Introduction to Linear Regression

Zoologie
et
Biodiversité Animale
Latimeria chalumnae

Hafaliana Christian Ranaivoson

Ecology and Evolution (E&E), University of Chicago

Mention Zoologie et Biodiversité Animale (MZBA), University of Antananarivo

Association Ekipa Fanihy (Efa), Madaggascar.

Adapted from Andrés Garchitorena, PhD
(E2M2, 2024)

E^2M^2 Workshop
Andasibe, May 2025

Objectives of the lecture

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

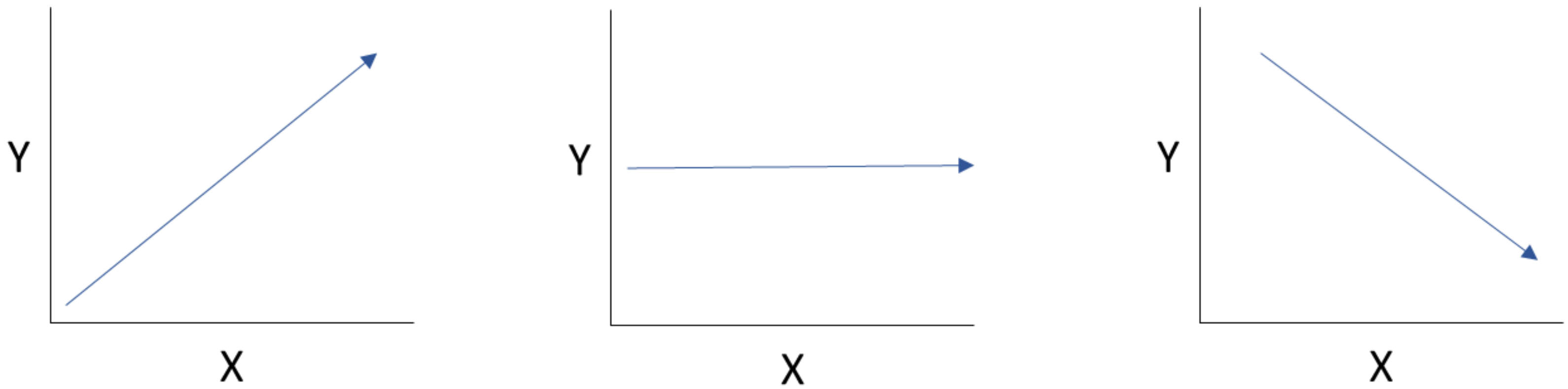
Being able to use a tool properly

‘ We are introducing to what is “available” (E2M2),
it is up to you to deepen what you think is “worth it” (C4C)’

SOME BASICS FIRST…

Linear Models

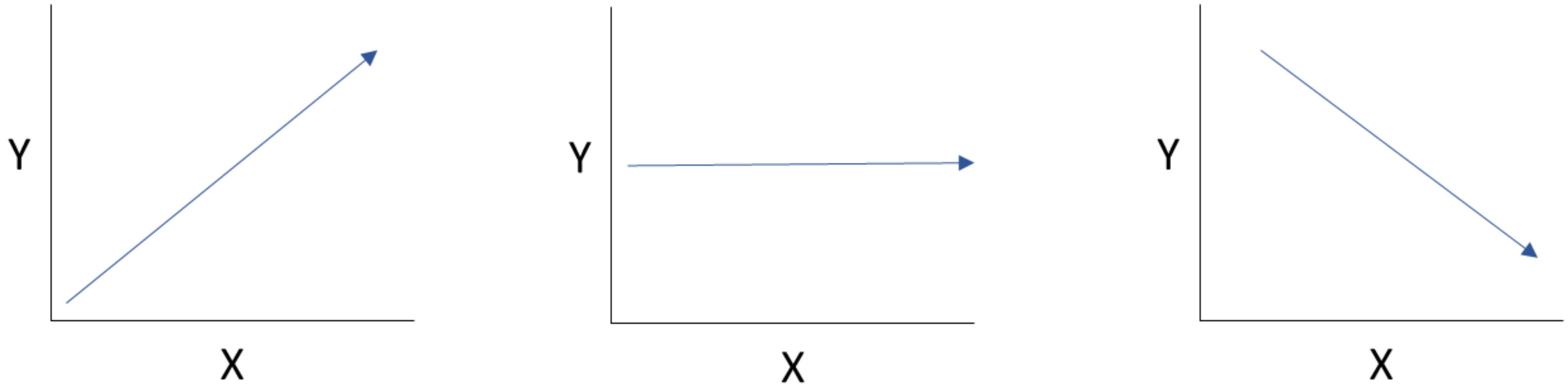
Think about a straight line to describe a trend



SOME BASICS FIRST…

Linear Models

Think about a straight line to describe a trend

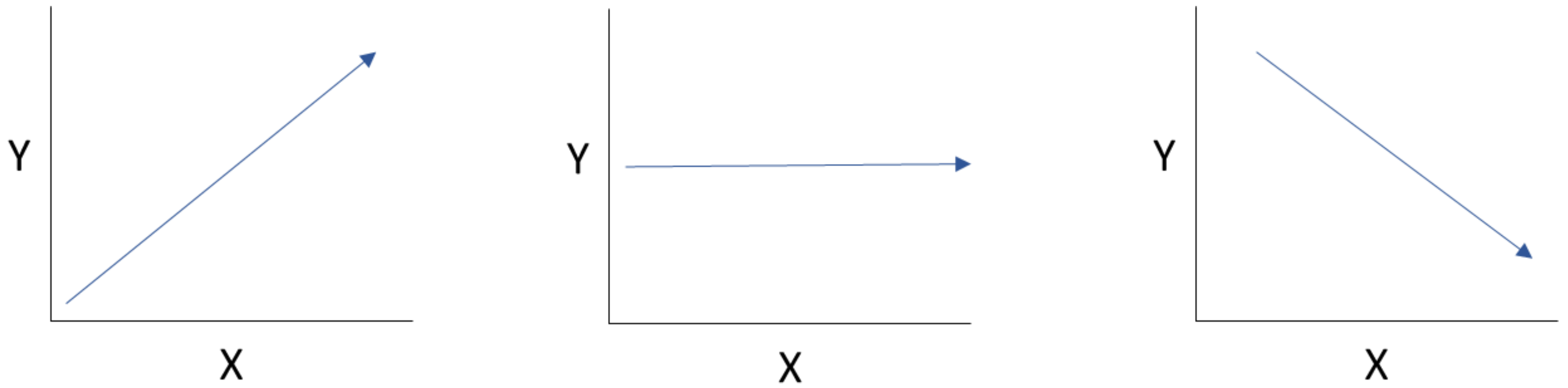


$$y = f(x)$$

SOME BASICS FIRST…

Linear Models

Think about a straight line to describe a trend



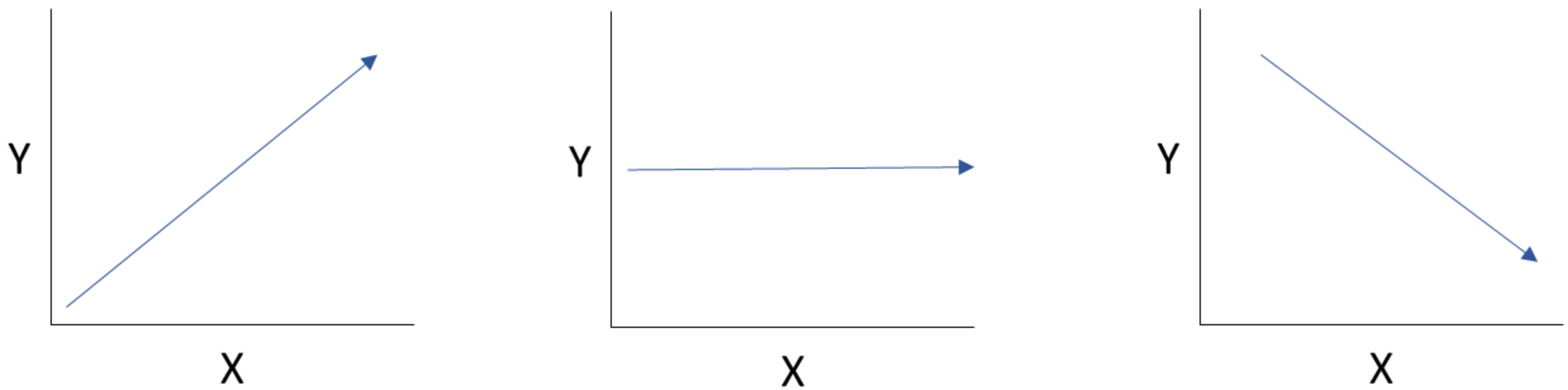
$$y = f(x)$$

Explanatory variable
Independent variable
(Variable explicative, axe des abscisses)

SOME BASICS FIRST…

Linear Models

Think about a straight line to describe a trend



$$y = f(x)$$

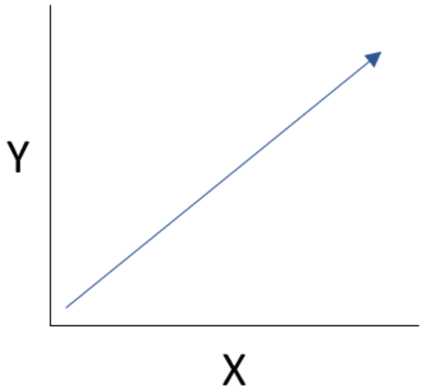
Response variable
Dependent variable
(Variable à expliquer, axe des ordonnées)

Explanatory variable
Independent variable
(Variable explicative, axe des abscisses)

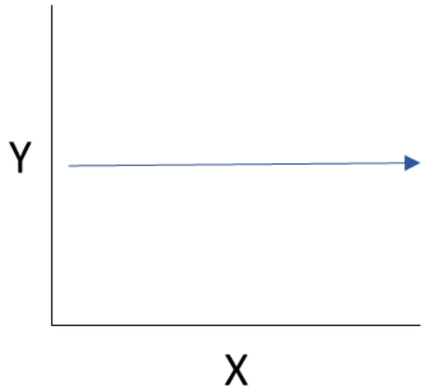
SOME BASICS FIRST…

Linear Models

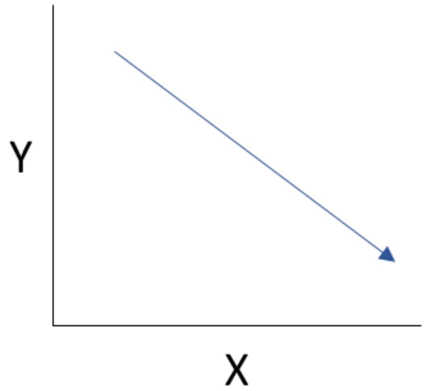
Think about a straight line to describe a trend



Going up (Miakatra)



Steady (Tsy miova)

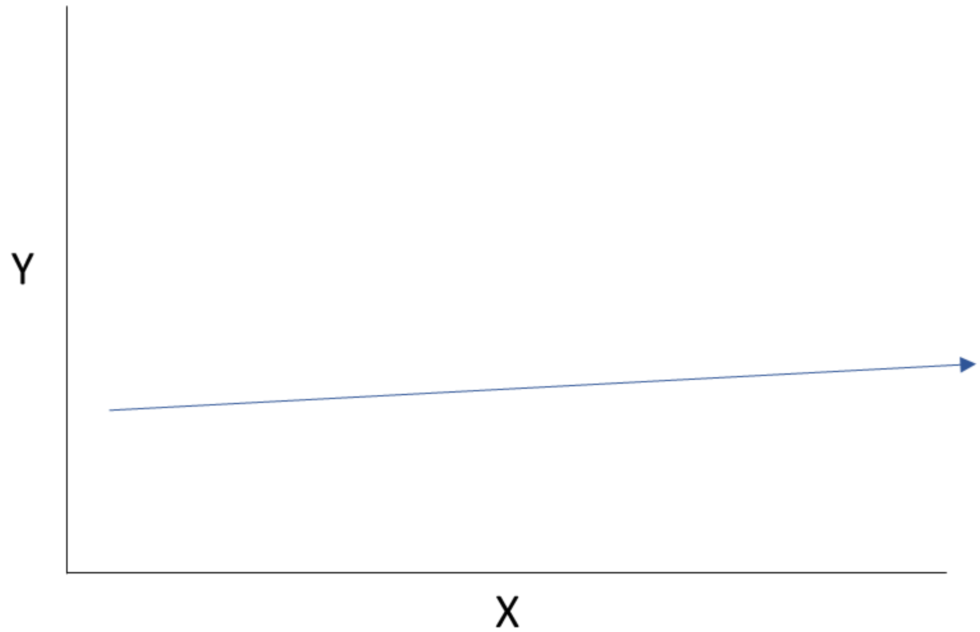


Going down (Mihidina)

SOME BASICS FIRST…

Linear Models

What about this one? (Dia ahoana izany ity?)



$$\mathcal{Y} = f(x)$$

SOME BASICS FIRST…

1. Univariate Linear Models

lm()

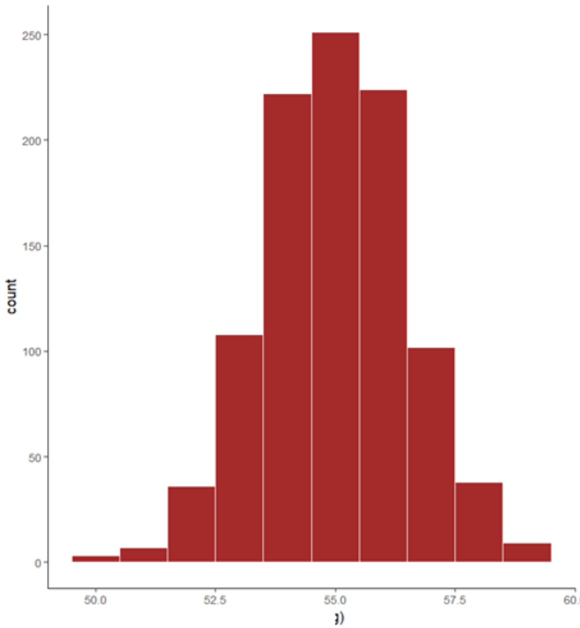
Considering the effect of **only one explanatory variable (x)** on a response variable **(y) with a normal distribution**

(Etudier « l’effet » une seule variable explicative sur une variable réponse ayant une distribution normale)

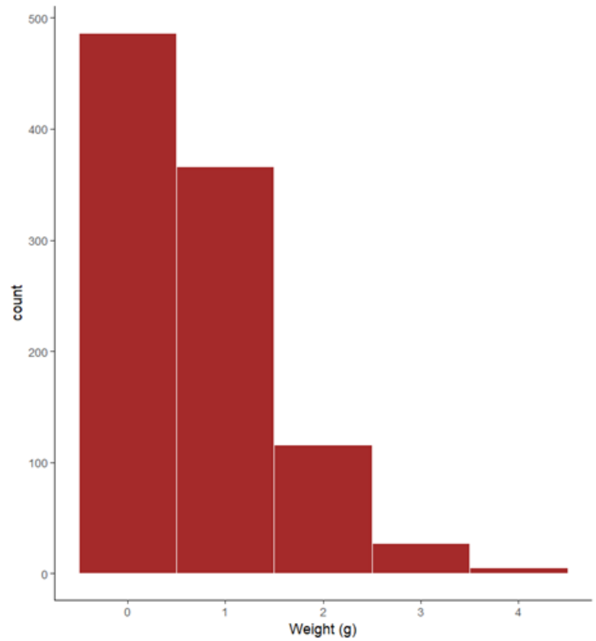
(Azo tsotsorina hoe fiakinan’ny “y” amin’ny “x” iray ihany)

