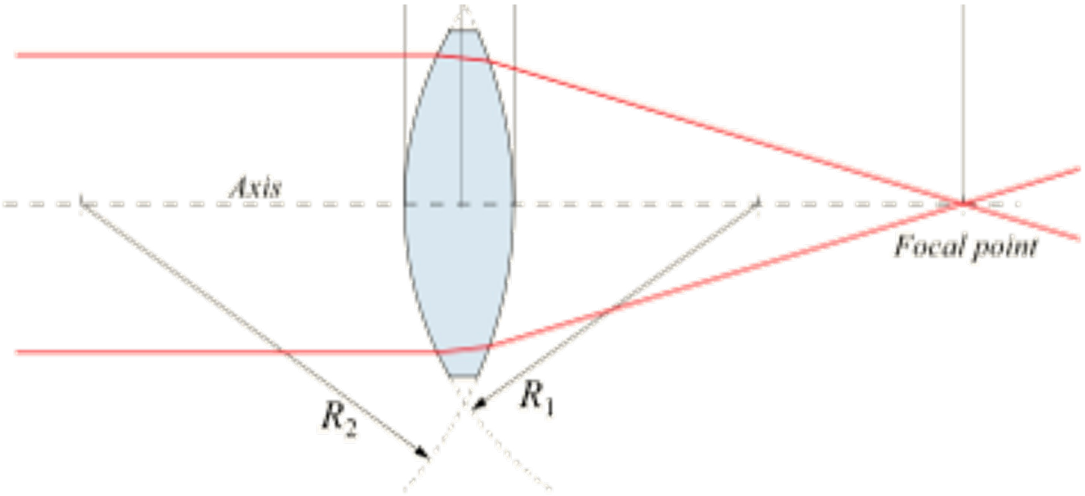
- A link function describes how the mean E(Y) = μ, depends on the linear predictor *v.*

$$\mu = m(\nu), \qquad \nu = m^{-1}(\mu) = l(\mu)$$

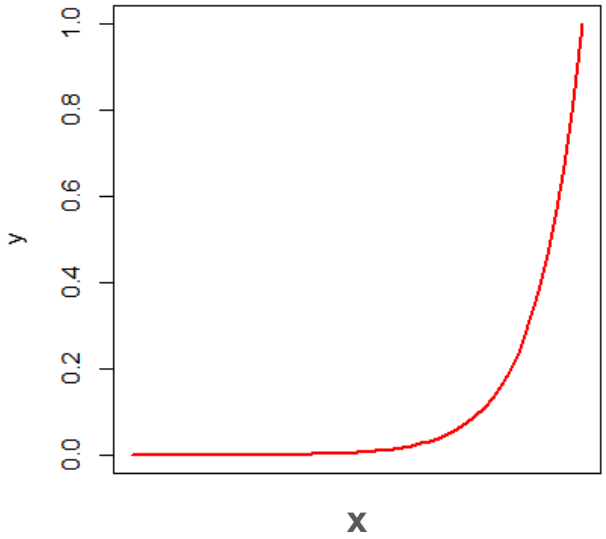
# Generalized linear modeling

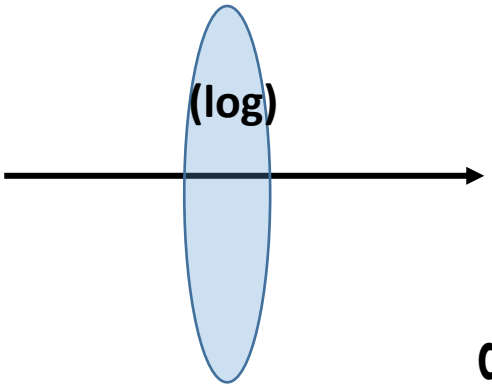
Most common families and links

- Gaussian: identity
- Poisson: log
- Binomial:  logit
- Negative binomial: log

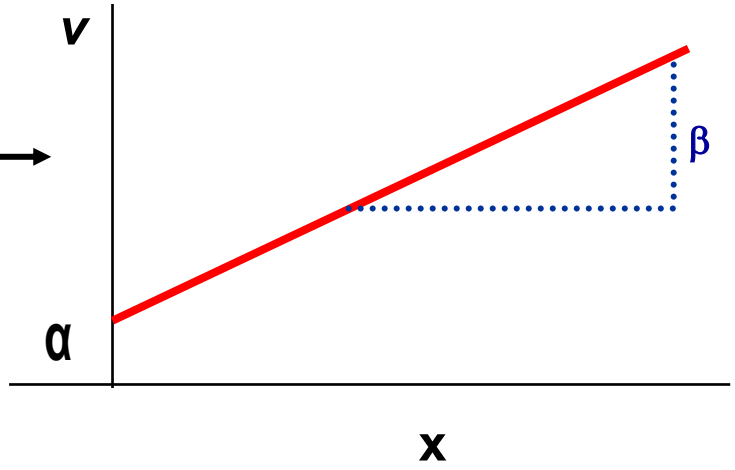



$Y = e^{(\alpha + \beta x)}$



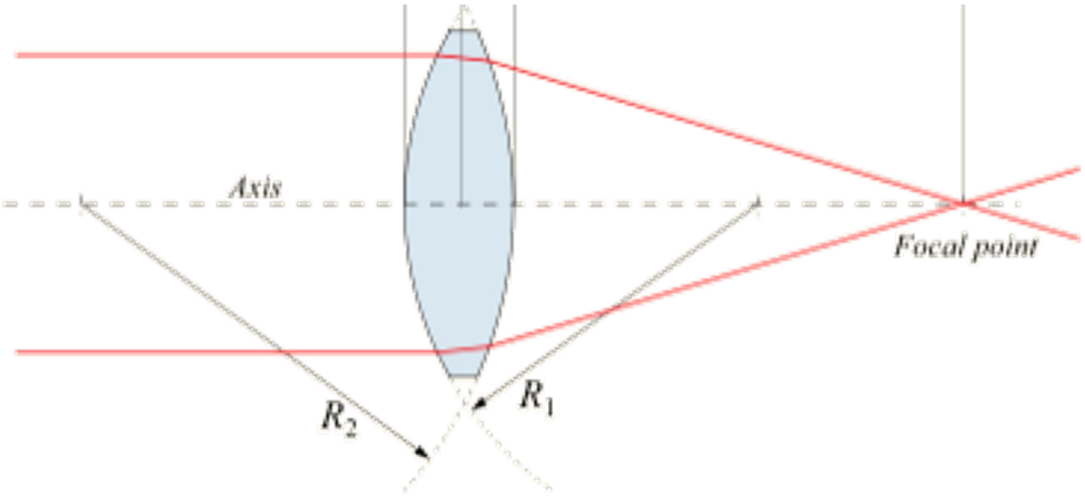

$V = \alpha + \beta x$



# Generalized linear modeling

Most common families and links

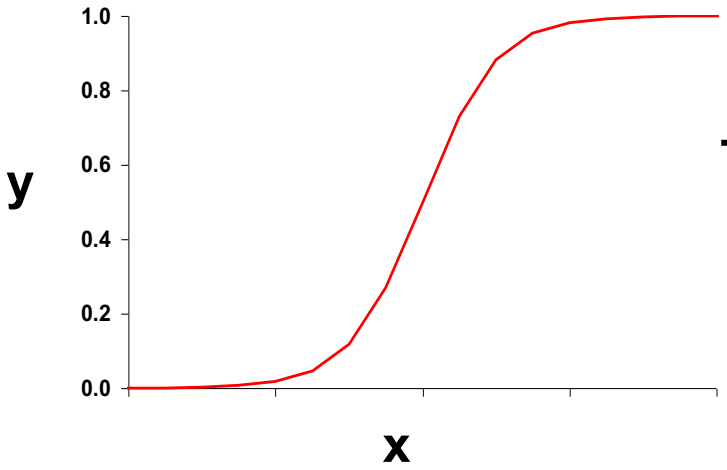
- Gaussian: identity
- Poisson: log
- Binomial:  logit
- Negative binomial: log