SOME BASICS FIRST…

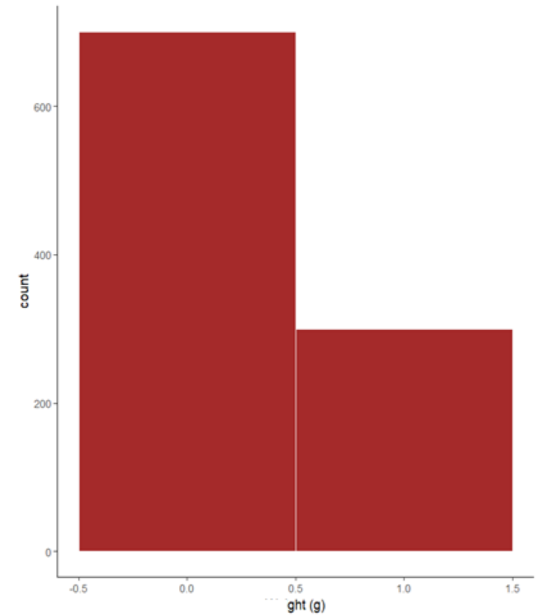
Variables and distributions



Var_

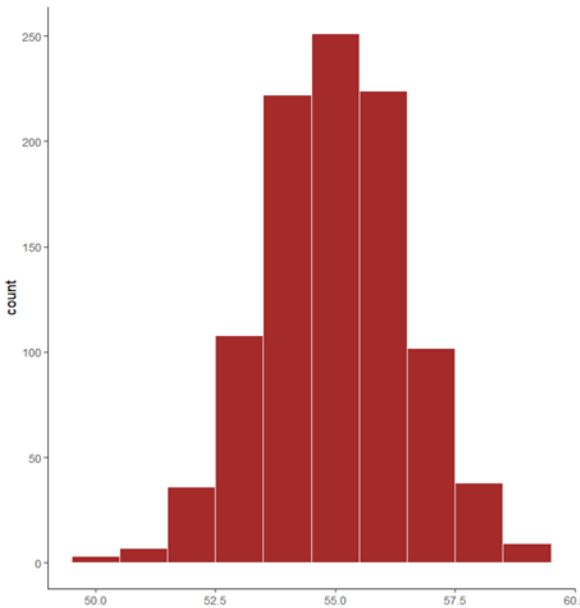


Var_

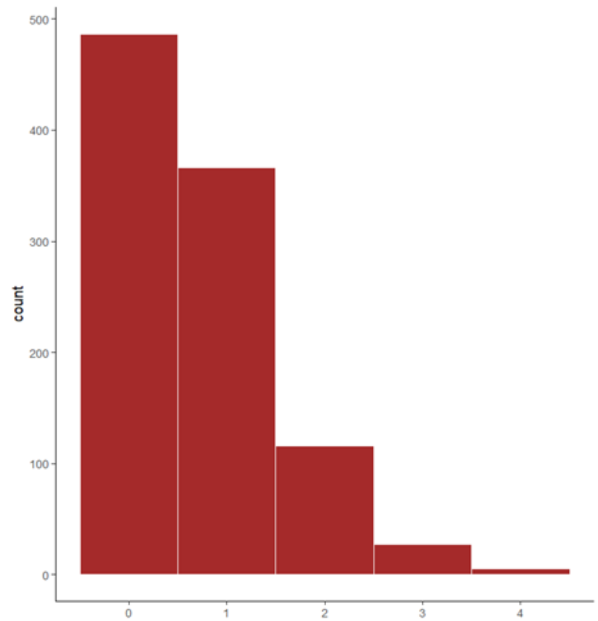


Var_

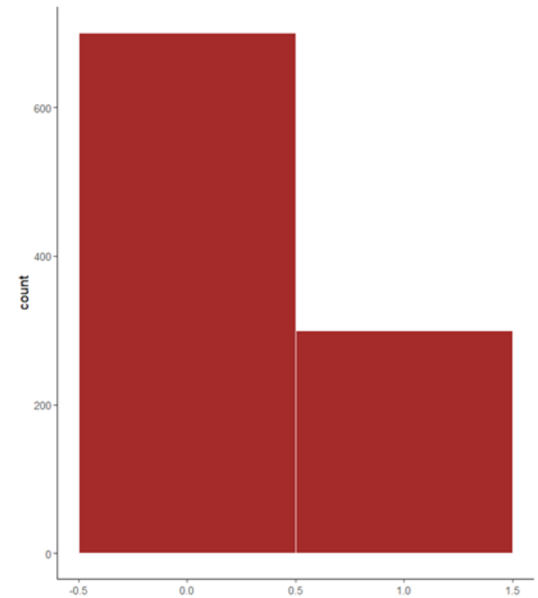
Variables and distributions



Var_norm

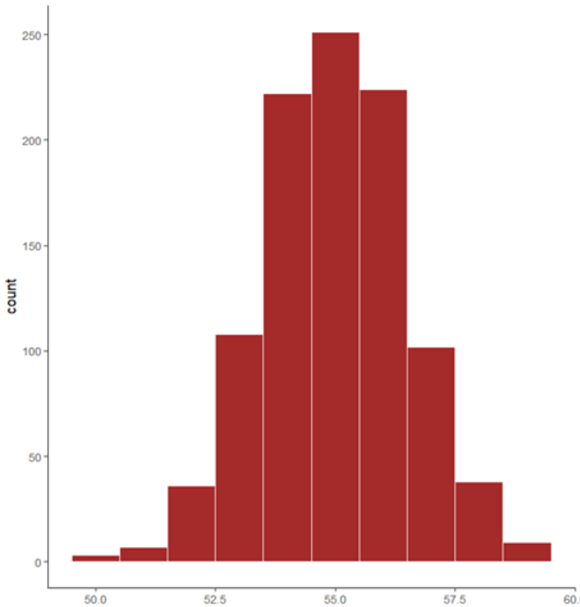


Var_poisson

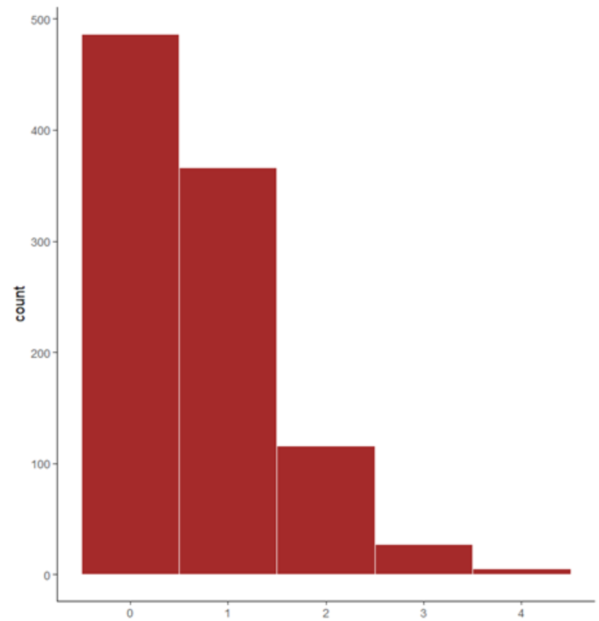


Var_binom

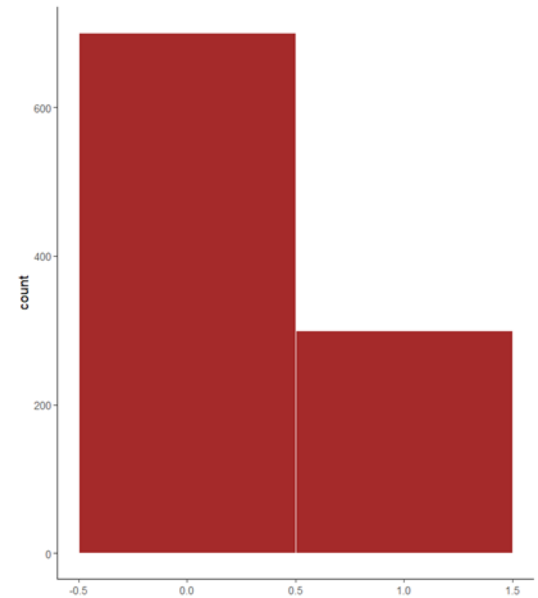
Variables and distributions



Var_norm



Var_poisson



Var_binom

Continuous

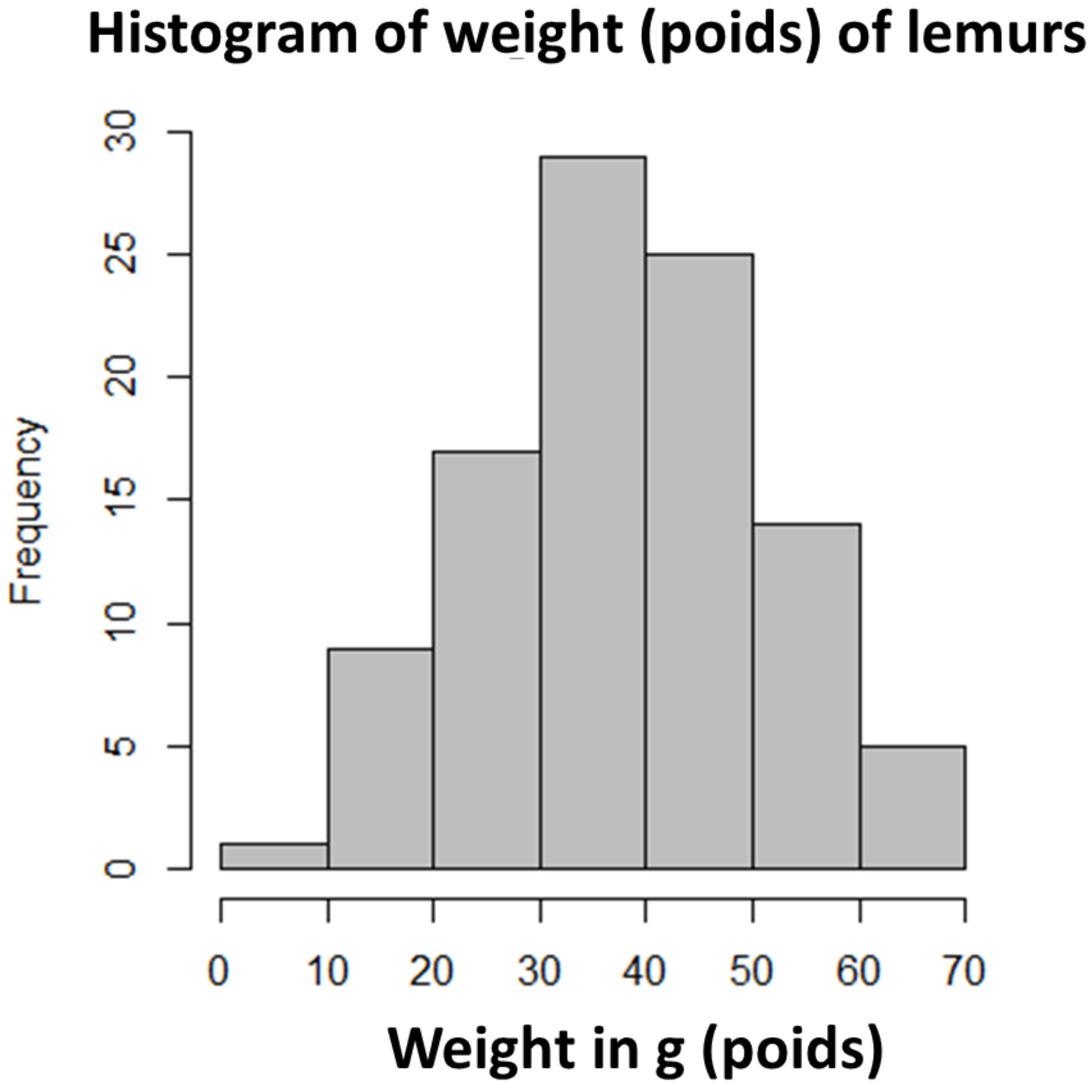
Integer (Count)

Presence/Absence

Let’s work through a cute example

Lemurs weight (poids) determinants

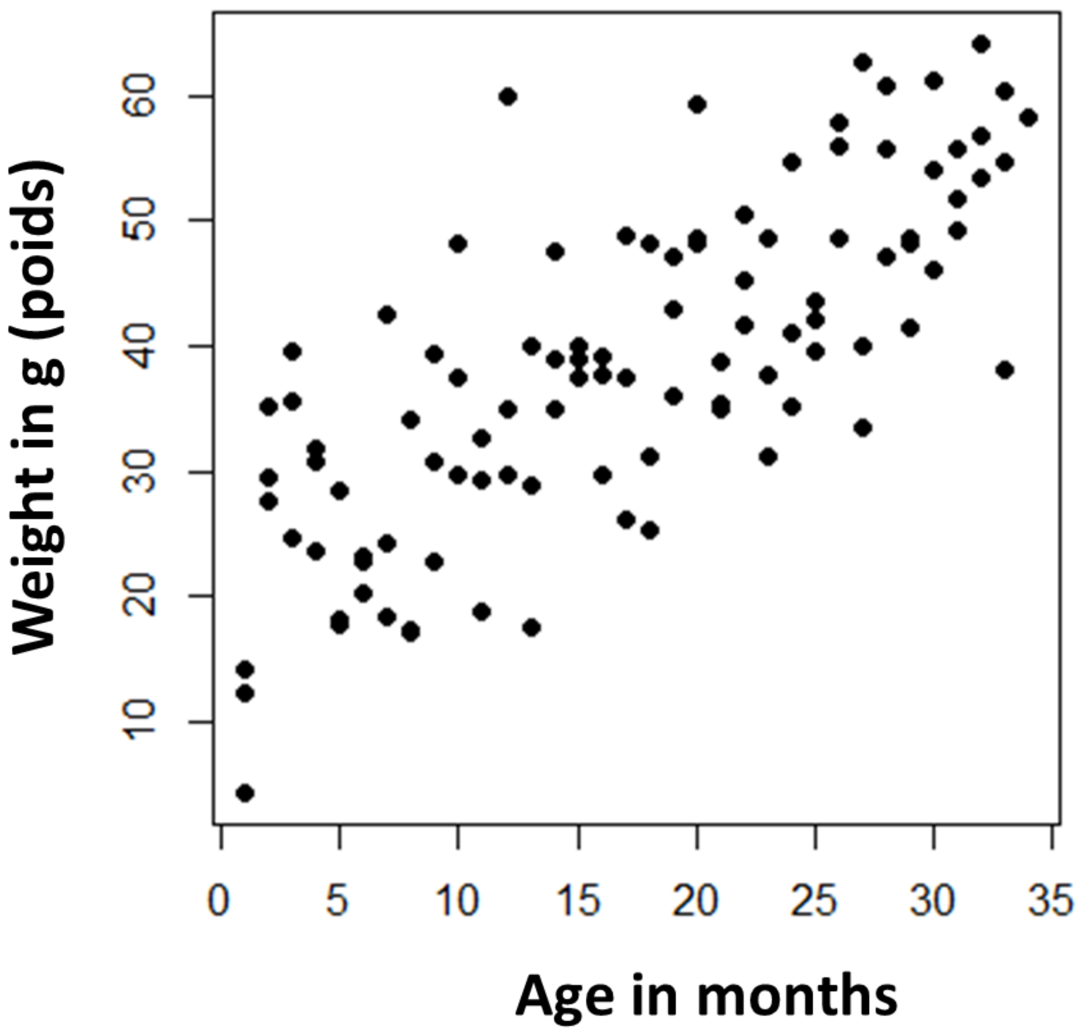
Lemur weight and determinants



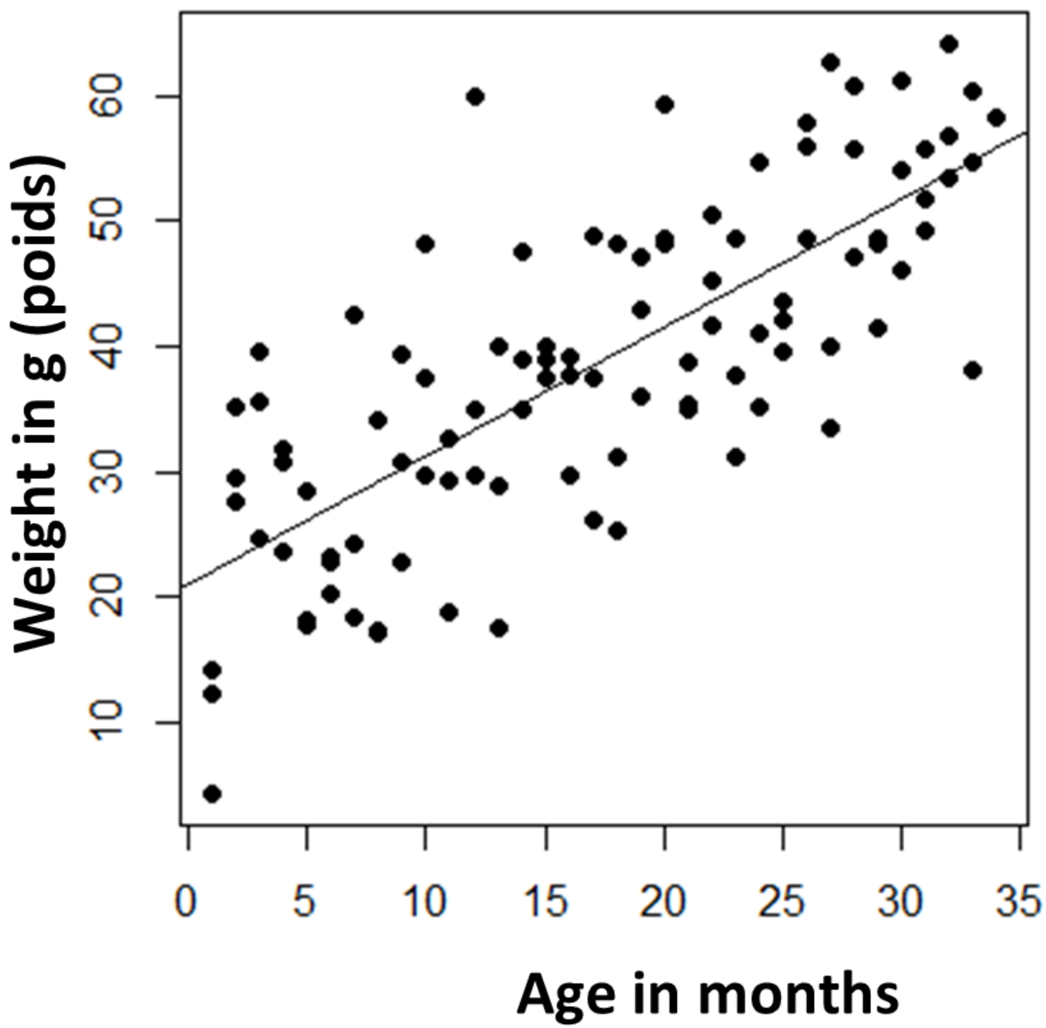
$$Zi = \frac{Xi - \bar{X}}{\sigma}$$

Z~N(0, 1)