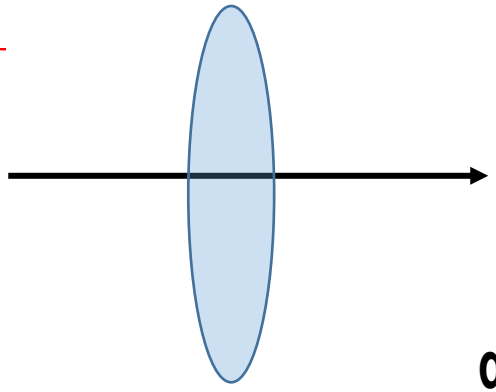


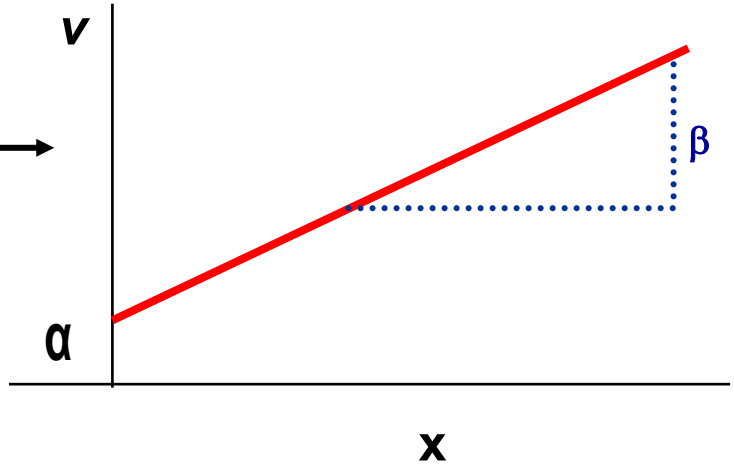

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

**(logit)**

$V = \alpha + \beta x$



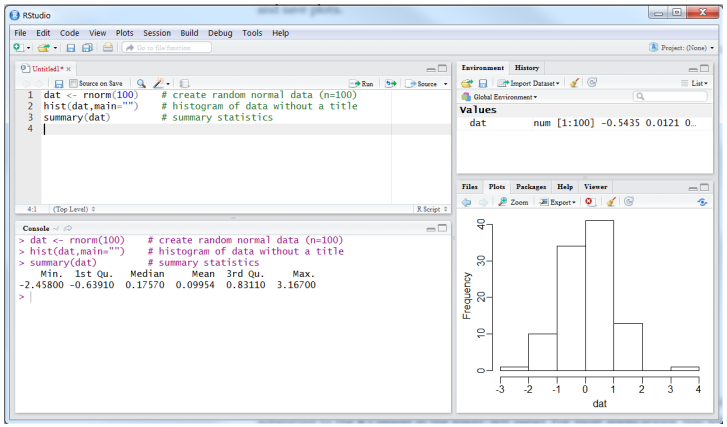





1. Univariate Linear Models

2. Multivariate Linear Models

3. Generalized Linear Models

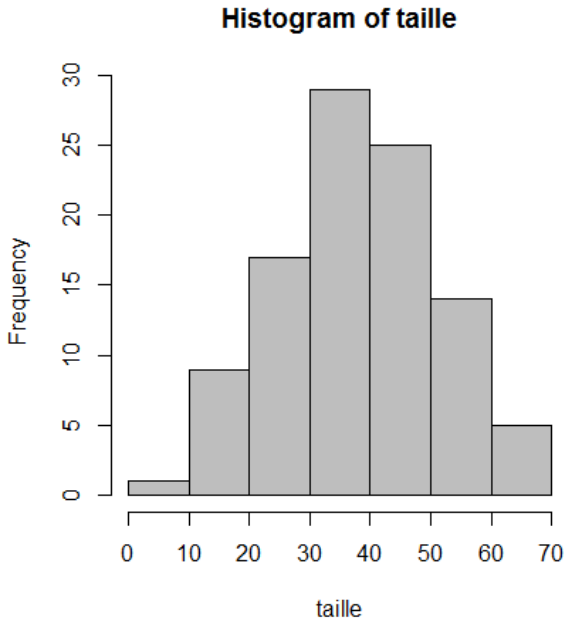



# STEPS IN DEVELOPMENT OF STATISTICAL MODELS (TUTORIAL)

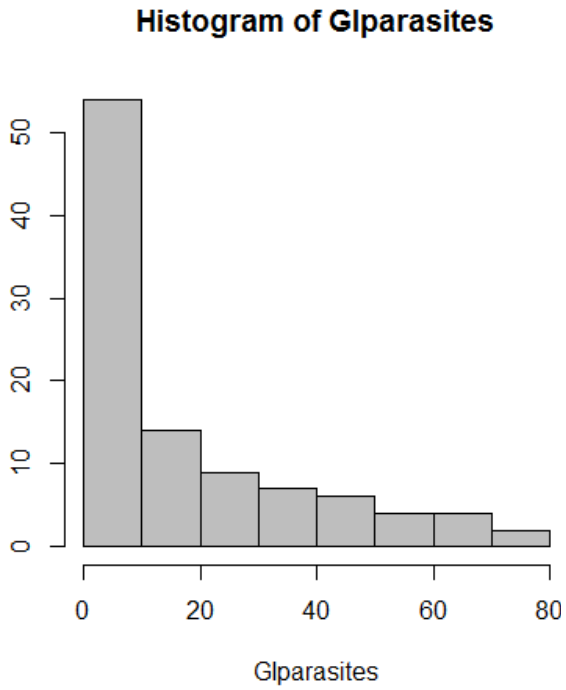
# Database construction and descriptive analyses

- Distribution of the response variable
- Distribution of the explanatory variables

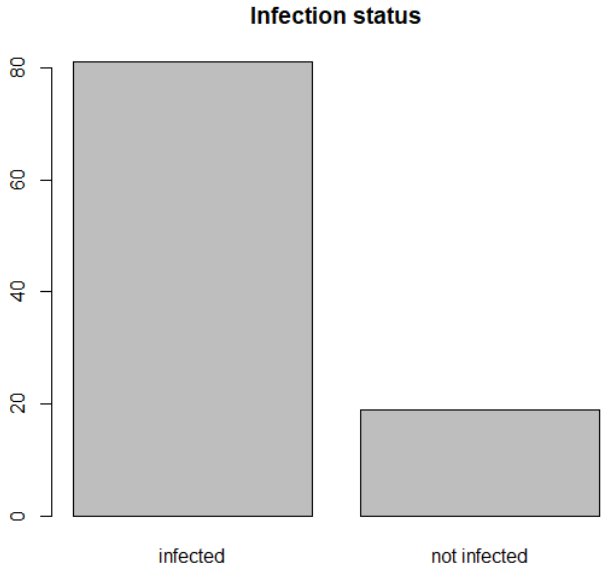
hist(mydata$var)




hist(mydata$var)



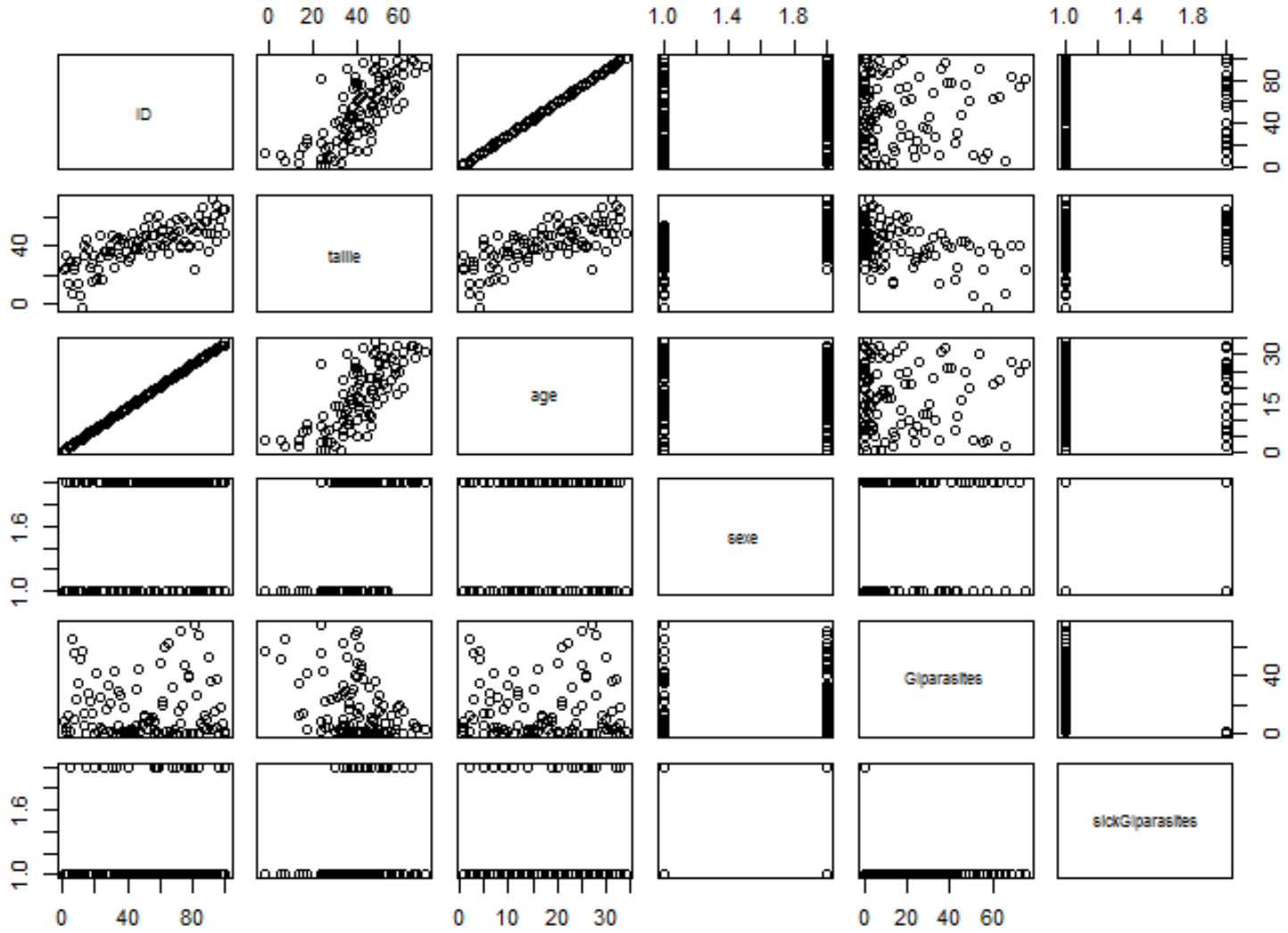

plot(mydata$var)



# Database construction and descriptive analyses

- Relationships between the variables

pairs(mydata)



# Univariate analyses

- Quantify the stregth of the relationship between the response variable and each explanatory variable
- Test the significance of the relationship between the response variable and each explanatory variable

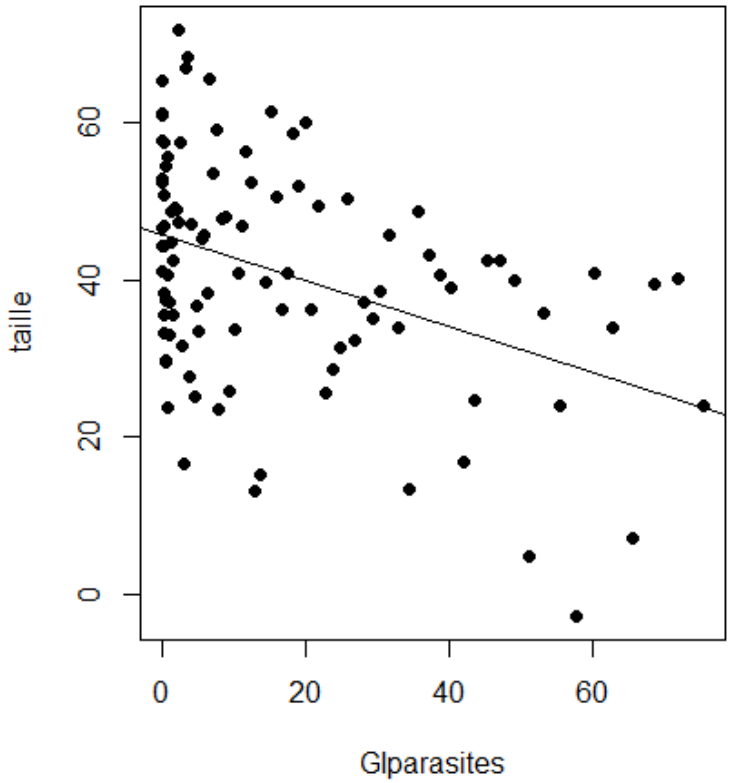
Model1 = lm(taille~GIparasites, data=mydata)
summary (Model1)

```
Call:
lm(formula = taille ~ GIparasites)

Residuals:
    Min      1Q  Median      3Q     Max
-31.605  -8.351   1.113   9.901  26.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.8267     1.7154  26.714  < 2e-16 ***
GIparasites  -0.2927     0.0651  -4.495 1.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.07 on 98 degrees of freedom
Multiple R-squared:  0.171,     Adjusted R-squared:  0.1625
F-statistic: 20.21 on 1 and 98 DF,  p-value: 1.906e-05
```



# Multivariate analyses

- Quantify the relationship between the response variable and a set of explanatory variables