Lemur weight and determinants

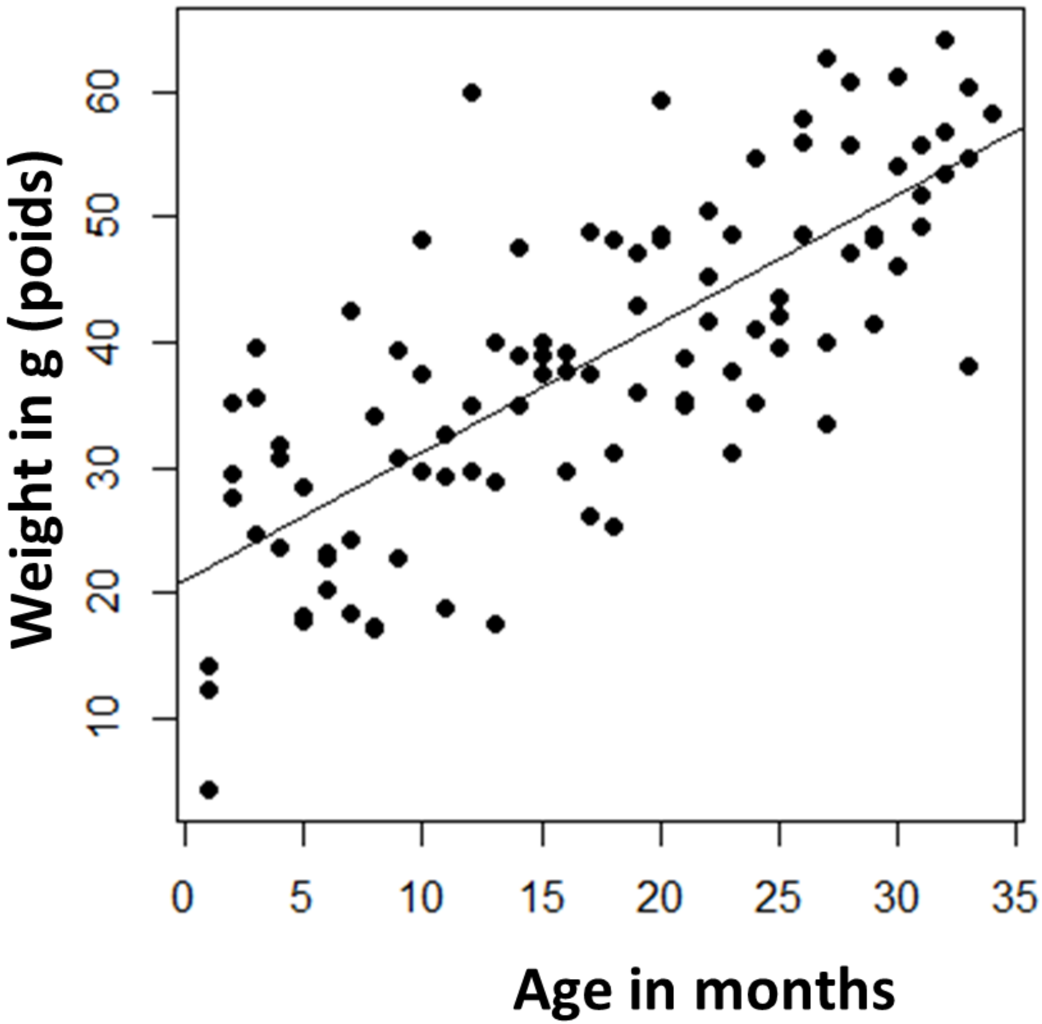


Lemur weight and determinants



lm(formula, data)

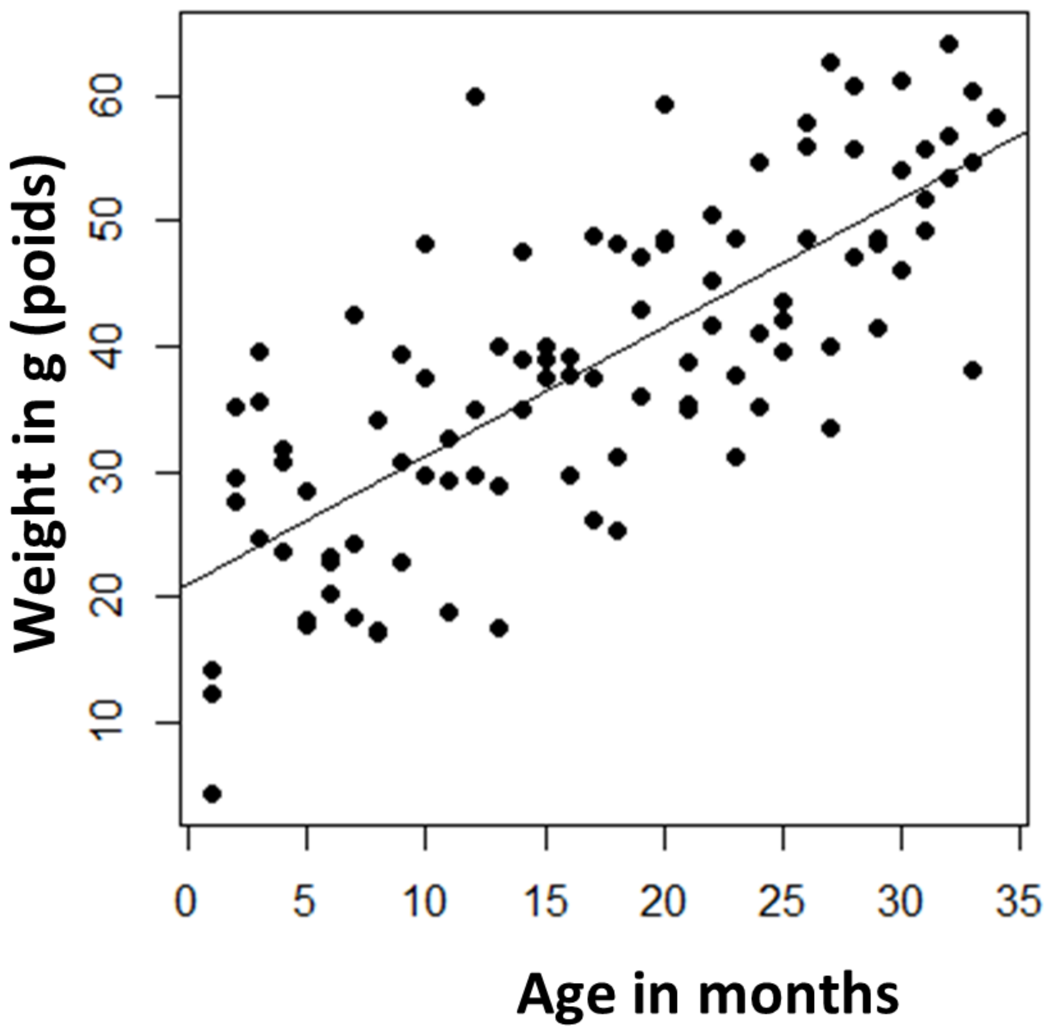
Lemur weight and determinants



lm(formula, data)

Formula : $\mathcal{Y} = f(x)$

Lemur weight and determinants

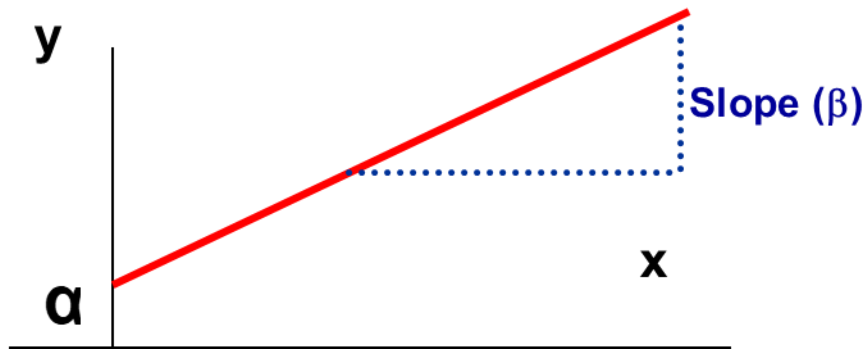


lm(weight ~ Age, data)

“lm(formula, data)”

Simple linear regression

- Relation between 2 continuous variables



- *Intercept (α)*
  - Value of y when x is 0
- *Regression coefficient β_1*
  - Measures association between y and x
  - Amount by which y changes on average when x changes by one unit
- *Error (ε)*
  - Difference between the predicted values and observed values of y

Simple linear regression

$$\boxed{f(x) = \alpha + \beta x + \varepsilon}$$

Response variable = Systematic component + Residual component

Systematic component: Intercept and explanatory variables

Residual component:

- Null mean
- Independence
- Fixed variance
- Normality

$$\boxed{y = \alpha + \beta x + \varepsilon}$$

The R function to fit a linear model is lm() which uses the form

fitted.model <- lm(formula, data=data.frame)

Simple linear regression

$$\boxed{f(x) = \alpha + \beta x + \varepsilon}$$

Response variable = Systematic component + Residual component

Systematic component → Intercept and explanatory variables

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

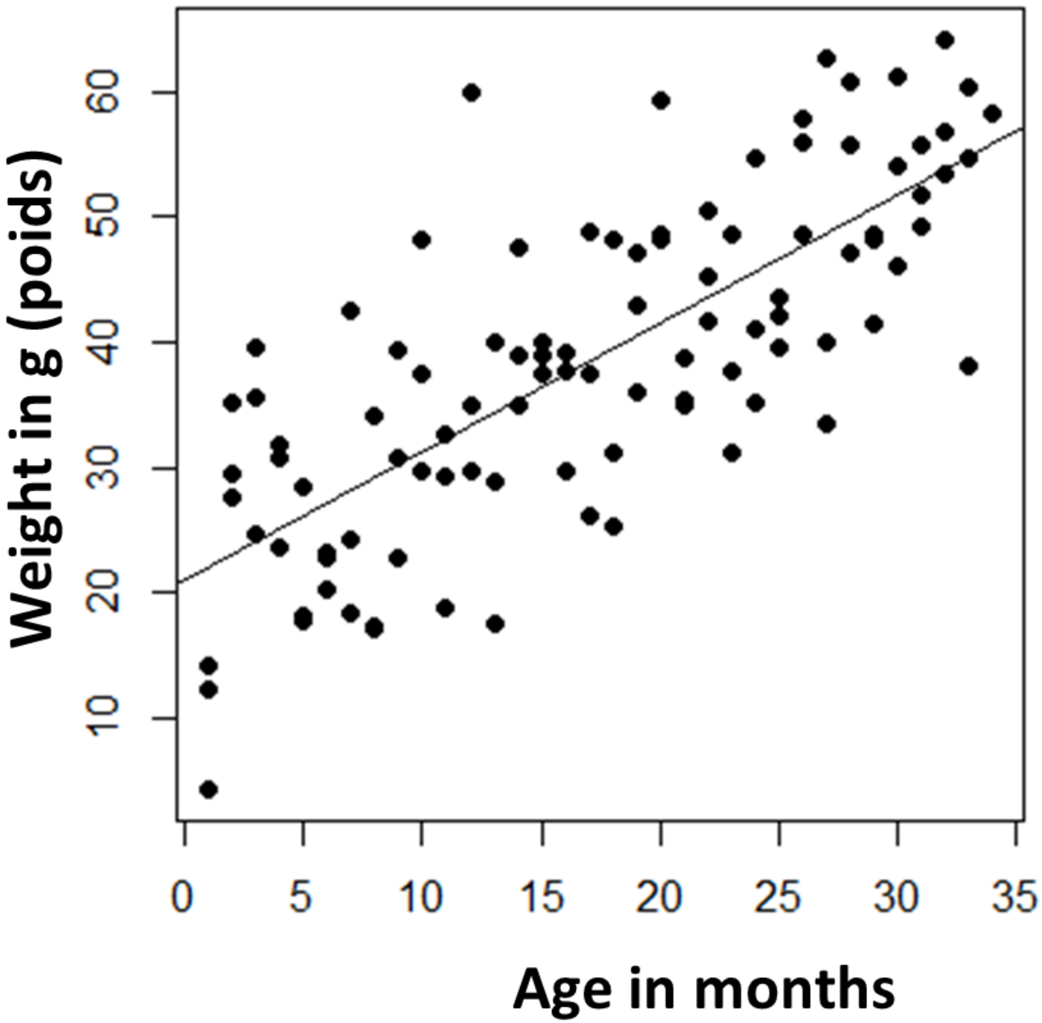
$$\boxed{y = \alpha + \beta x + \varepsilon}$$

$$\boxed{\textbf{weight} = \alpha + \beta.\text{age} + \varepsilon}$$

The R function to fit a linear model is lm() which uses the form

fitted.model <- lm(formula, data=data.frame)

Lemur weight and determinants

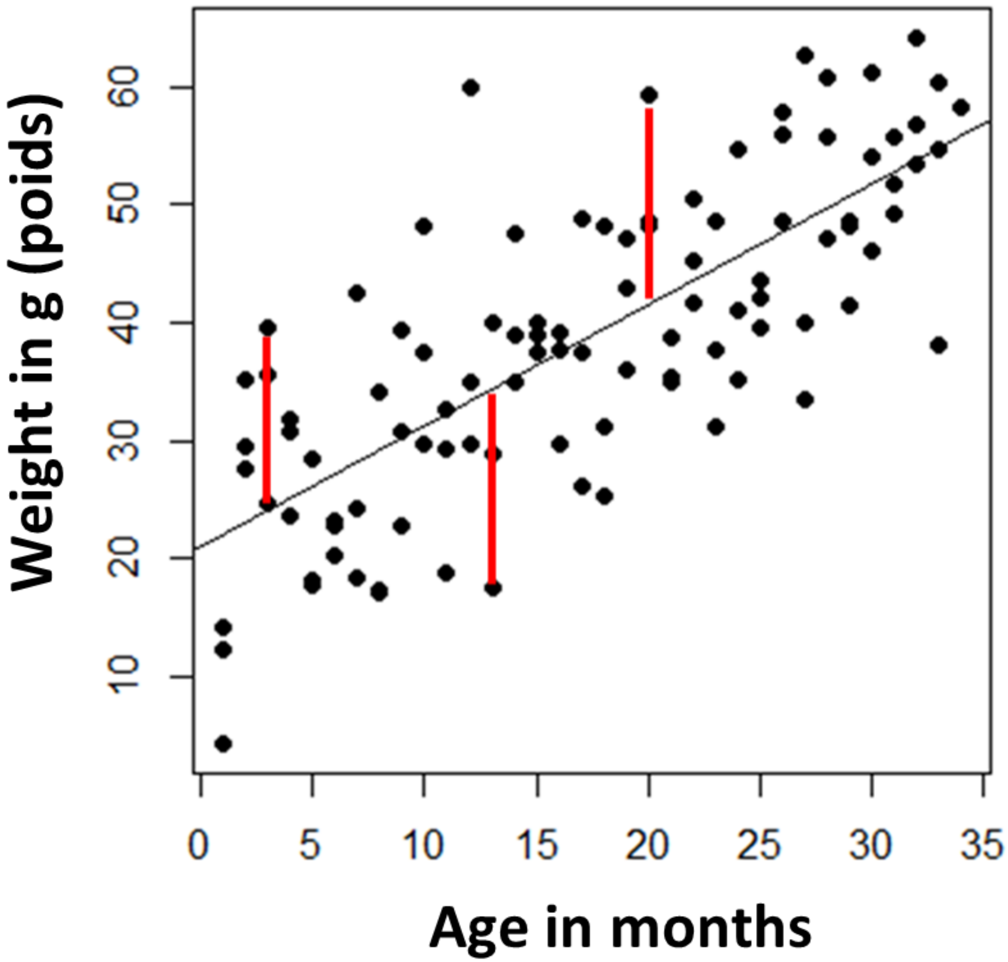


lm(weight ~ Age, data)