Model1 = lm(taille~age+sexe+GIparasites, data=mydata)
summary (m1)

```
Call:
lm(formula = taille ~ age + sexe + GIparasites, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-16.9962  -2.6011  -0.1584   3.7331  12.0600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.94145    1.28143   17.12   <2e-16 ***
age          1.02365    0.05584   18.33   <2e-16 ***
sexeMale    10.88561    1.09295    9.96   <2e-16 ***
GIparasites -0.29930    0.02652  -11.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.323 on 96 degrees of freedom
Multiple R-squared:  0.8653,    Adjusted R-squared:  0.8611
F-statistic: 205.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

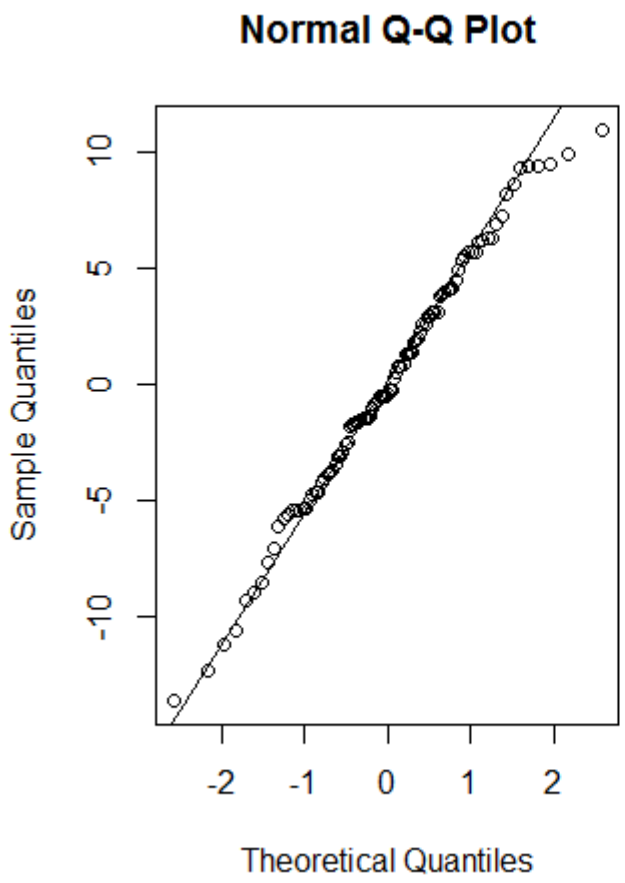