Formula : $y = f(x)$

$$\boxed{y = \alpha + \beta x + \varepsilon}$$

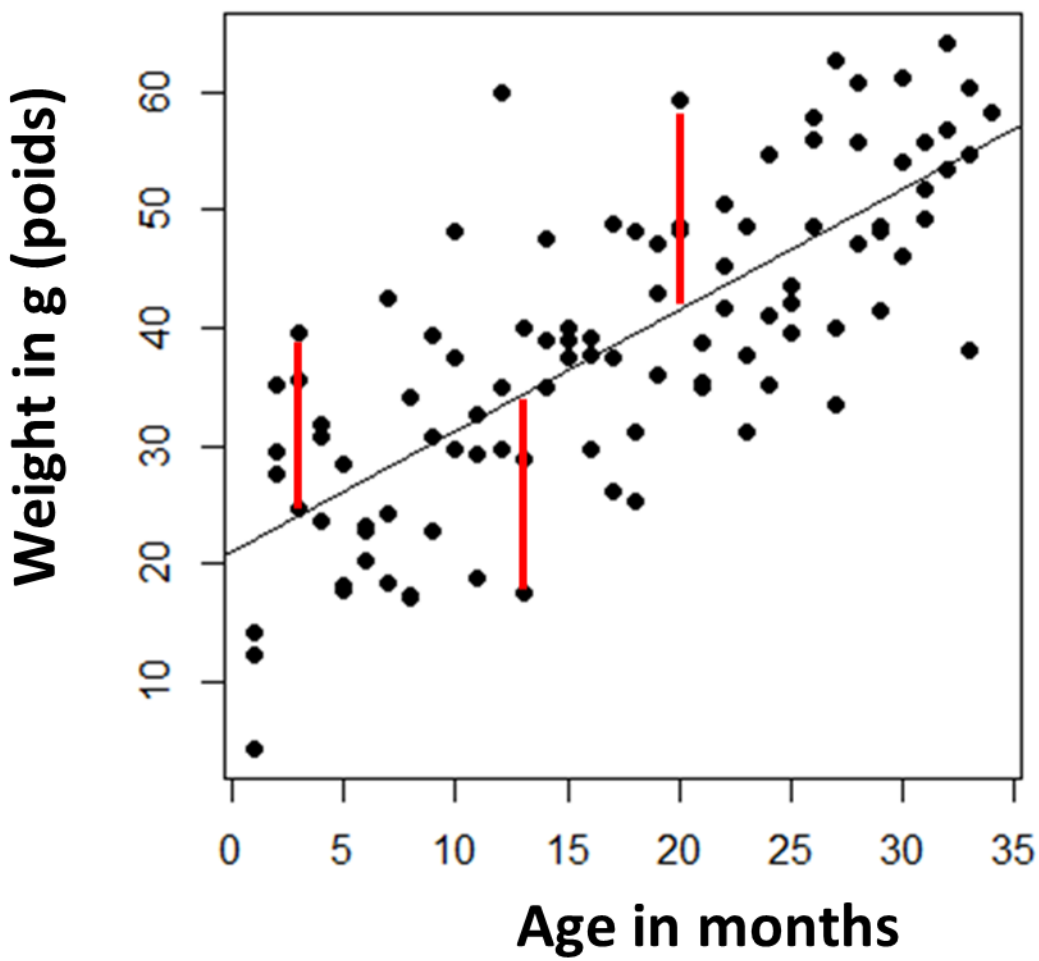
Lemur weight and determinants



$$\boxed{y = \alpha + \beta x + \varepsilon}$$

Lemur weight and determinants

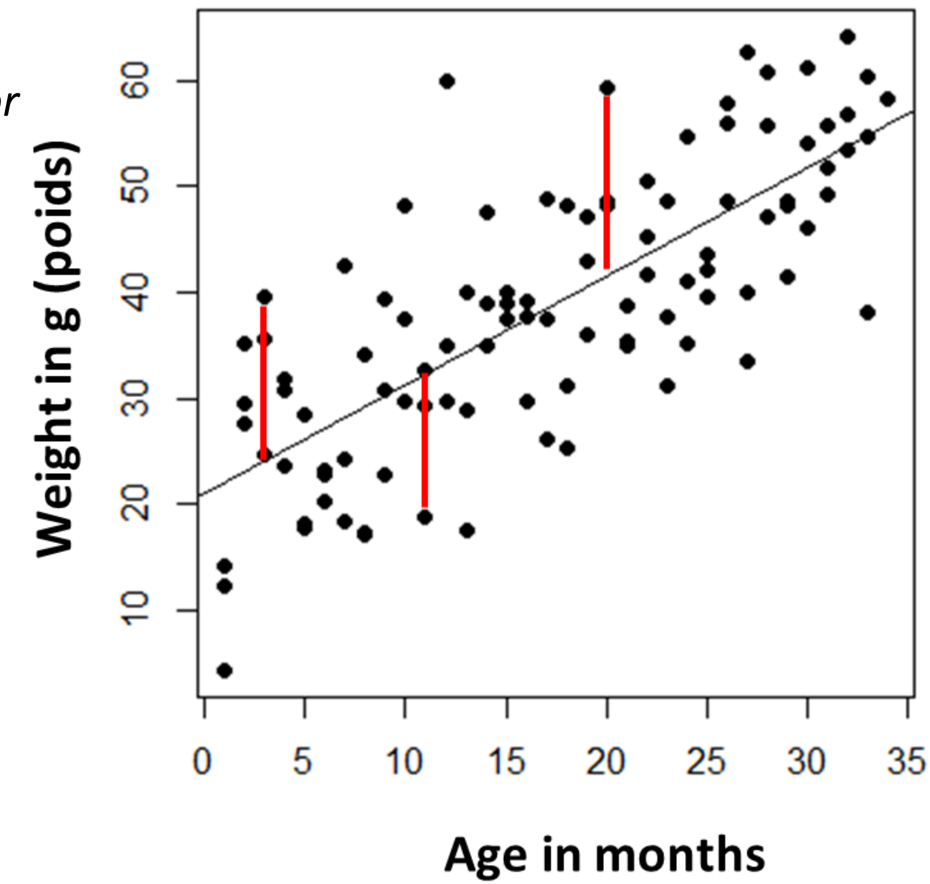
The goal is to minimize the difference between what we predict and what we observe
(Méthode des moindres carrés ry reto an!)



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

Simple linear regression

weight = 20 + 1.15 x Age (months) + Error



A process is generally the result of several others…

Mety hoe tsy ny mahabe ny taona ihany no mampavesatra ny lanja an!

1. Univariate Linear Models — lm()

↓

2. Multivariate Linear Models — lm()

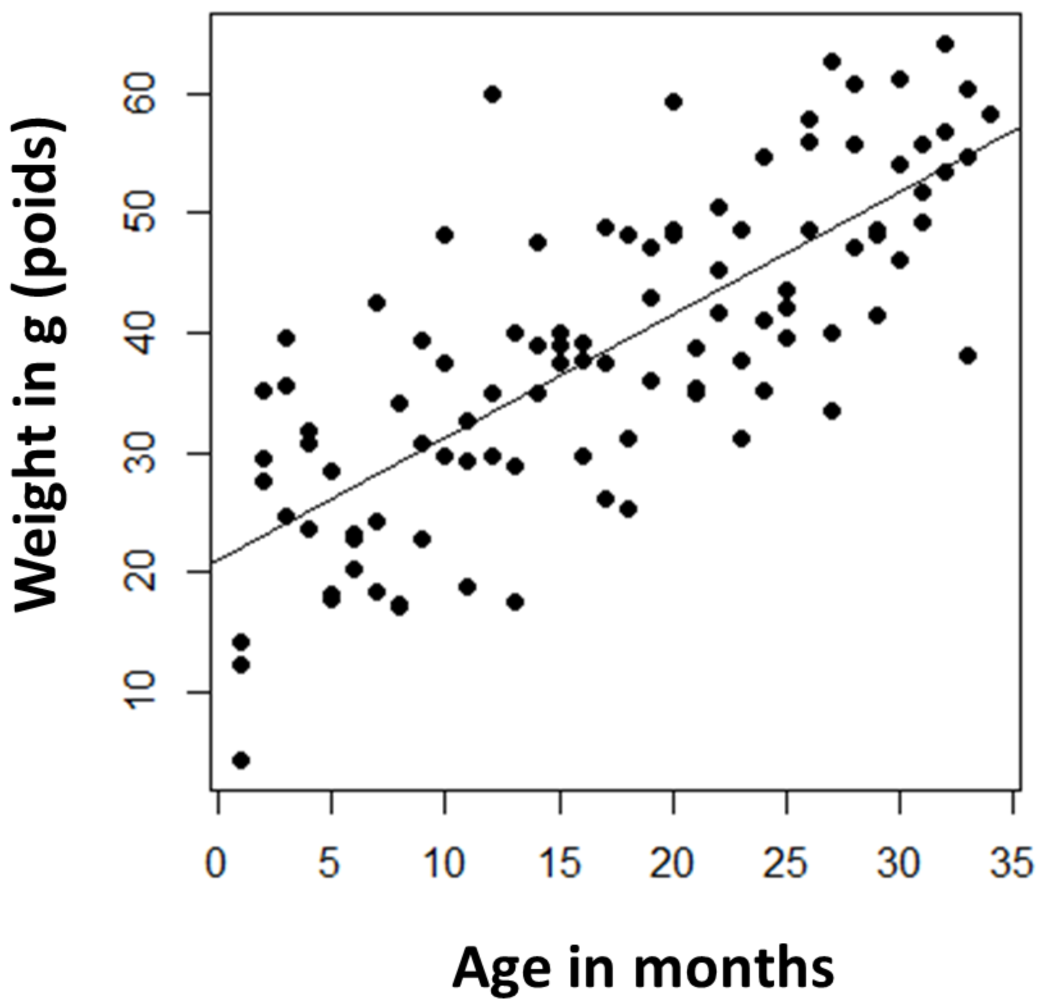
Considering the effect of **multiple explanatory variables (x_1, x_2…x_n) on** a response variable **(y) with a normal distribution**

(Etudier « l’effet » de plusieurs variables explicatives sur une variable réponse ayant une distribution normale)

(Azo tsotsorina hoe fiakinan’ny “y” amin’ny “x” maromaro)