- Select the set of predictors that best explains the response variable (backwards, forward, stepwise)

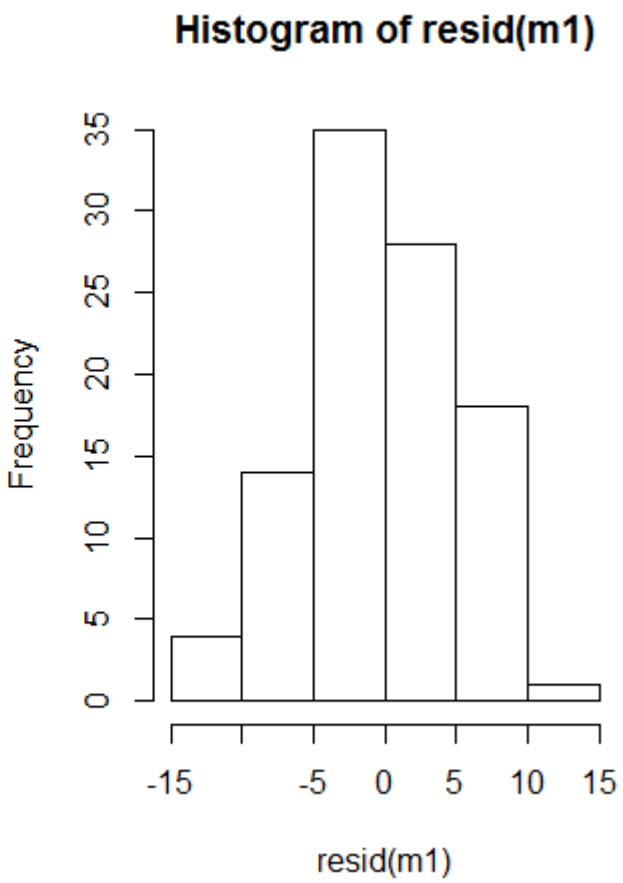
drop1 (m1)  add1 (m1)  step (m1)

# Model validation

- Check that model assumptions have not been violated

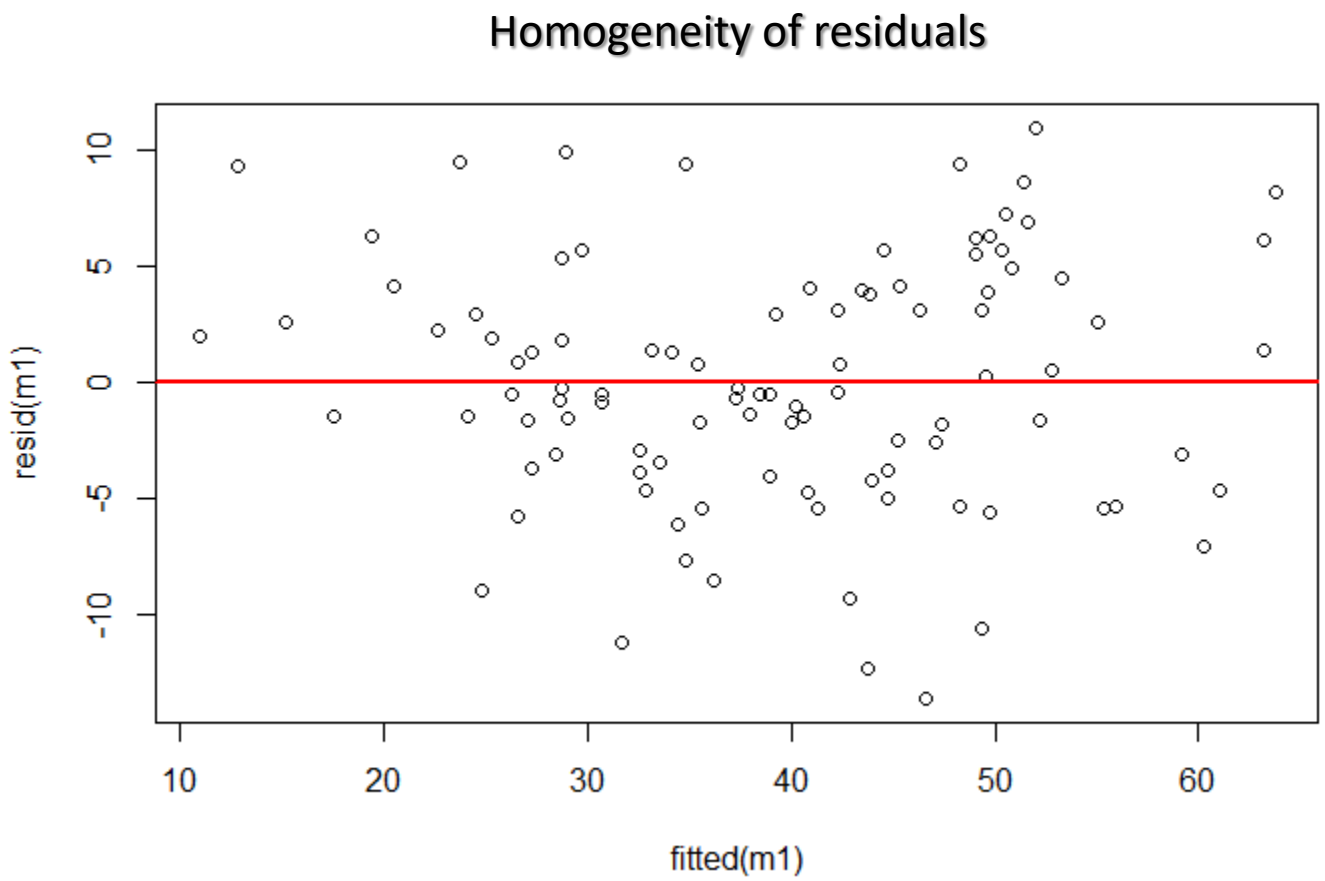
Normality of residuals

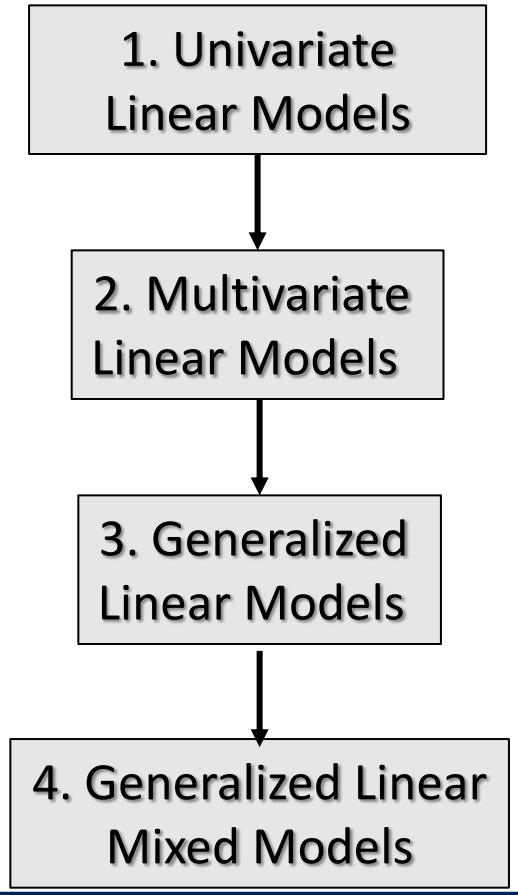


# Model validation

- Check that model assumptions have not been violated



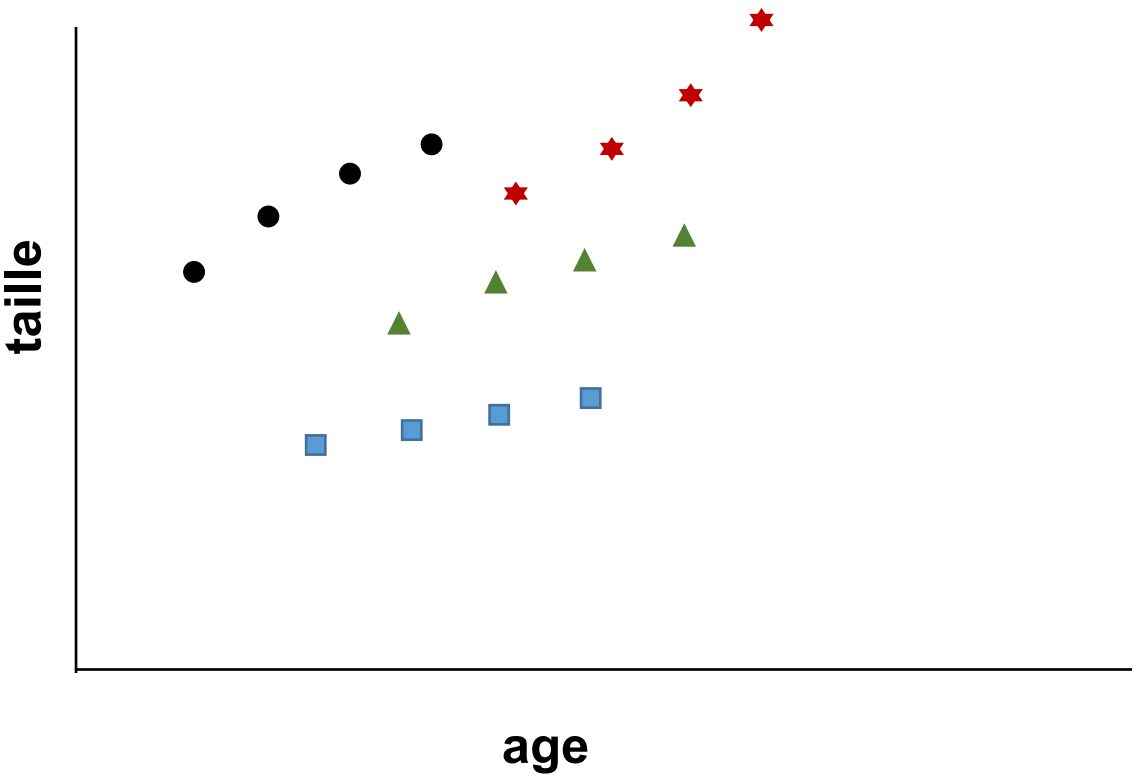



# INTRODUCTION TO GENERALIZED LINEAR MIXED MODELS

# Assumption and limitation of glms

all observations are considered **independent**

# What if…?




# Why use GLMMs?

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

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