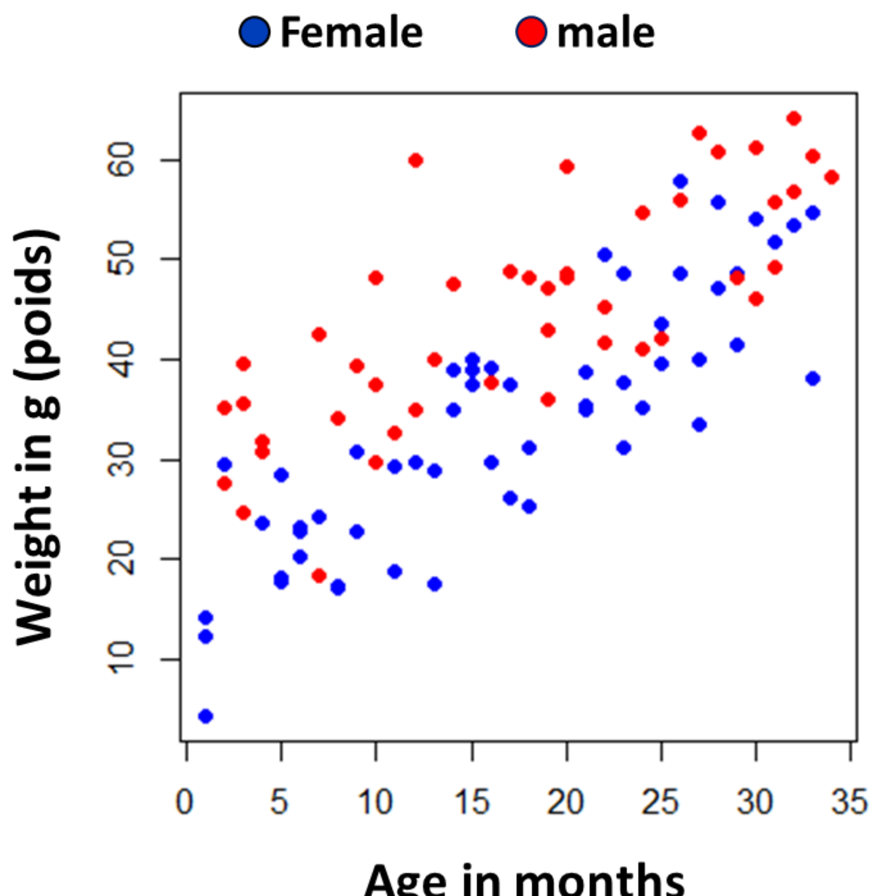
INTRODUCING MULTIVARIATE LINEAR MODELS

Lemur weight and determinants



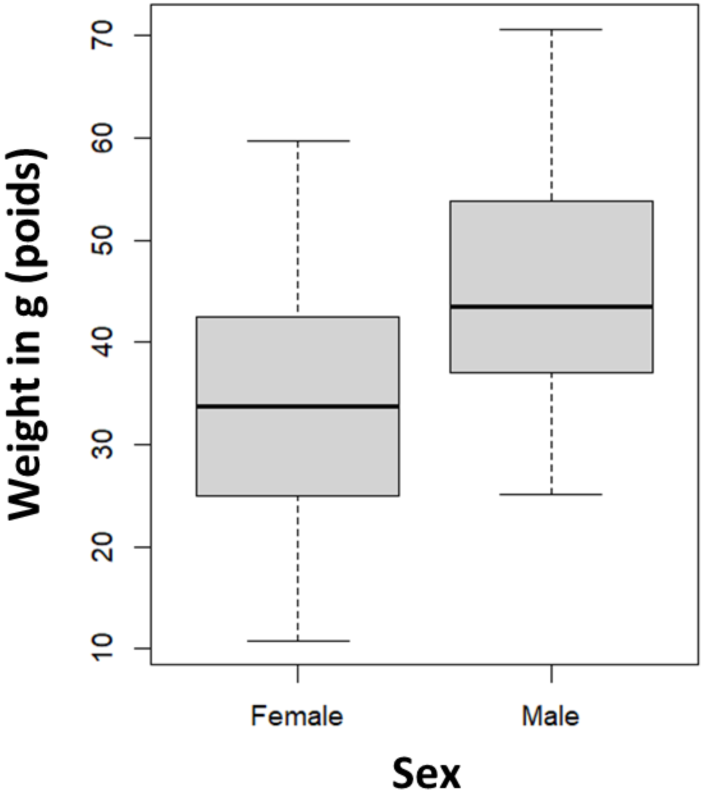
Lemur weight and determinants

The effect of gender



Lemur weight and determinants

The effect of gender

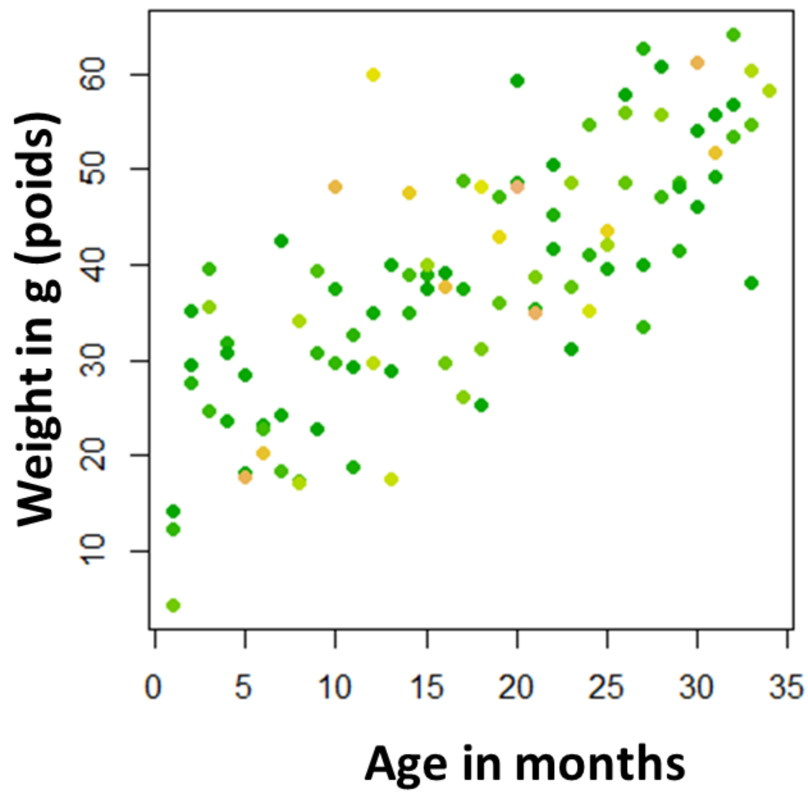


Taille = 15 + 1.15 x Age (months) + 15 x Sexe (male) + Error

Lemur weight and determinants

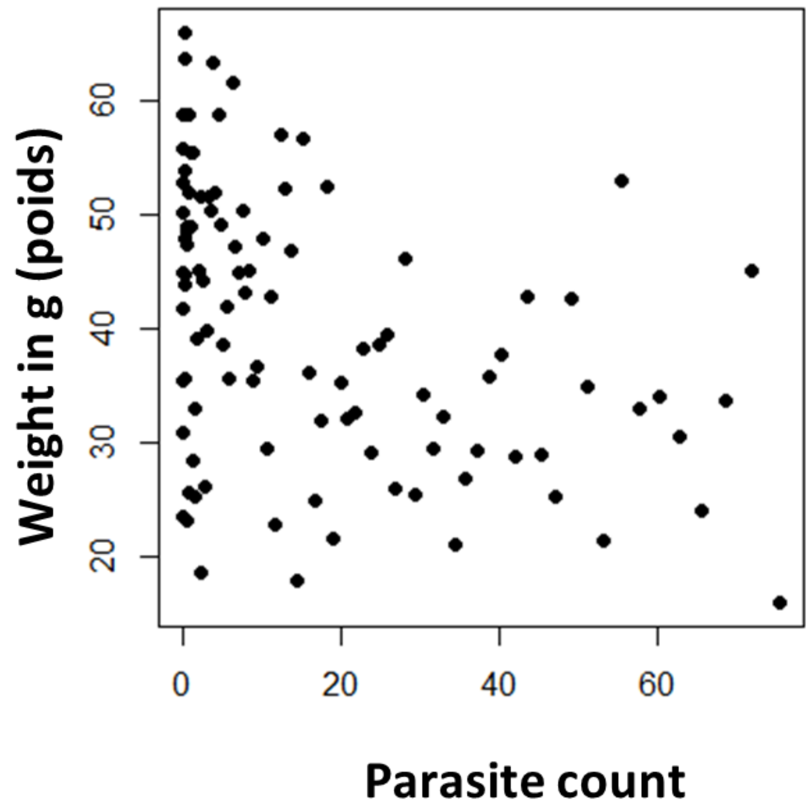
The effect of parasites

Green: low parasite burden
Yellow: high parasite burden



Lemur weight and determinants

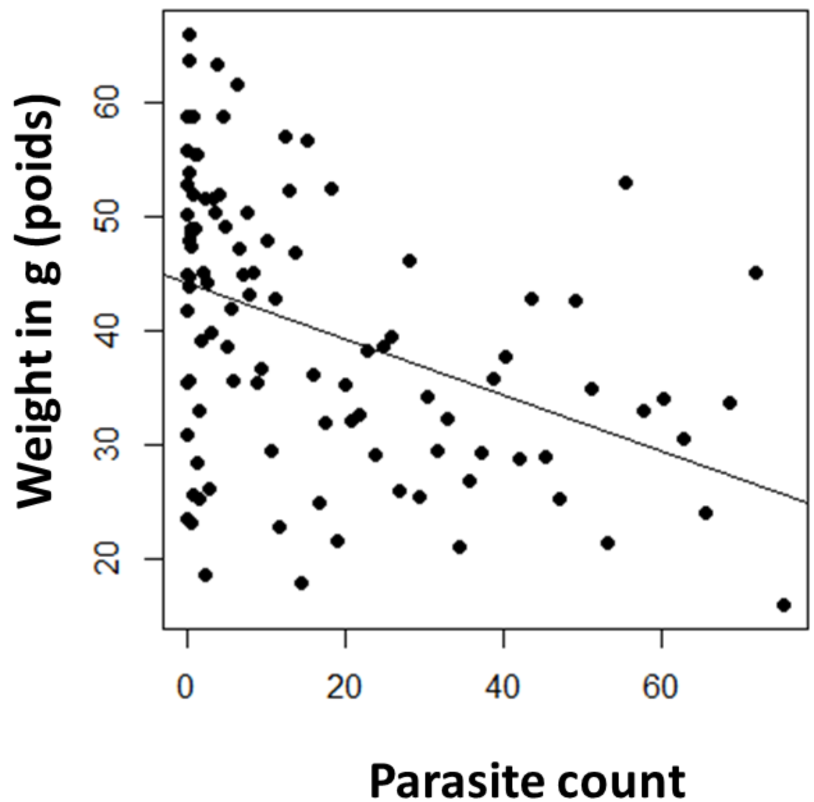
The effect of parasites



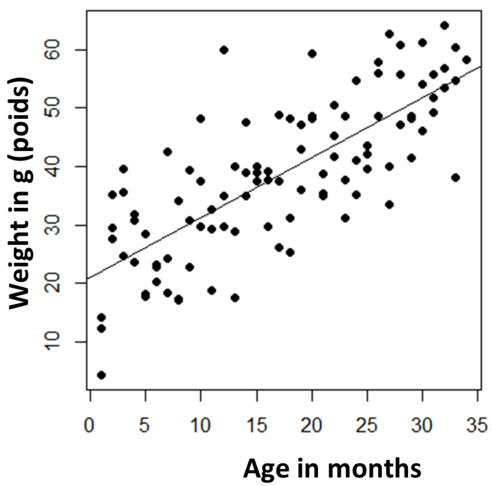
Lemur weight and determinants

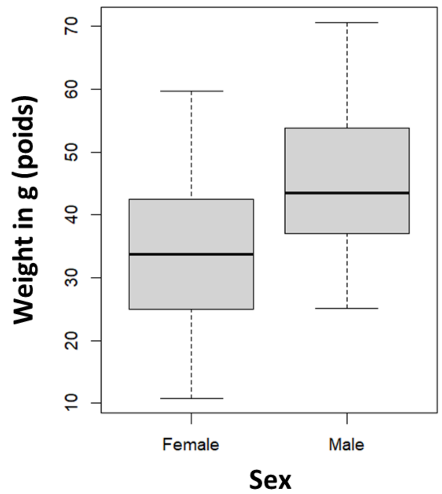
The effect of parasites

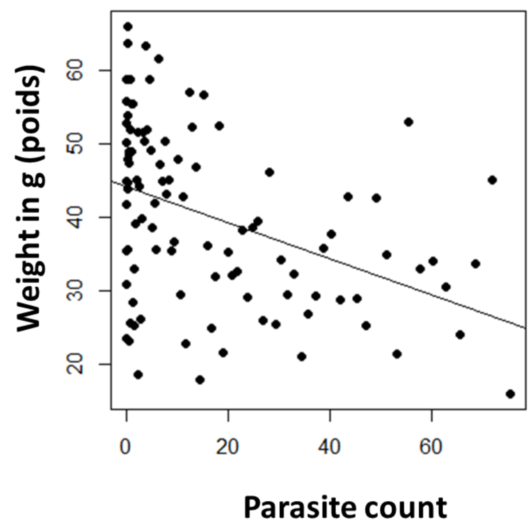
Taille = 45 - 0.3 x Nb Parasites + Error



Lemur weight and determinants







lm(taille~age+sexe+GIparasites, data)

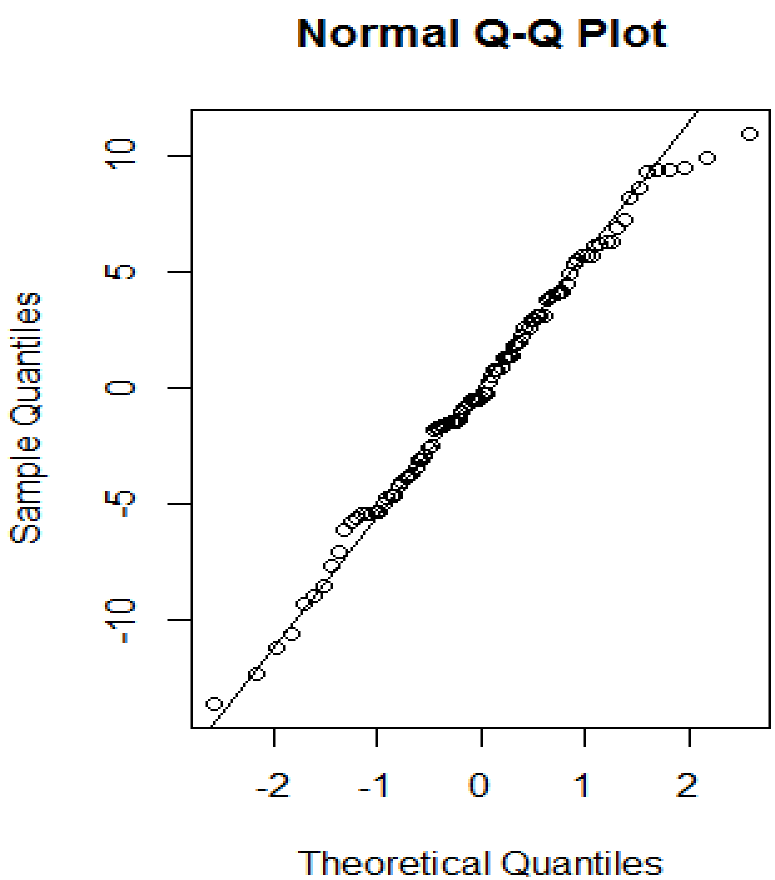
Multiple linear regression

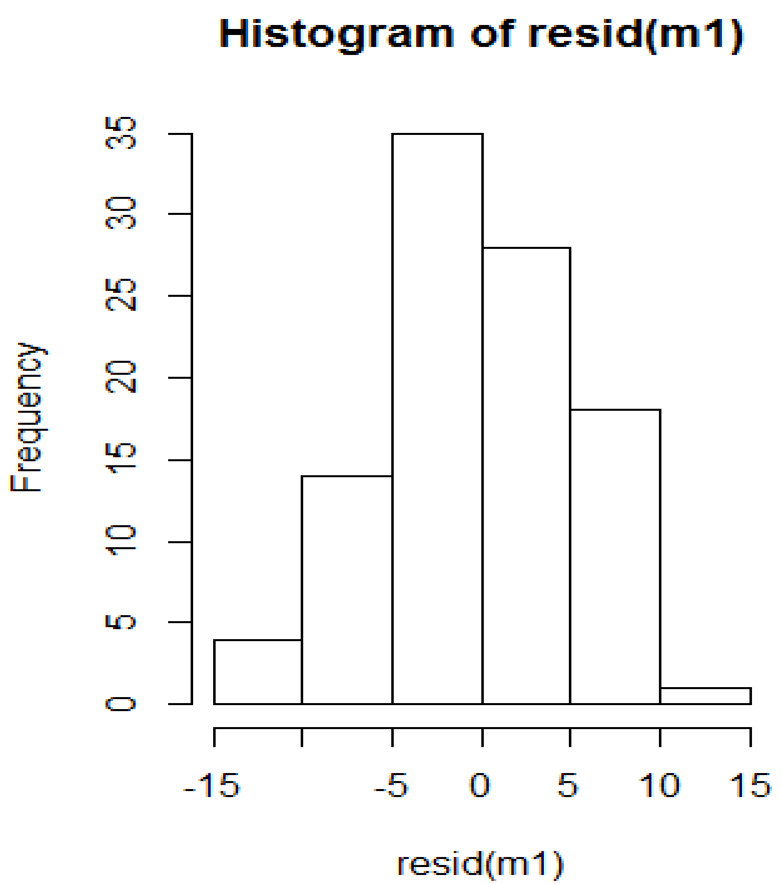
- Linear regression with multiple explanatory variables
- To describe the relationship between
  - The response variable, y
  - The explanatory variables, $x = (x_1, x_2, \ldots, x_n)$
- The model: $y = \alpha + \beta_1 * x_1 + \ldots + \beta_n * x_n + \varepsilon$
  with $\varepsilon \sim N(0, \sigma^2)$
- We generally select the model that best fits the data (best explains observed patterns) with the smallest number of variables

Model validation (case of linear regression y~N(0,1))

- Check that model assumptions have not been violated

Normality of residuals

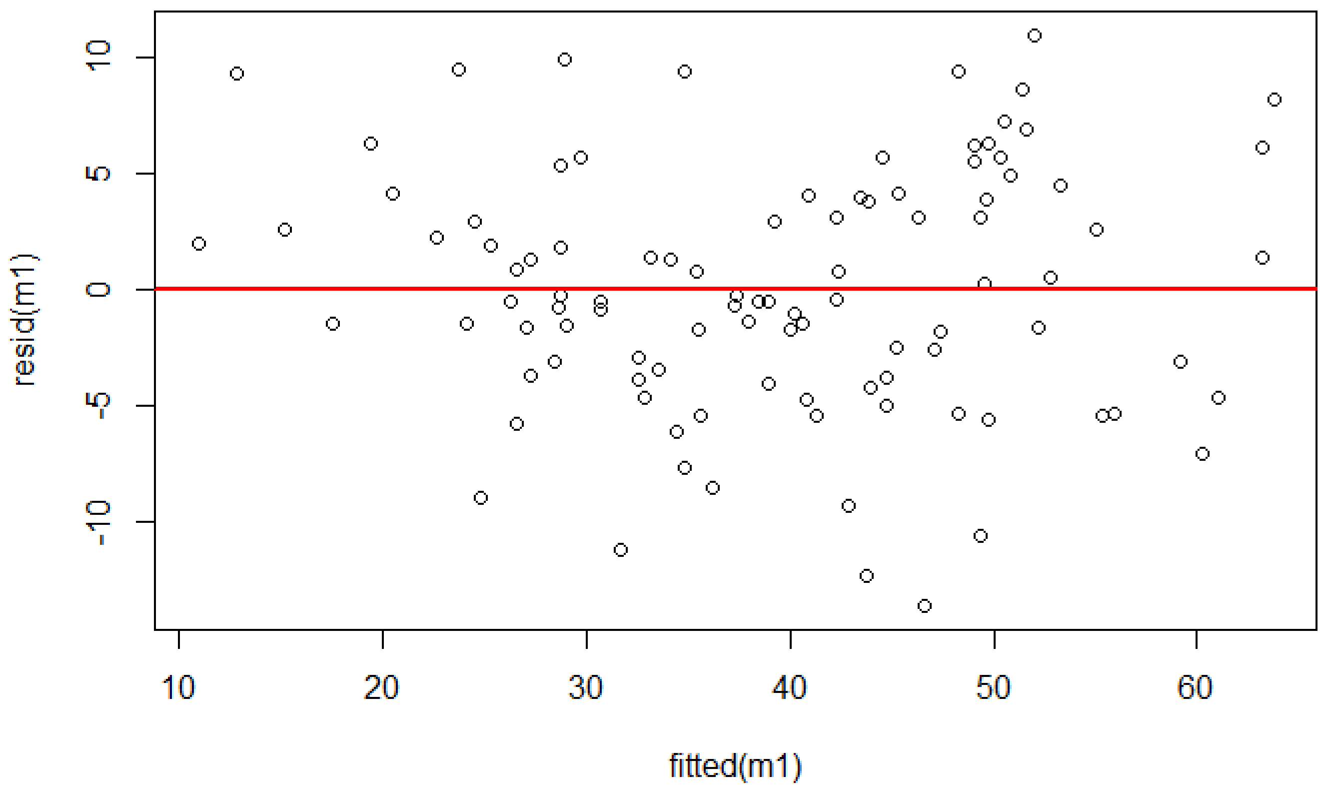




Model validation (case of linear regression y~N(0,1))

- Check that model assumptions have not been violated

Homogeneity of residuals



Unfortunately, not all things in life are normal…

1. Univariate Linear Models — **lm()**

↓

2. Multivariate Linear Models — **lm()**

↓

3. Generalized Linear Models — **glm()**

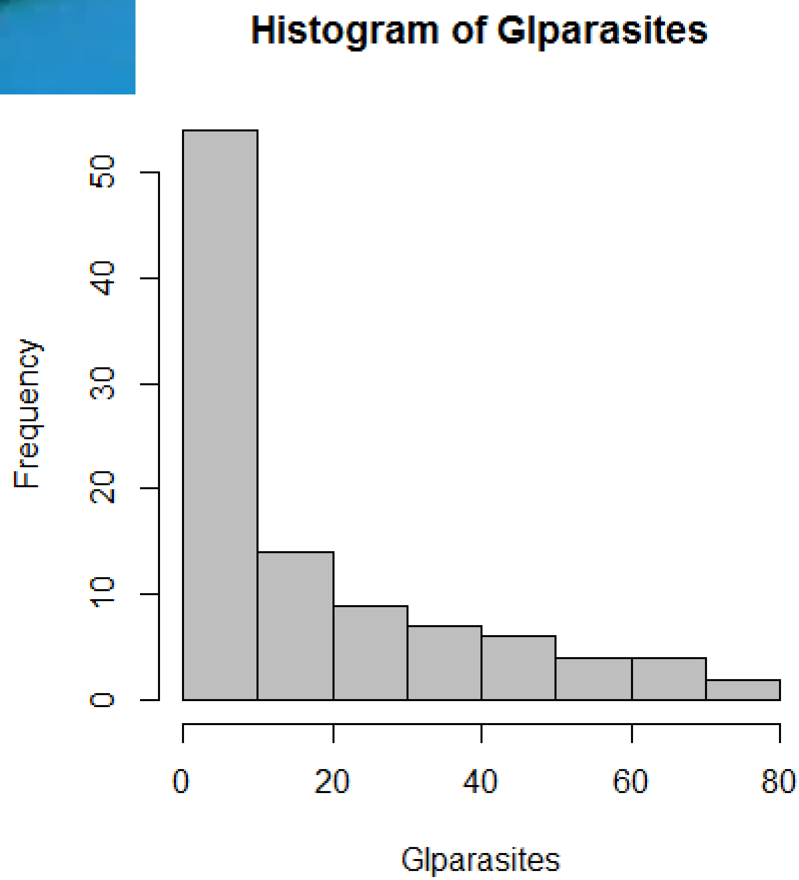
Case when the response variable do **not follow a Normal distribution**

(Dans le cas ou la variable réponse ne suit pas une distribution normale)

Dia ahoana raha tsy manaraka ny “distribution normale” ny “variable response”?

INTRODUCING GENERALIZED LINEAR MODELS

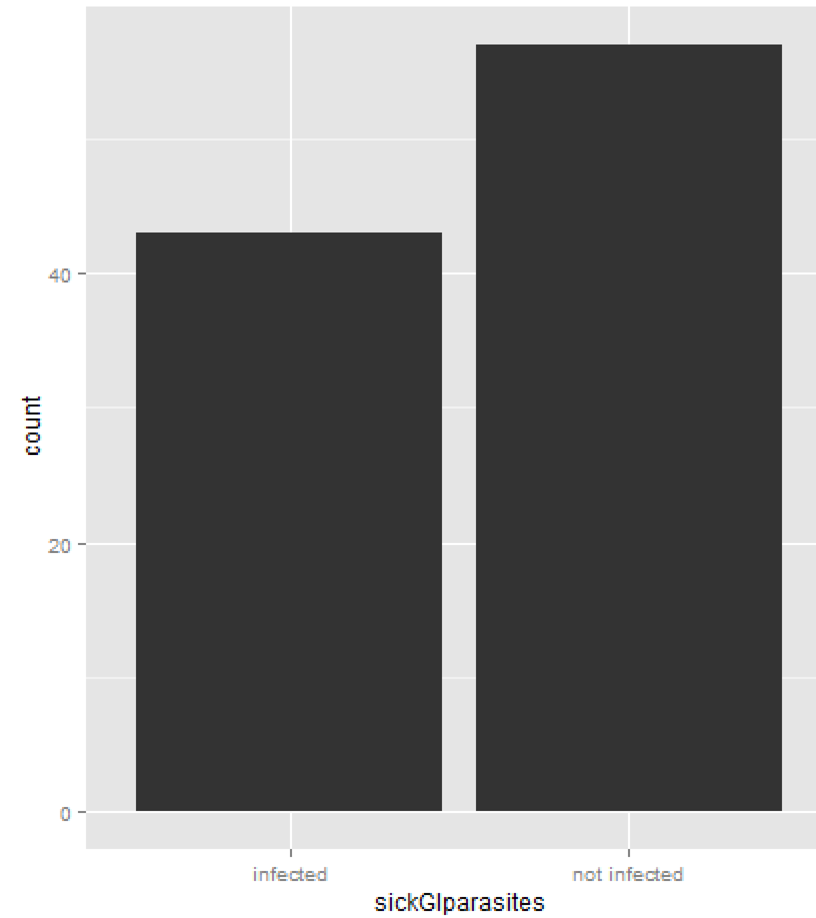
Count data



- 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

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"

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

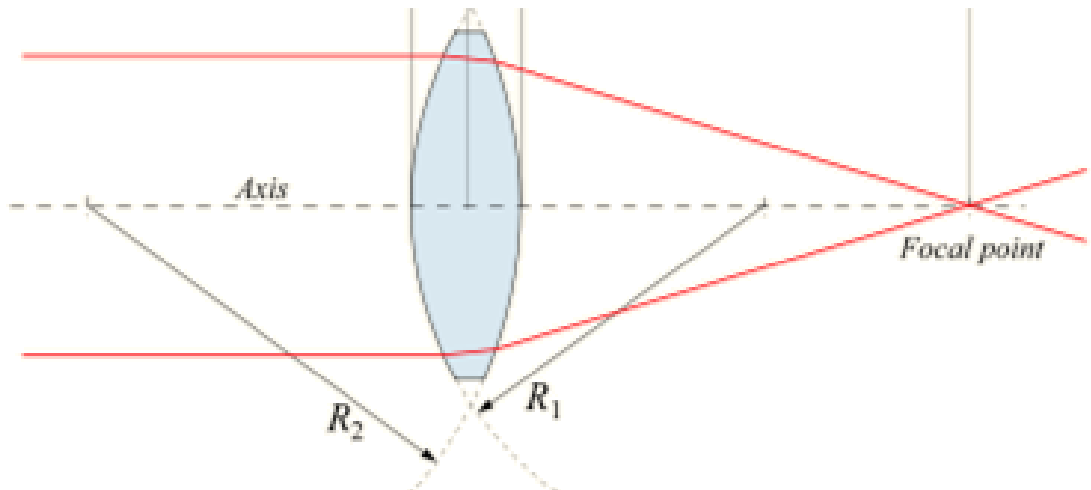
- A link function describes how the mean E(Y) = μ, depends on the linear predictor *v.*

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

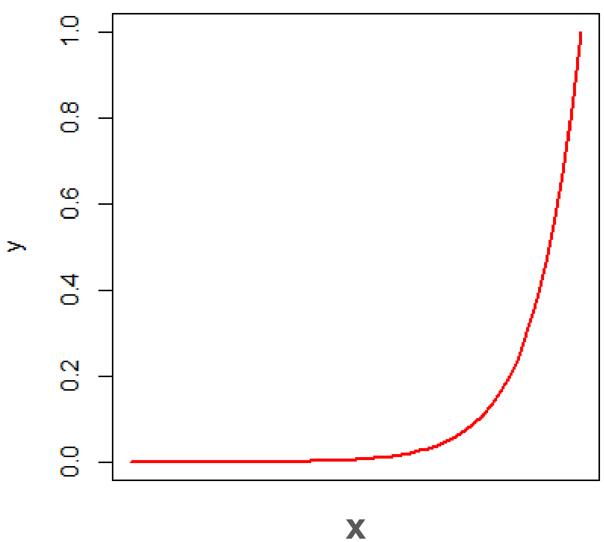
Generalized linear modeling

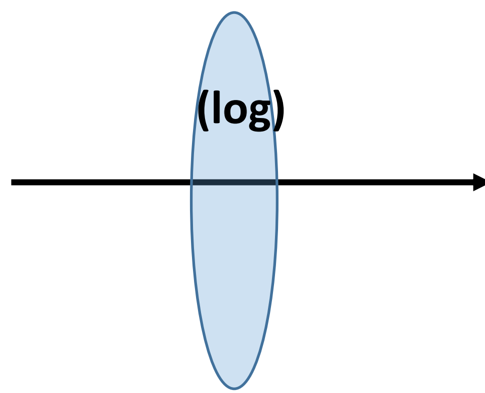
Most common families and links

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

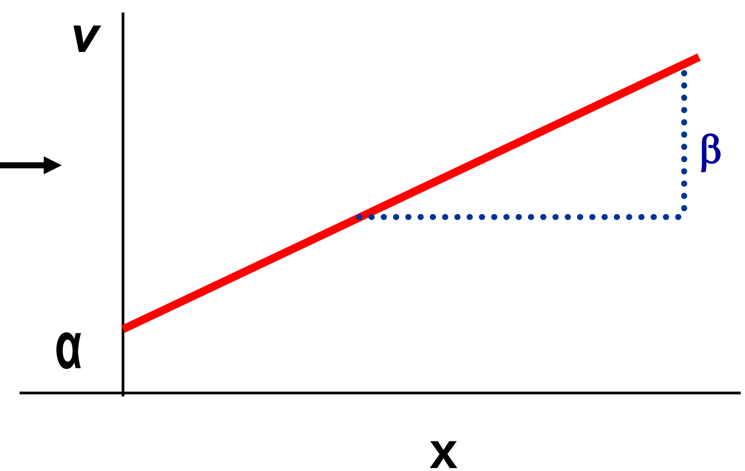


$Y = e^{(\alpha + \beta x)}$





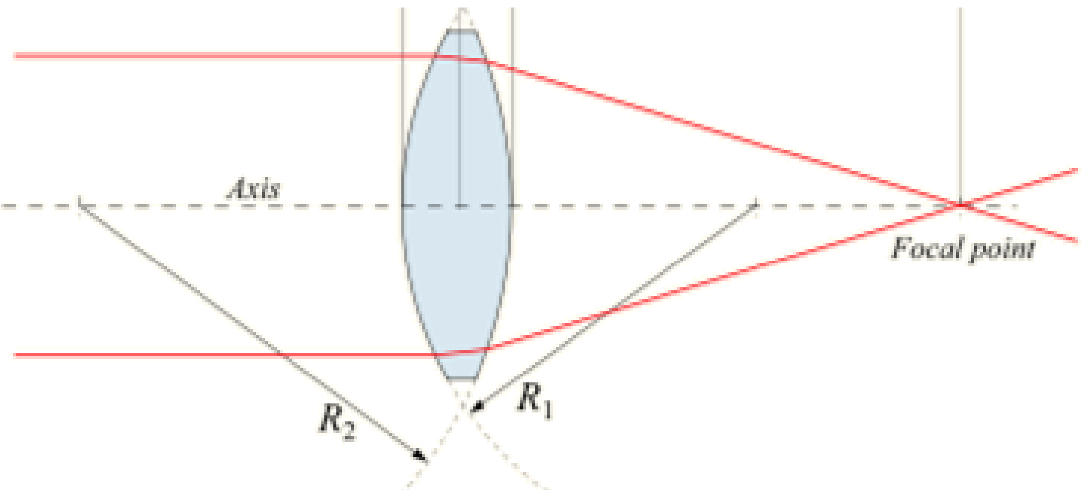
$V = \alpha + \beta x$



Generalized linear modeling

Most common families and links

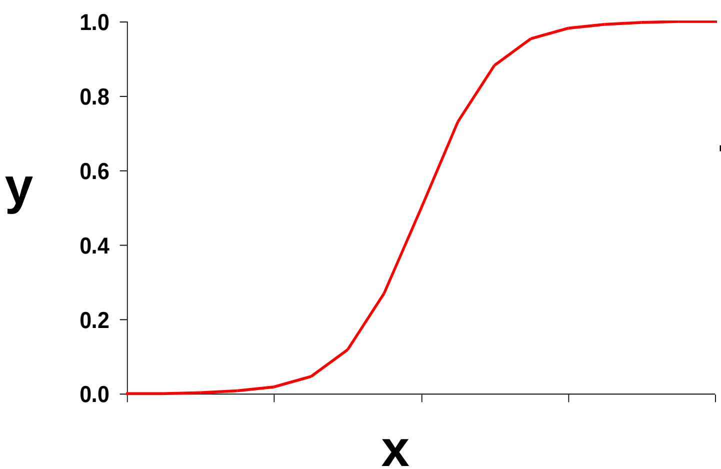
- Gaussian: identity
- Poisson: log
- Binomial: logit
- Negative binomial: log

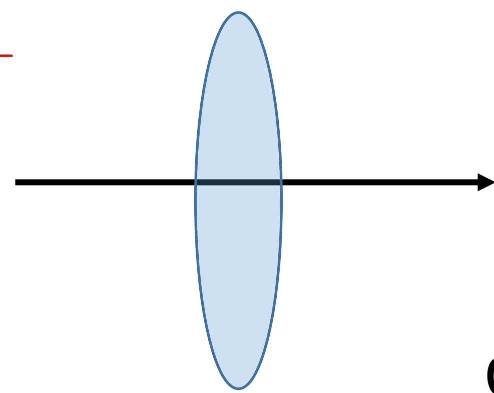


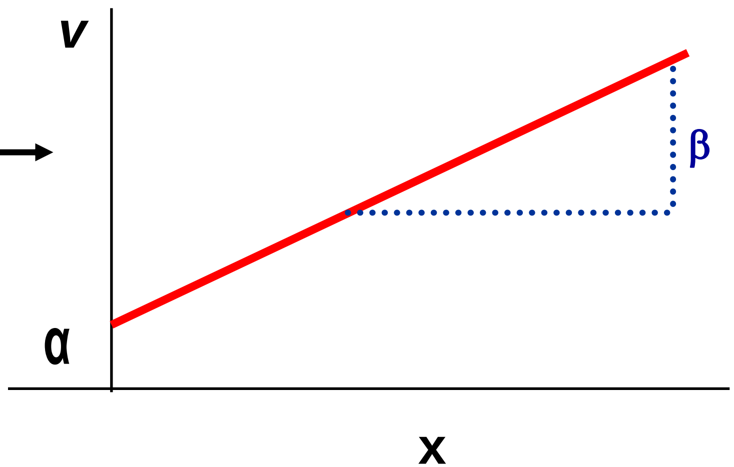
$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

(logit)

$V = \alpha + \beta$ x







Most common families and links

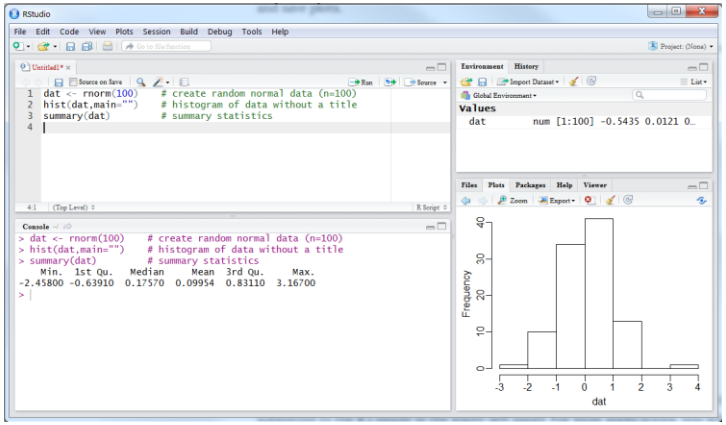
- Gaussian: identity
- Poisson: log
- Binomial: logit
- Negative binomial: log