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i$$

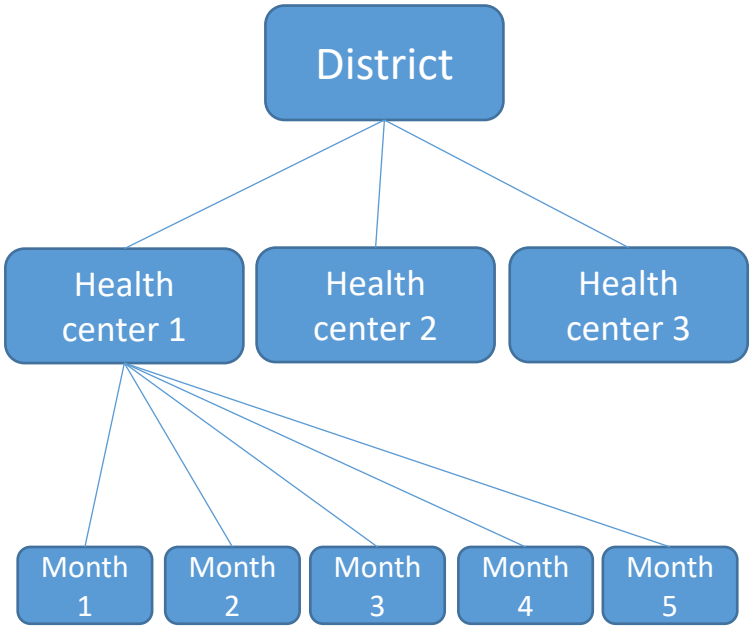
*Fixed Effects*

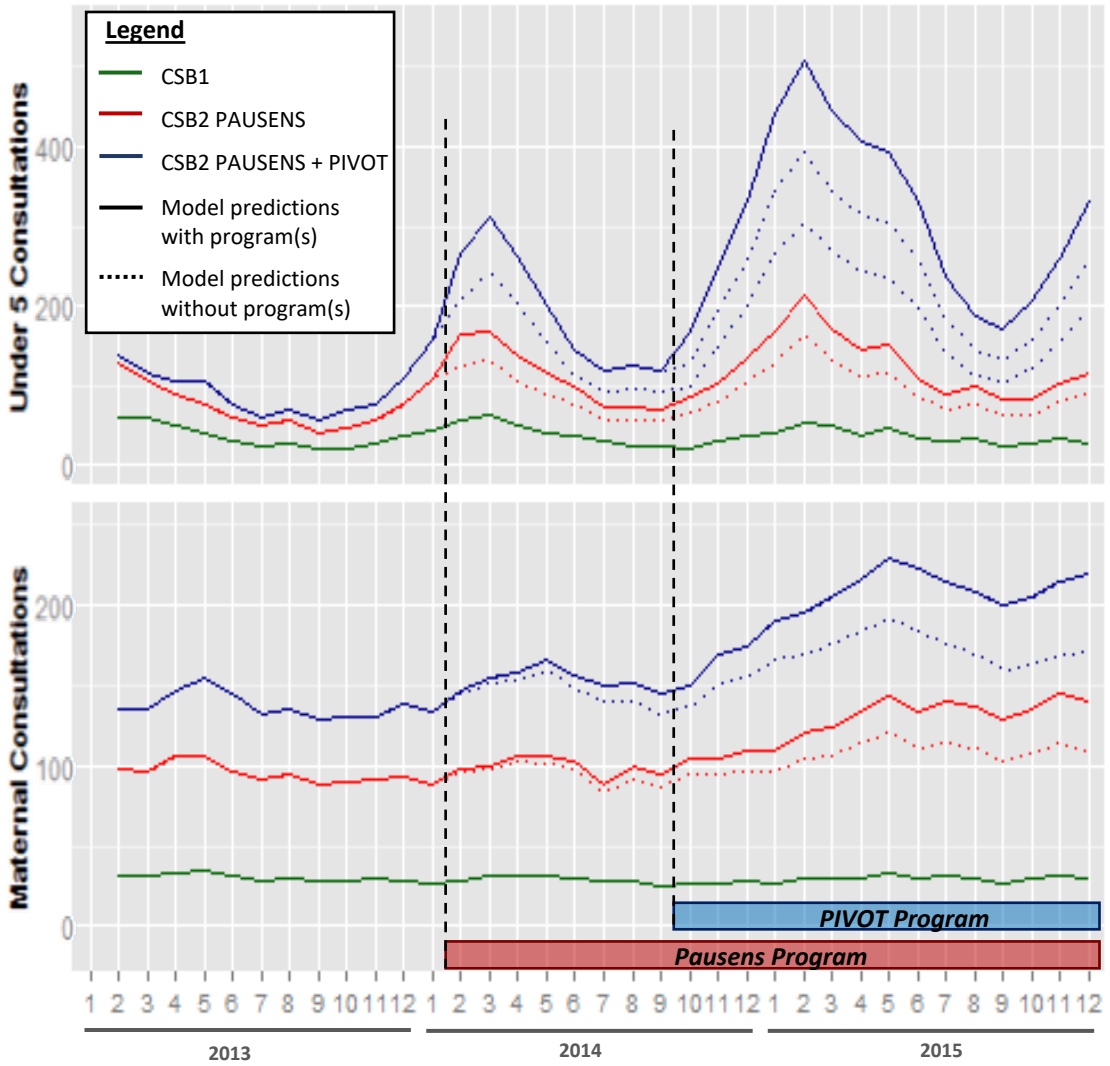
*Random Effects*

The R function to fit a generalized linear mixed model is glmer() which uses the form

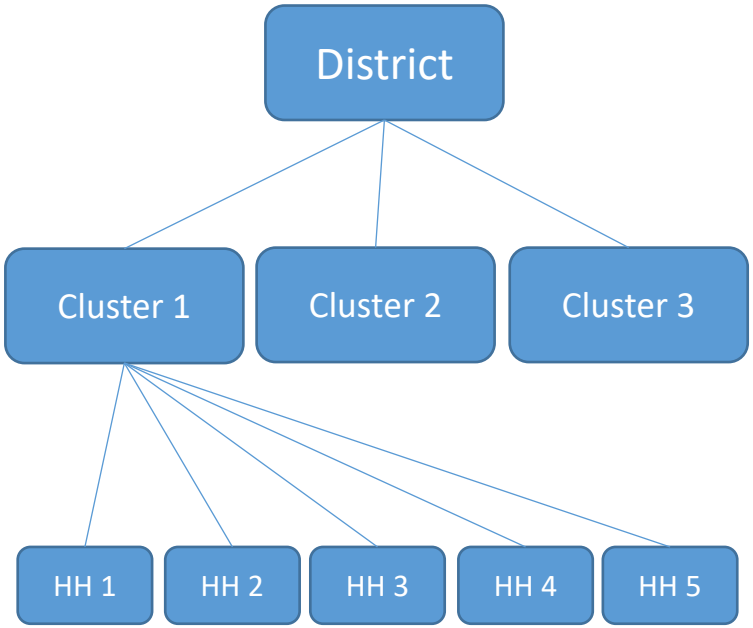
**fitted.model <- glmer(formula, family="model family", data=data.frame)**

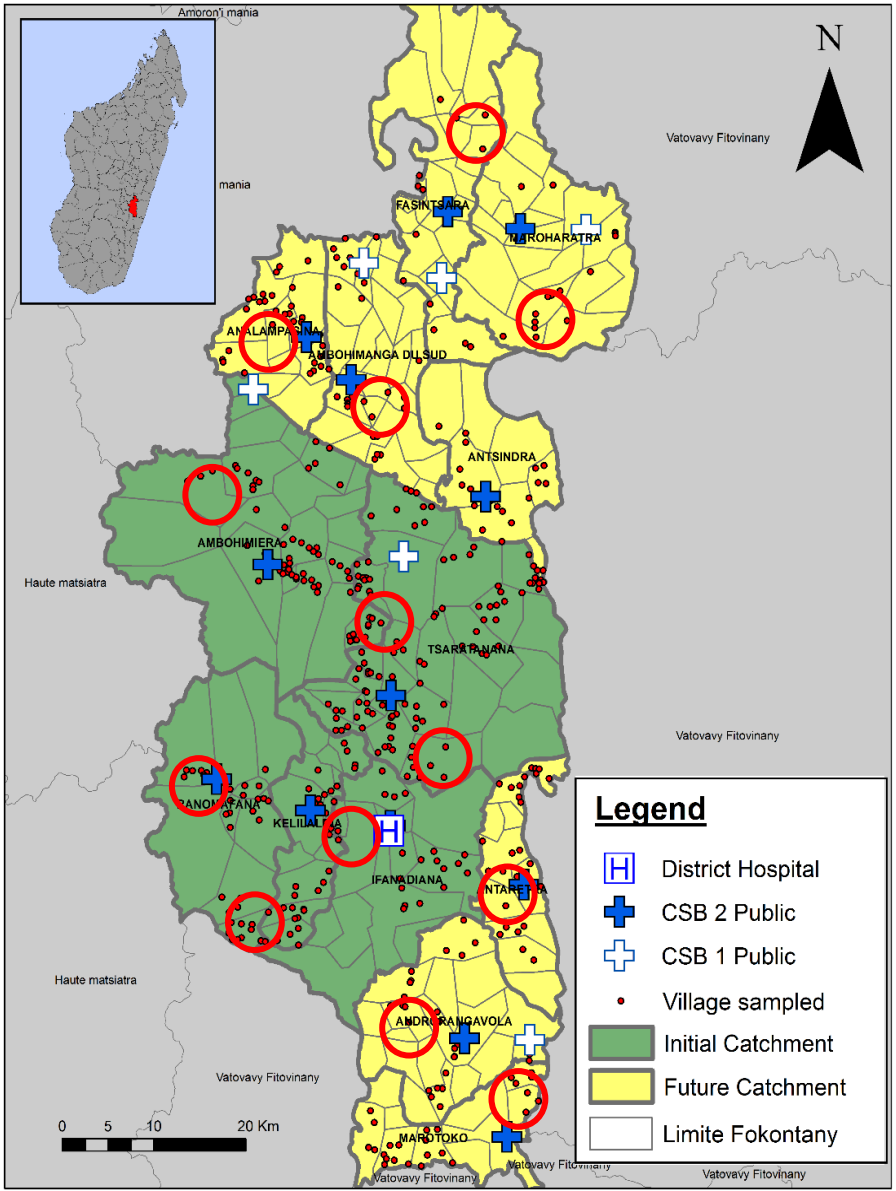
# Repeated measures



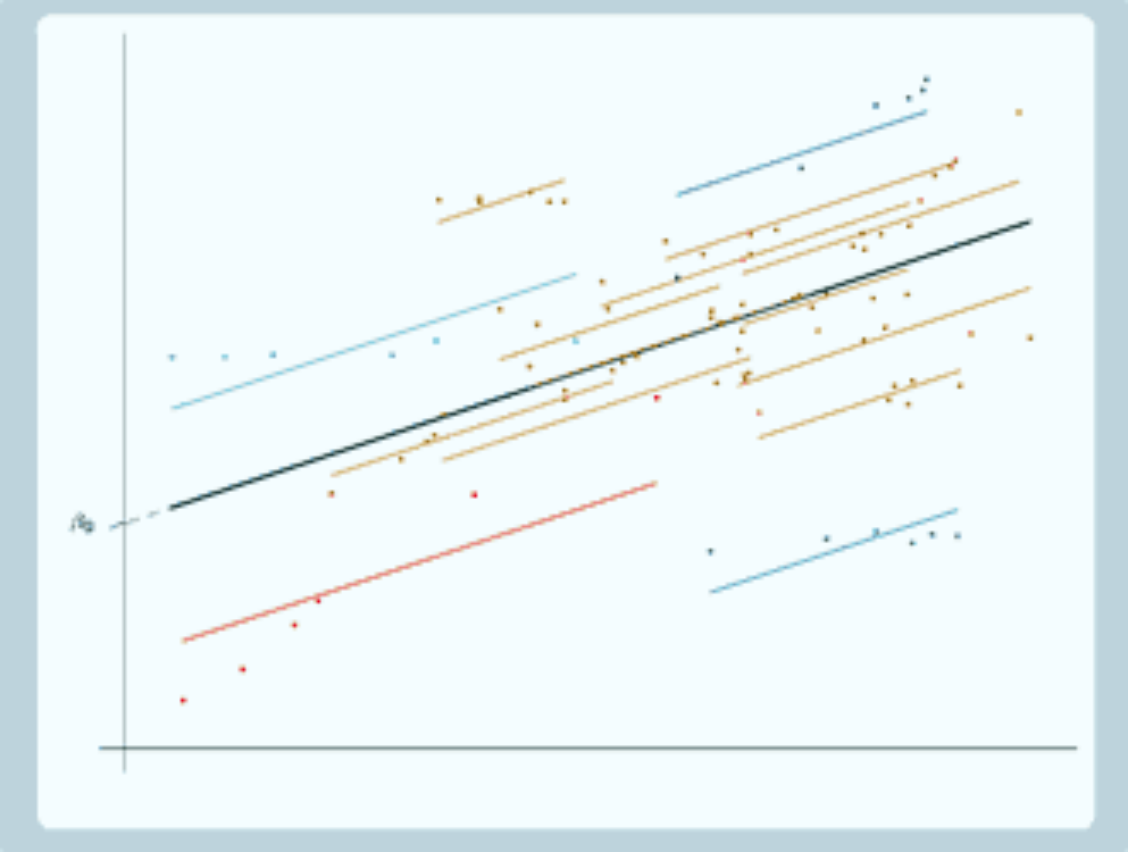
# Spatial correlation



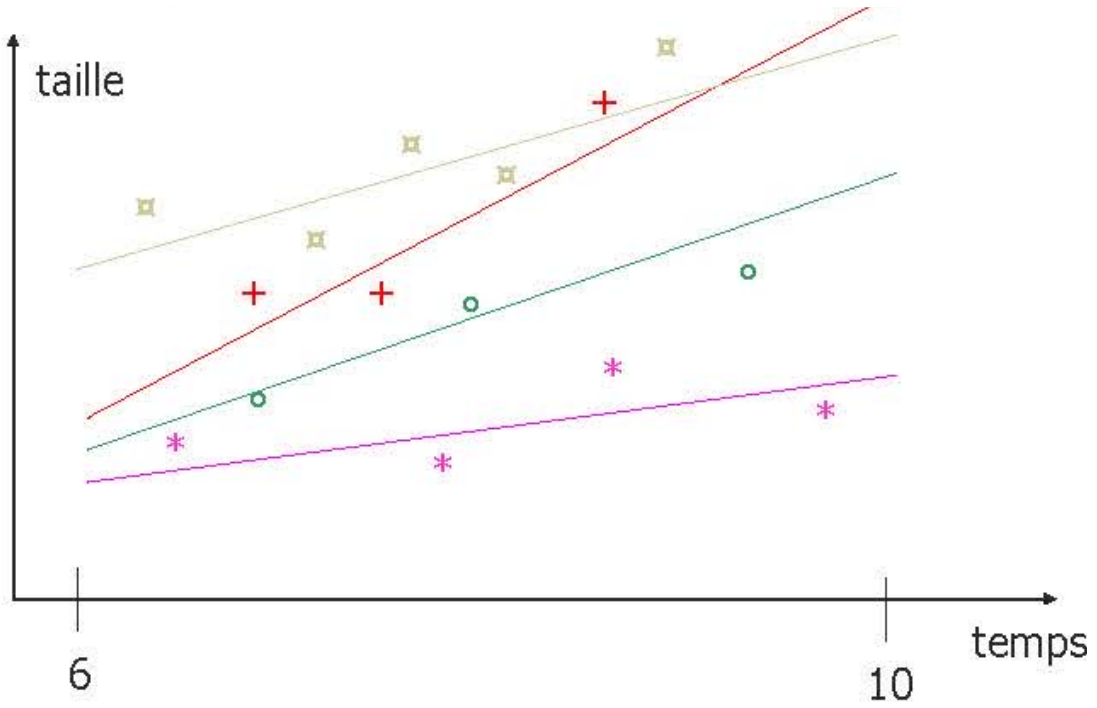
# Random intercept

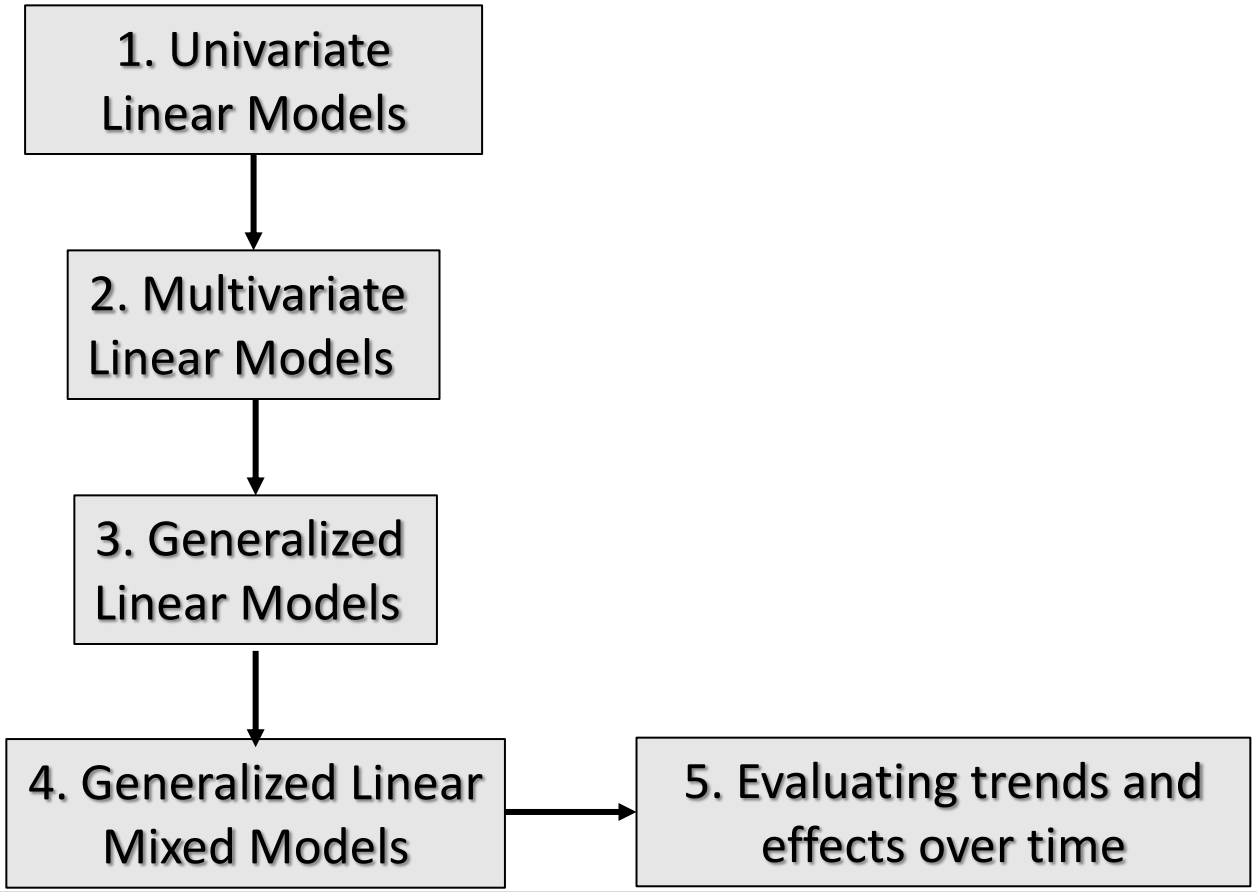
- The intercept is different for each individual/site
- Accounts for baseline differences in the response variable between individuals/sites



# Random slope

- The effect of a variable (b) is different for each individual/site
- Accounts for baseline differences in the relationship response-explanatory variable between individuals/sites





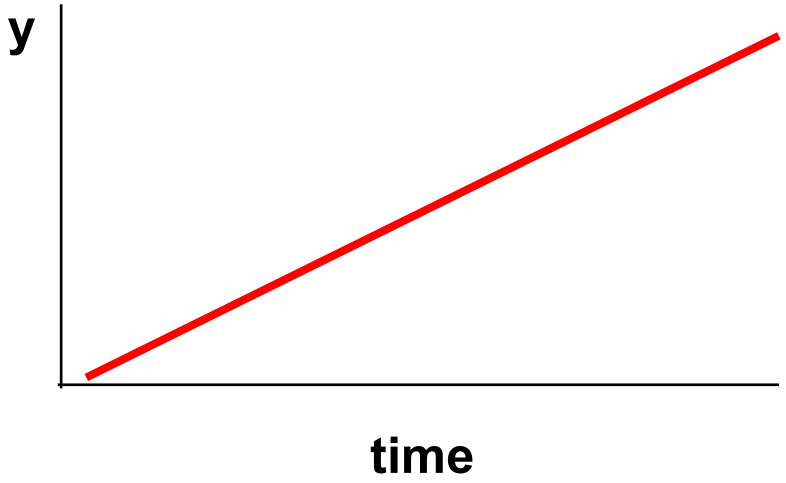

# NOW THAT WE CAN MODEL REPEATED OBSERVATIONS OVER TIME…

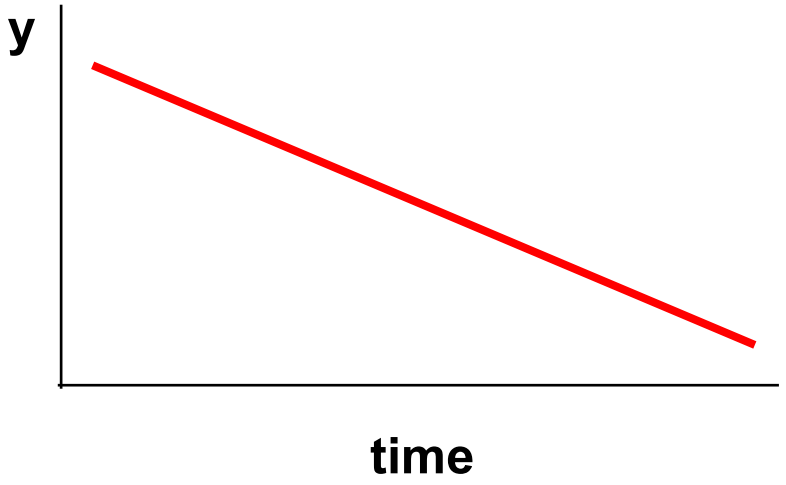
# Introducing time-dependent trends

a) Linear trends (days, months, years)

Time = 1, 2, 3, 4, …, N

*Where N is the total number of observations for each individual or site (including NAs)*



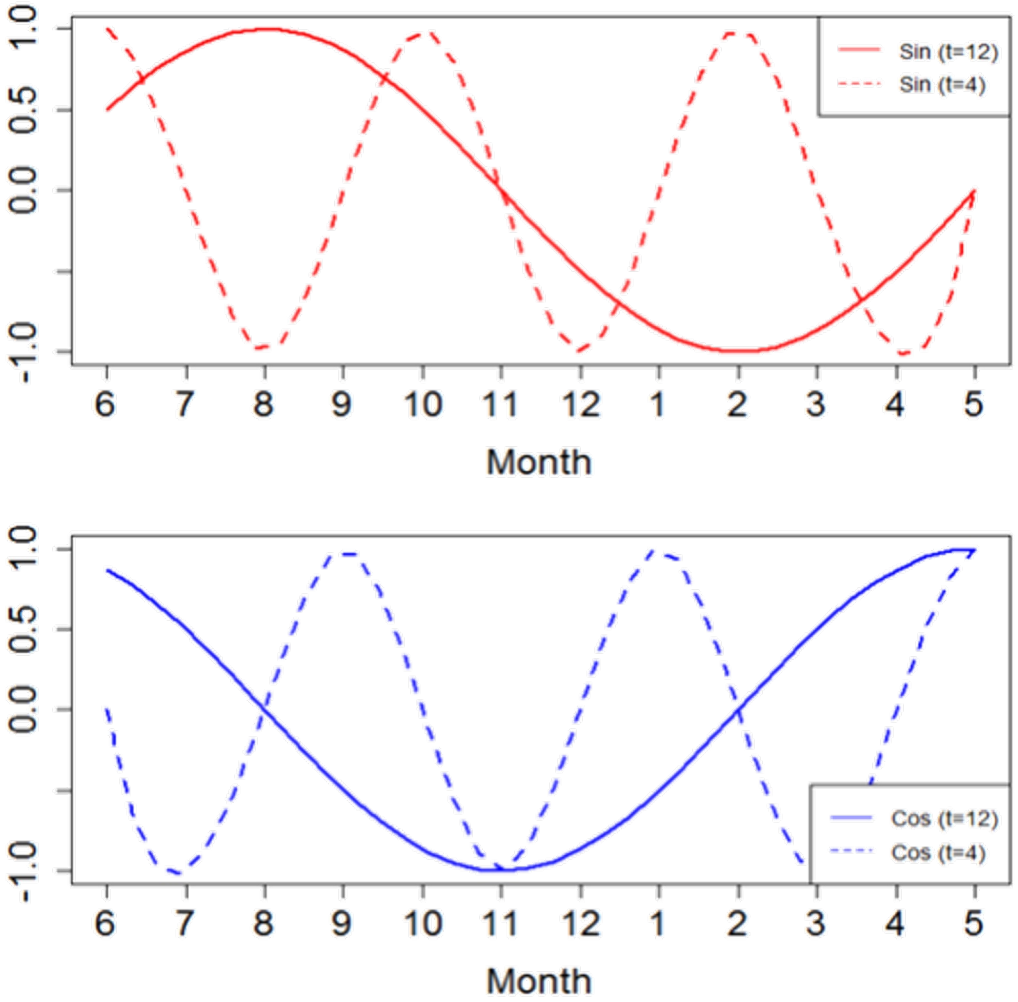
# Introducing time-dependent trends

a) Linear trends

b) Seasonal trends

Season=sin(2*pi*(month$_i$-shift)/period)

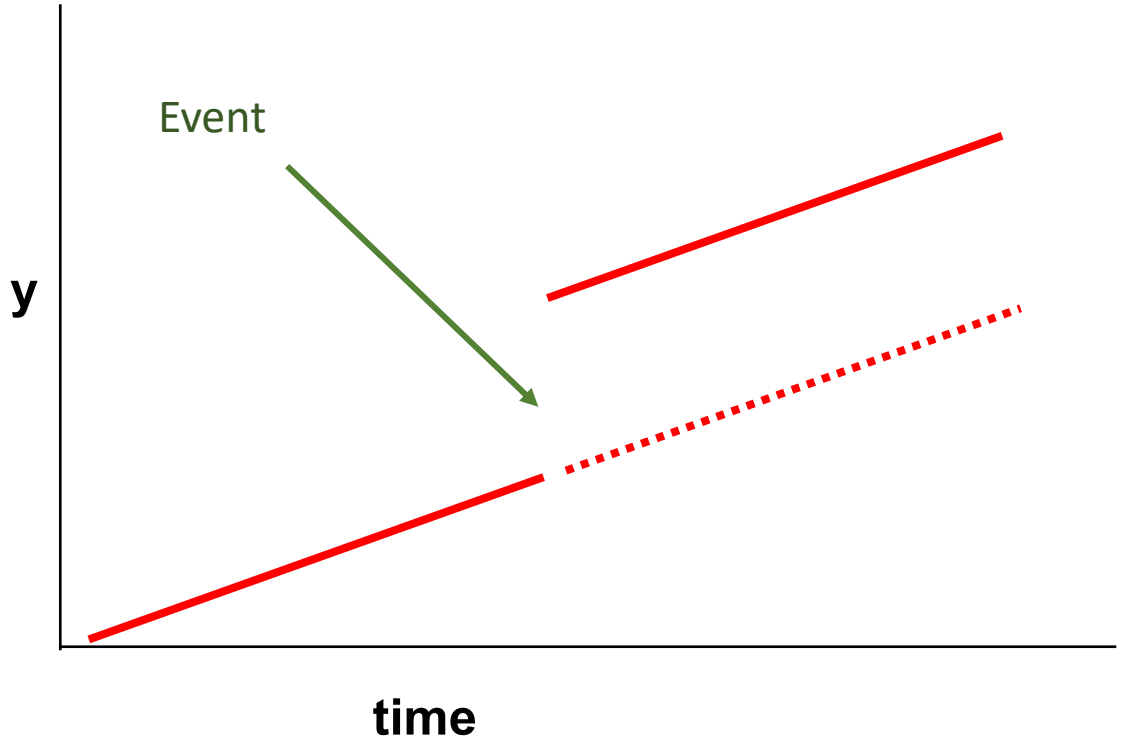