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)

1. Univariate Linear Models

2. Multivariate Linear Models

3. Generalized Linear Models

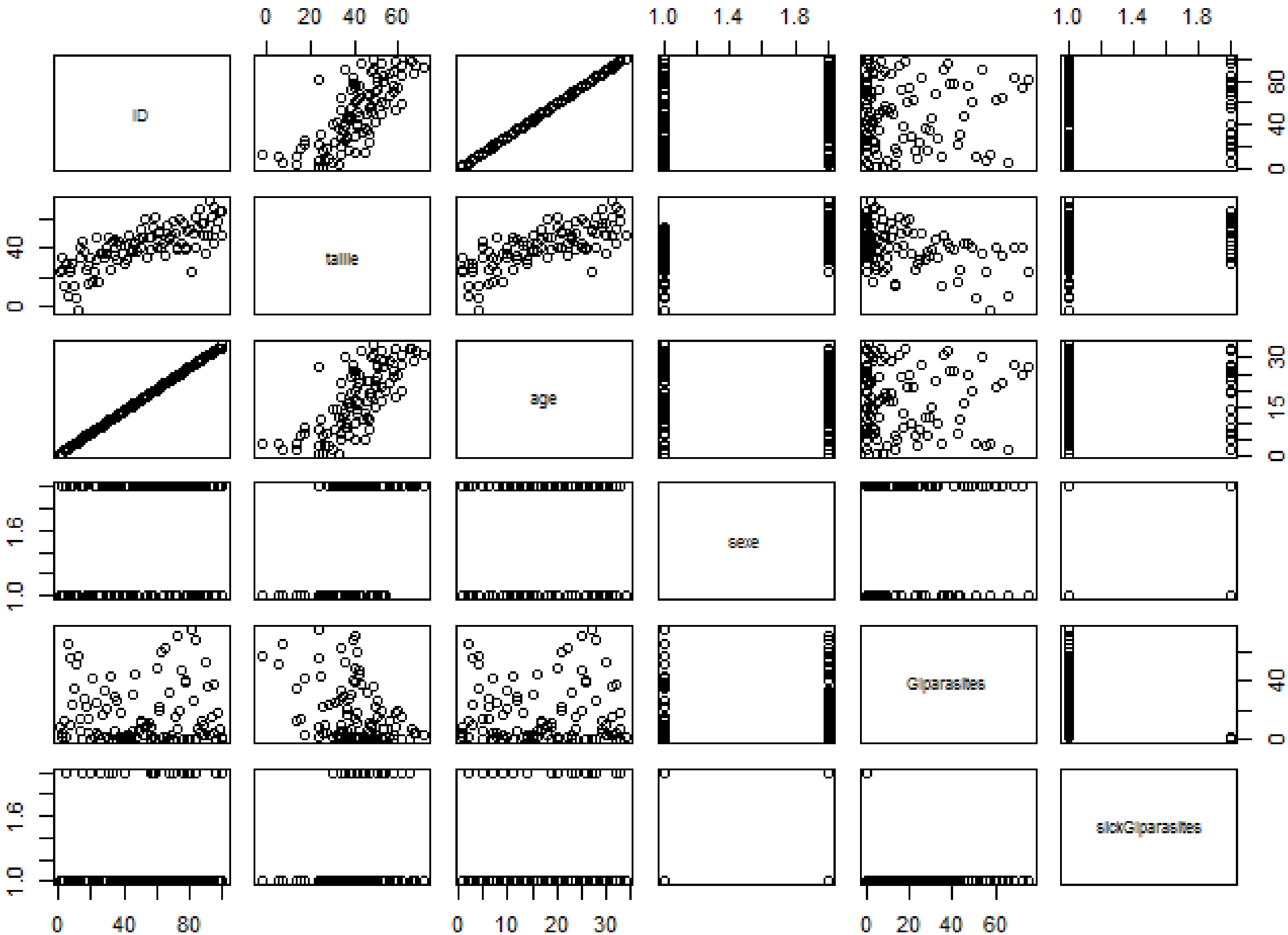


STEPS IN DEVELOPMENT OF STATISTICAL MODELS (TUTORIAL)

Database construction and descriptive analyses

- Relationships between the variables

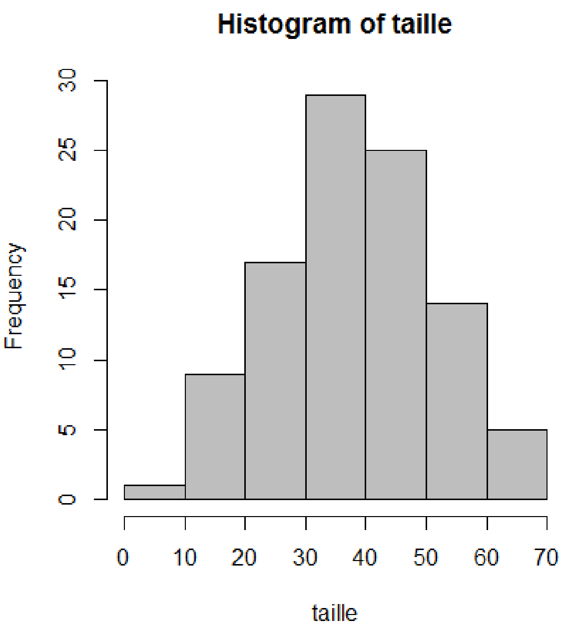
pairs(mydata)



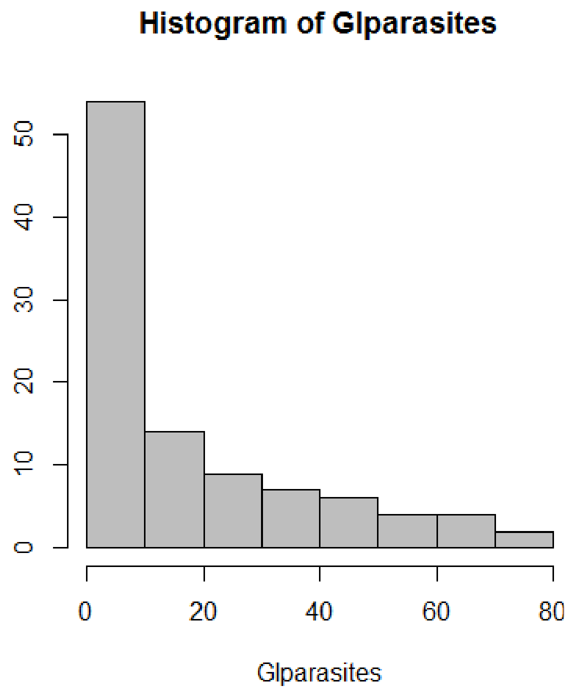
Database construction and descriptive analyses

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

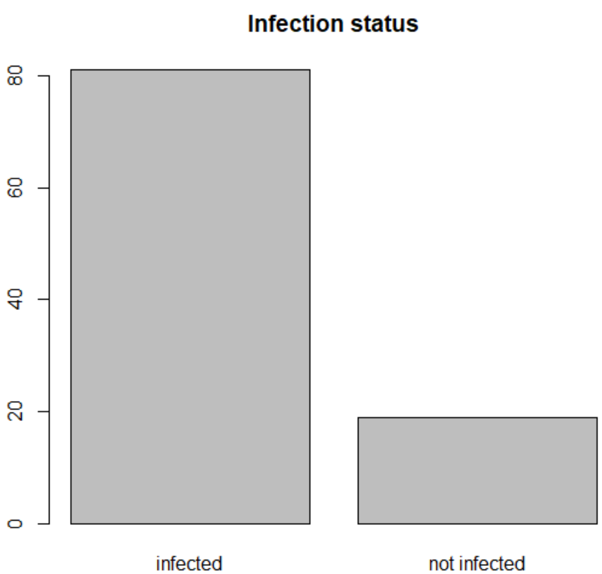
hist(mydata$var)



hist(mydata$var)



plot(mydata$var)

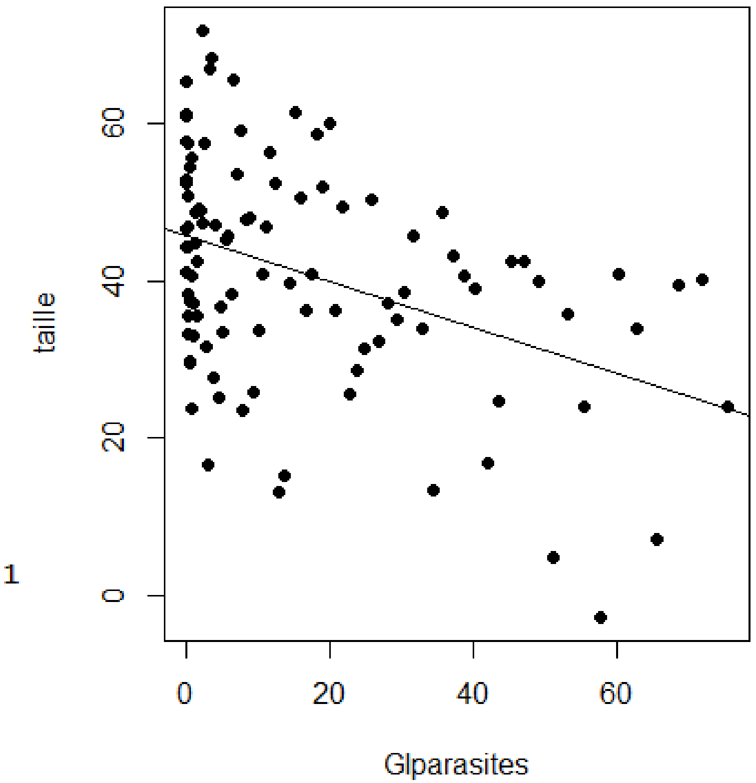


Univariate analyses

- 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)

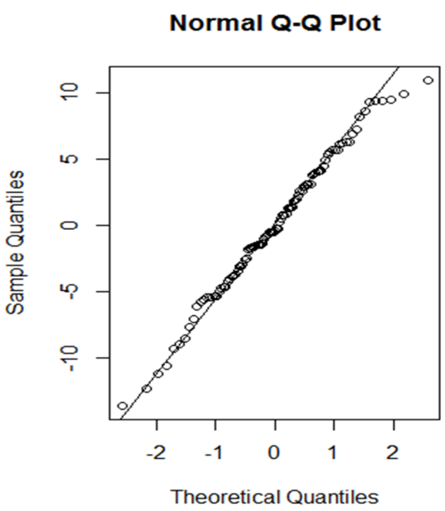
```
Call:
lm(formula = taille ~ age + sexe + GIparasites, data = mydata)

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
```

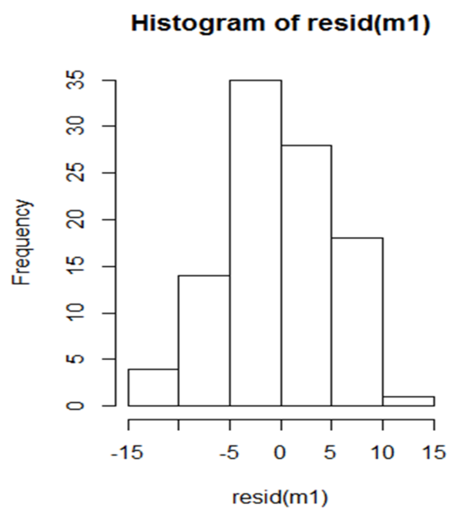
- Select the set of predictors that best explains the response variable (backwards, forward, stepwise)

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

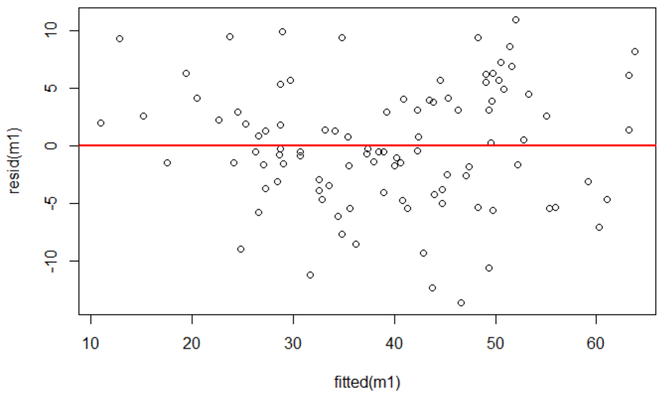
Model validation (case of linear regression y~N(0,1))