Season=cos(2*pi*(month$_i$-shift)/period)



# Evaluating abrupt and progressive changes over time

a) Immediate impact

Impact =
- 0 before the event happened
- 1 after the event happened



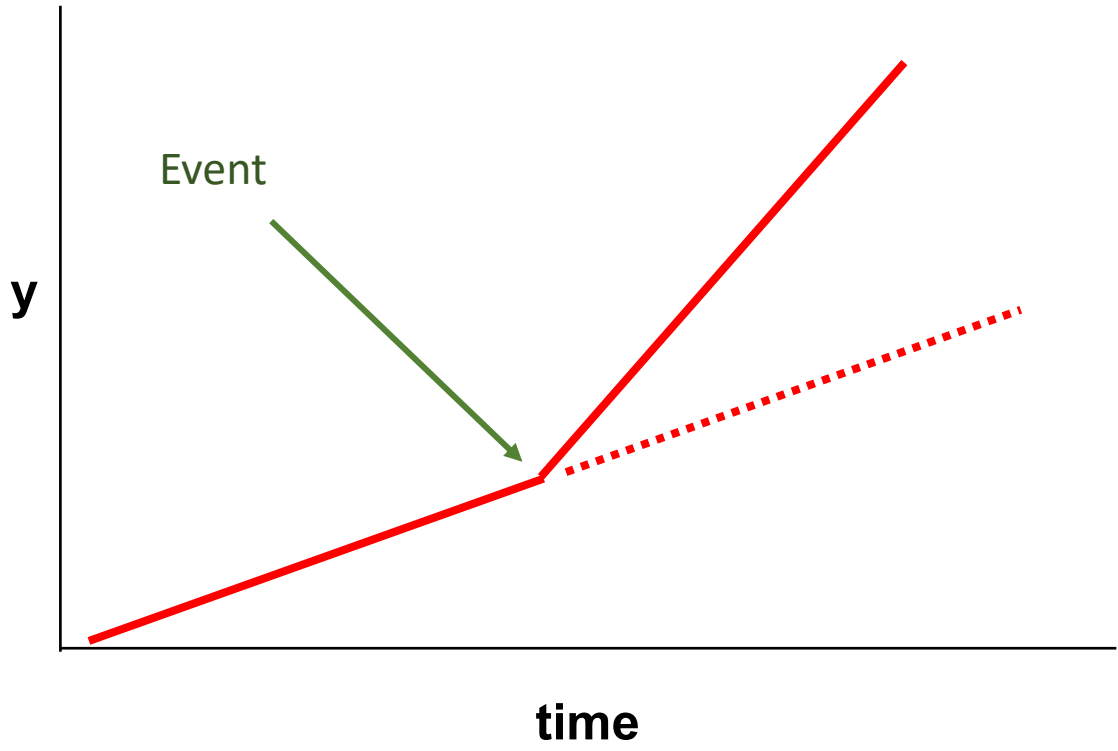
# Evaluating abrupt and progressive changes over time

a) Immediate impact

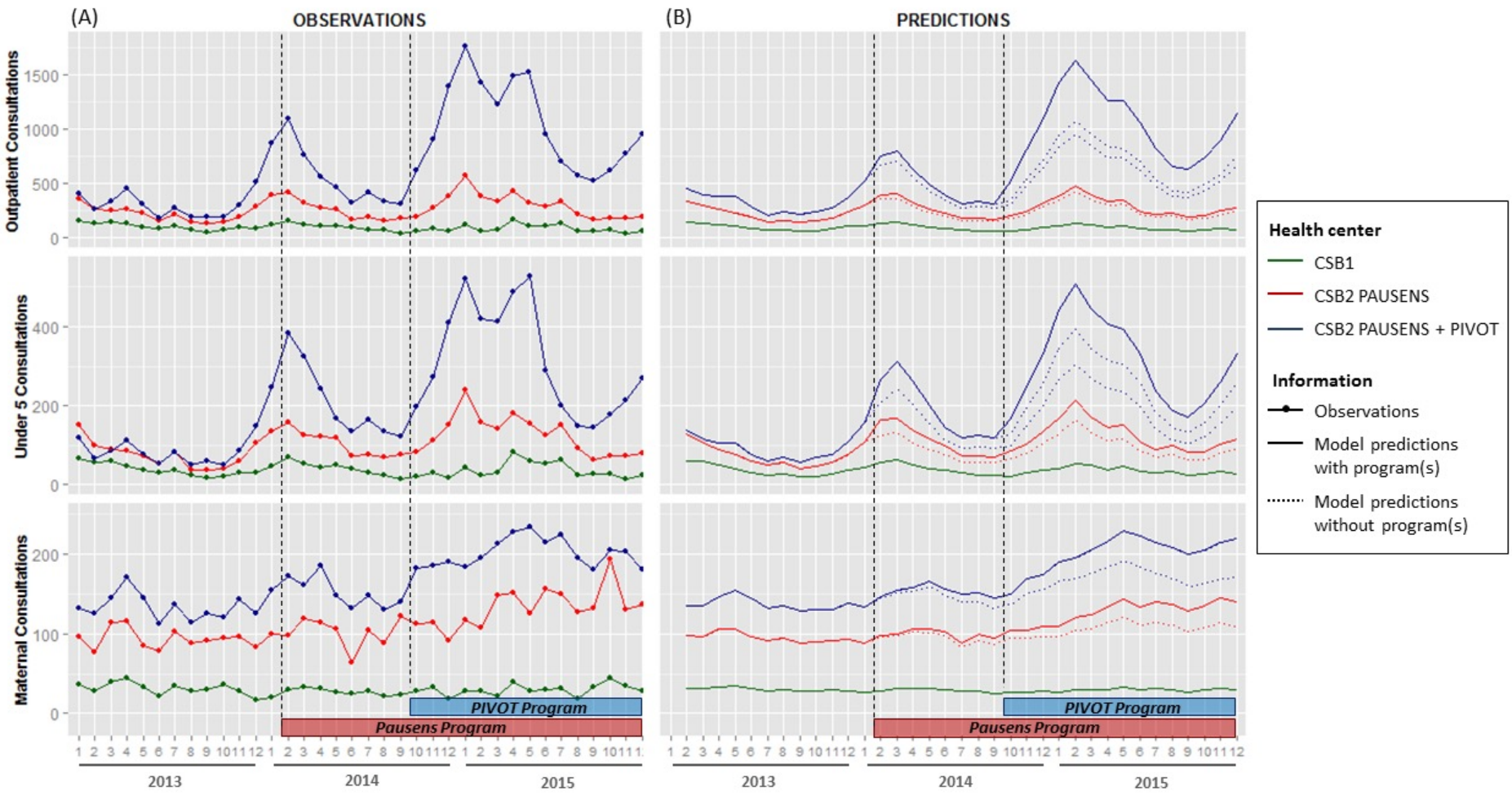
b) Progressive impact

Impact =
- 0 before the event happened
- 1, 2, 3, 4, …, N

after the event happened



# Back to the initial example



# The use of Generalized Linear Mixed Models for the study of dynamical systems

Andrés Garchitorena

Researcher, Institut de Recherche pour le Développement

Research Advisor, PIVOT Madagascar

*$E^2M^2$ Workshop*
*Ranomafana, January 2019*