- Check that model assumptions have not been violated



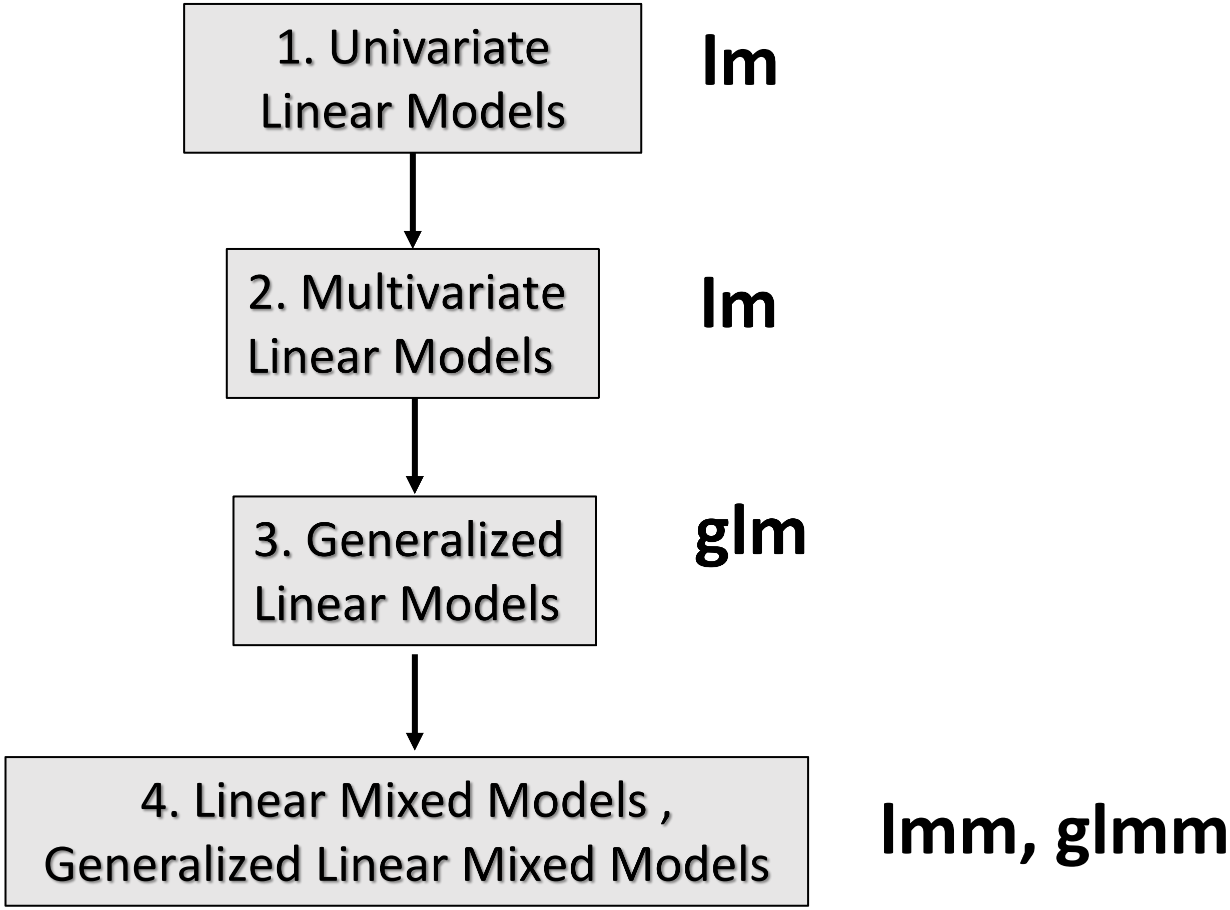


Homogeneity of residuals



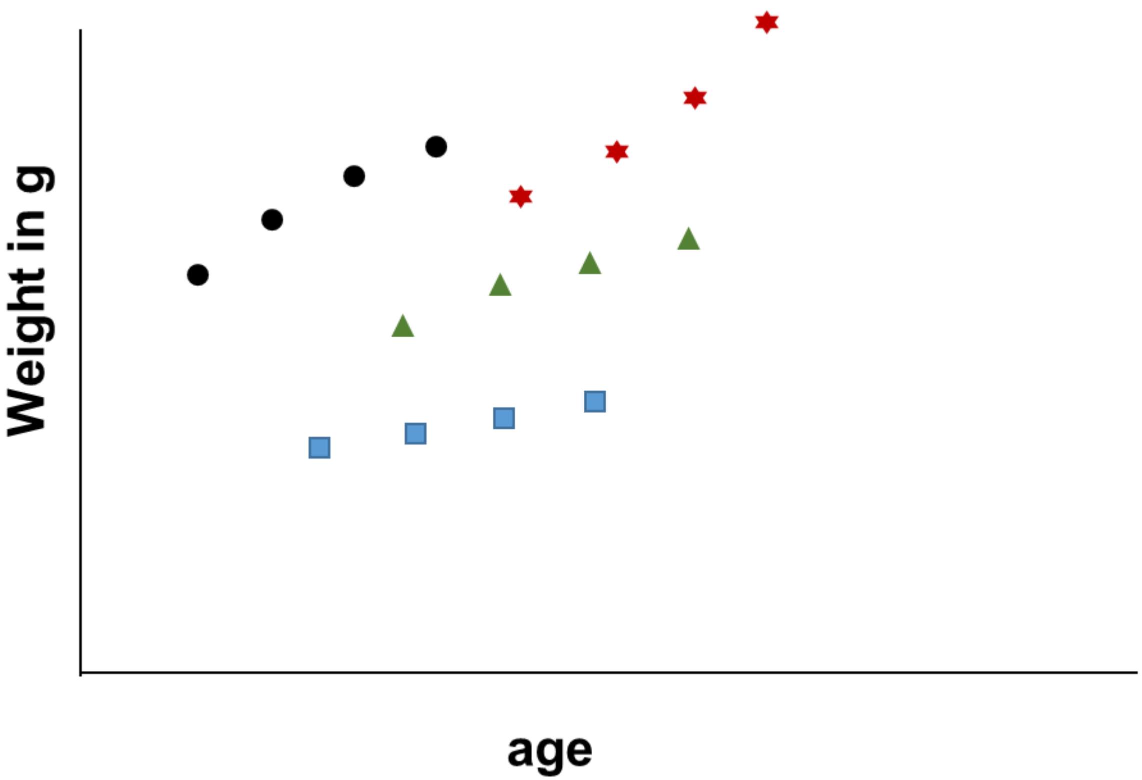
Assumption and limitation of lm, glm

all observations are considered ***independent,***
or they are often nested

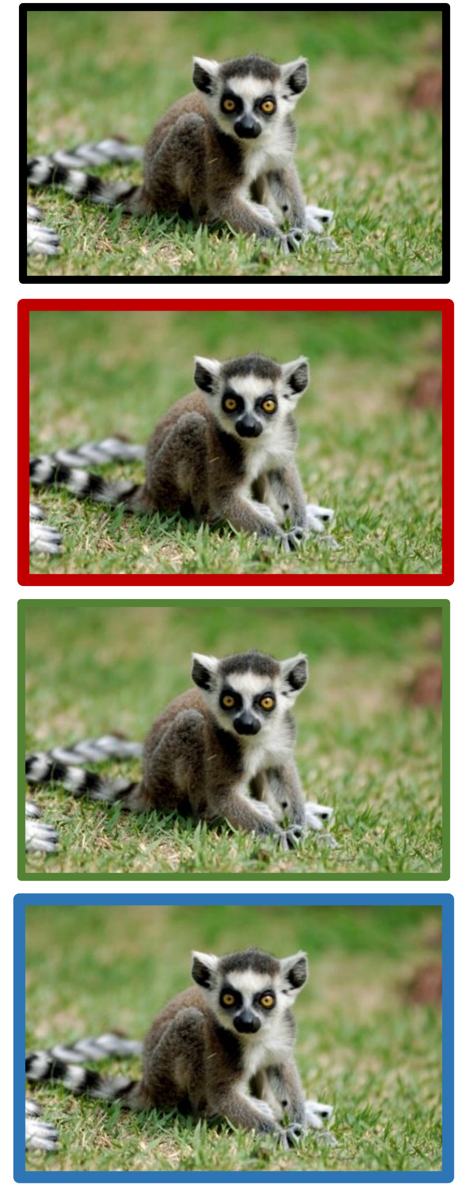


INTRODUCTION TO GENERALIZED LINEAR MIXED MODELS

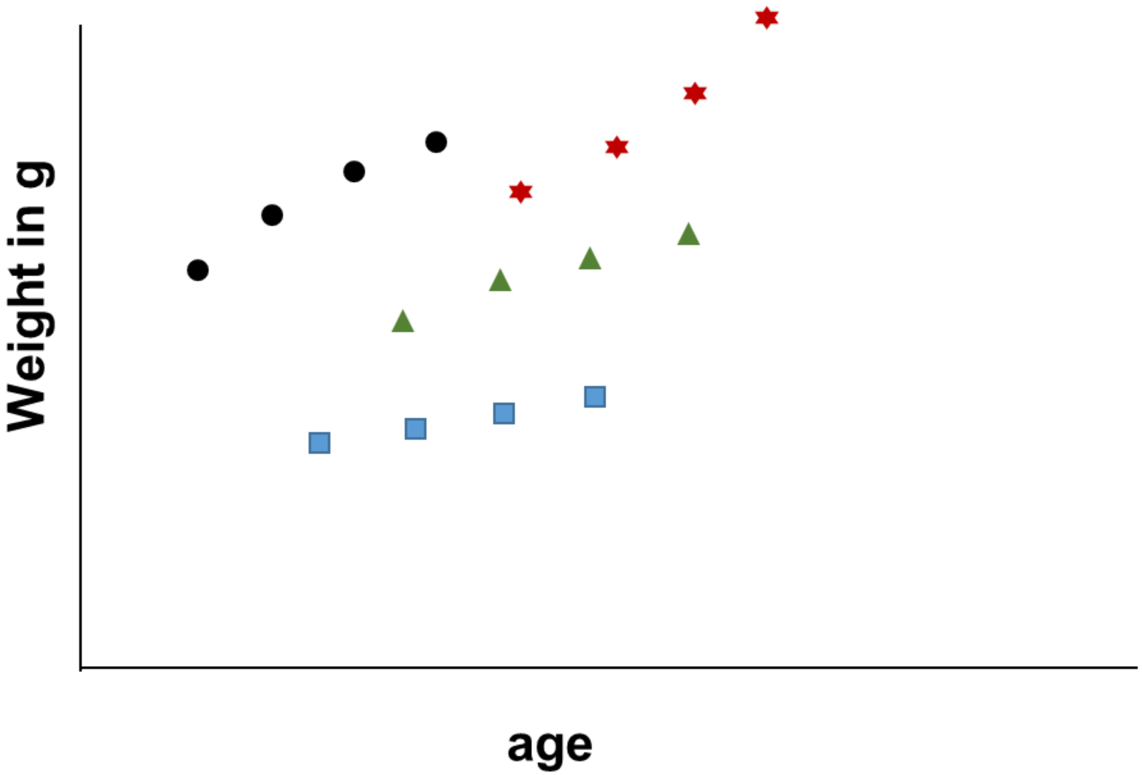
What if…?



Identity (IDs)

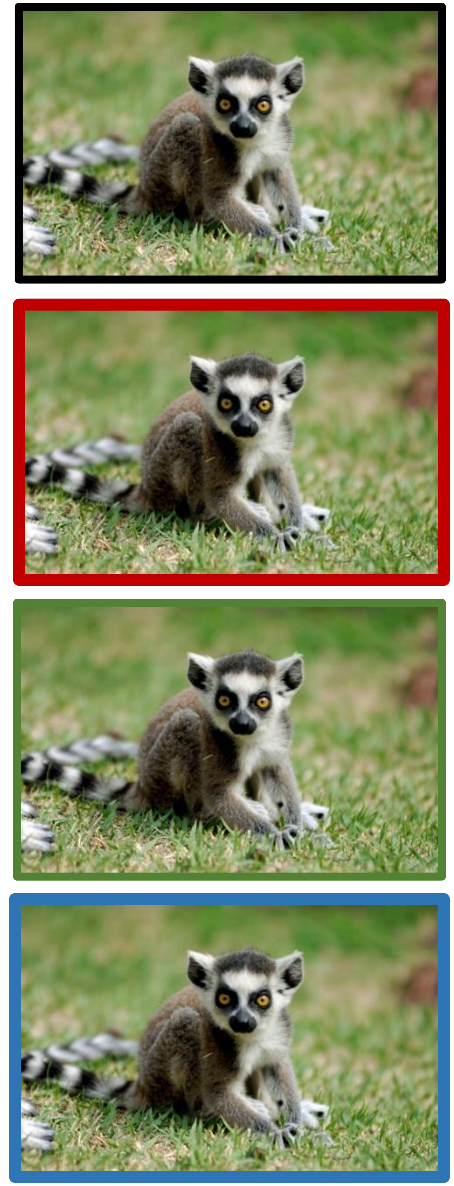


What if…?

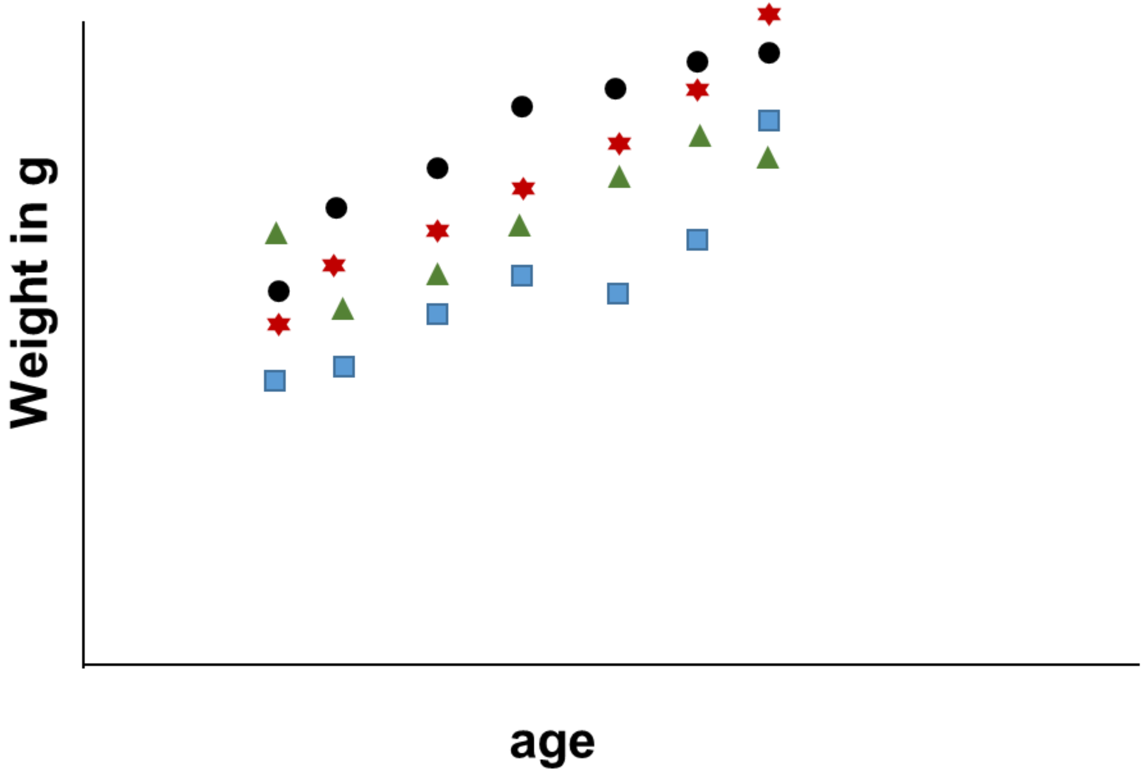


Repeated measure of same individual

Identity (IDs)



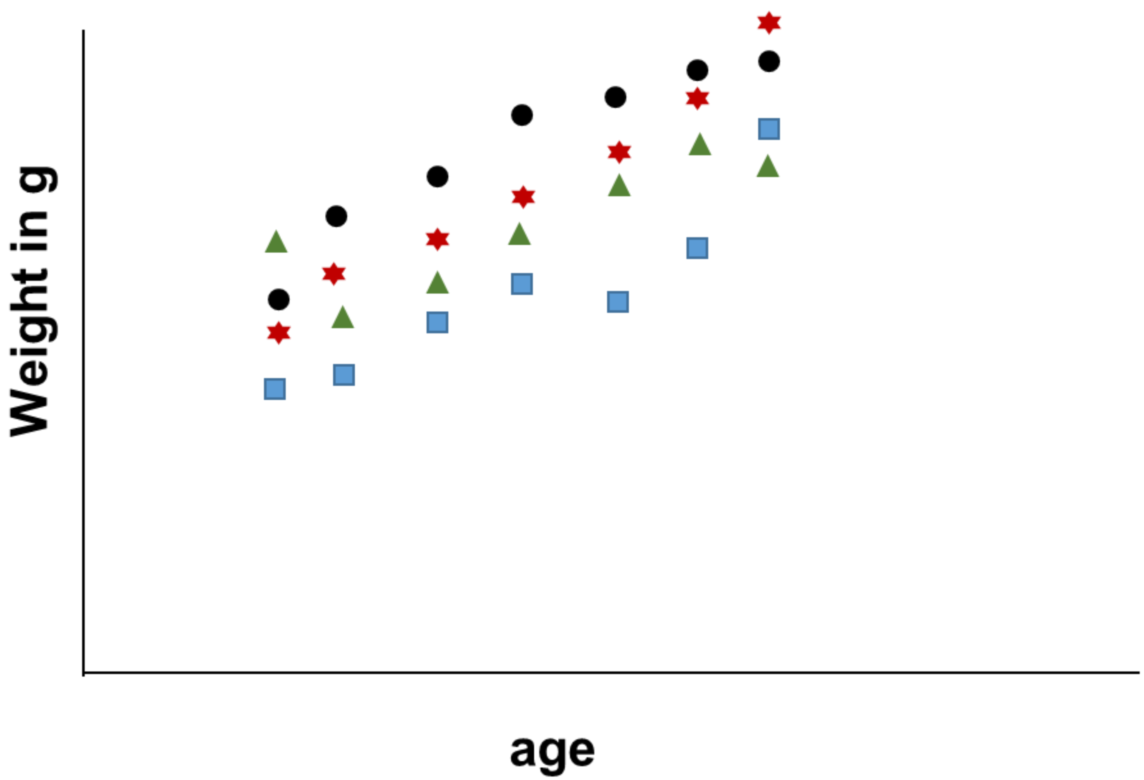
Or…?



Site

- Andasibe
- Ambohipo
- 67Ha
- Ambohipo

Or…?



Site

- Andasibe
- Ambohipo
- 67Ha
- Ambohipo

Independent but clustered

Why use LMM, GLMM?

All explanatory variable we have seen in previous models are called variable that have fix effect

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

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

$$\boxed{y = \underbrace{\beta_i X_i}_{\textit{Fixed Effects}} + \underbrace{b_i Z_i}_{\textit{Random Effects}} + \varepsilon}$$

Random intercept (clustered measures)

$$y = (\alpha + a_i) + \beta_i X_i + \varepsilon$$

Random Effects *Fixed Effects*

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

fitted.model <- lmer(formula, data=data.frame)

Formula : weight ~ age + 1|sites

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

Formula : weight ~ age + 1|sites

Introduction to Linear Regression

Thank you very much!

Misaotra betsaka!

Hafaliana Christian Ranaivoson
Ecology and Evolution (E&E), University of Chicago
Mention Zoologie et Biodiversité Animale (MZBA), University of Antananarivo
Association Ekipa Fanihy (Efa), Madaggascar.

Adapted from
Andrés Garchitorena, PhD (E2M2, 2024)

E²M² Workshop
Andasibe, May 2025