

David Makowski

david.makowski@inrae.fr

david.makowski@universite-paris-saclay.fr

Palaiseau, 2025

Generalized linear model (GLM)

an extension of the linear model

<https://github.com/davemakowski>

Outline

- Some reminders
- GLM: why and what
- Logistic and binomial models
- Poisson models

Some reminders

What is a statistical model?

- A particular type of mathematical model
- A model that includes observables (the measured variables)
- ... and unobservable elements (the parameters and some hidden variables)
- Some of these elements are random variables.

What is a statistical model?

$$Y = f(X, \theta, \varepsilon)$$

Diagram illustrating the components of a statistical model:

- Response variable (points to Y)
- Input variable(s) (points to X)
- Parameter(s) (points to θ)
- Residual (points to ε)

What is a statistical model?

$$Y = X\theta + \varepsilon$$

Diagram illustrating the components of a statistical model:

- Response variable (points to Y)
- Input variable(s) (points to X)
- Parameter(s) (points to θ)
- Residual (points to ε)

What is a linear model?

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1P} \\ x_{21} & x_{22} & \dots & x_{2P} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{NP} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_P \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{pmatrix}$$

N observations of Y

Matrix of $N*P$ values of the P input variables

P parameters

N residuals

$$y_2=x_{21}\theta_1+x_{22}\theta_2+\ldots+x_{2P}\theta_P+\varepsilon_2$$

Many types of linear models

- A continuous explanatory variable
Simple linear regression
- Several continuous explanatory variables
Multiple linear regression
- One categorical explanatory variable
One-way analysis of variance (ANOVA)
- Several categorical explanatory variables
Analysis of variance with 2, 3... factors
- Continuous and categorical explanatory variables
Analysis of covariance (ANCOVA)

Why are they useful?

- Testing a relationship between a response variable (Y) and one or more explanatory variables (X)
Statistical test
- Quantifying the effect of X on Y
Estimation and confidence interval
- Predicting Y as a function of X
Prediction

Other types of statistical models

- **Generalized linear model (GLM)**: useful to deal with categorical response variable Y
- Non-linear models: useful to deal with nonlinear responses
- Non-parametric models: avoid strong assumption
- Mixed models: useful to deal with repeated/longitudinal data

What is a linear model?

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1P} \\ x_{21} & x_{22} & \dots & x_{2P} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{NP} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_P \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{pmatrix}$$

N observations of Y

Matrix of $N*P$ values of the P input variables

P parameters

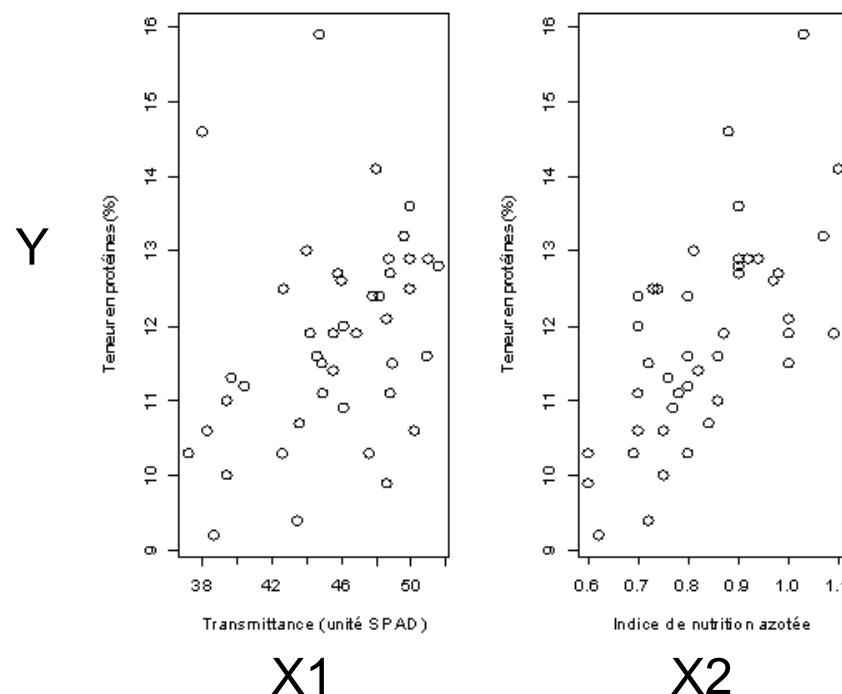
N residuals

Why linear models are so popular?

- Relevant in many cases
- Parameters are easy to estimate
- Many statistical tools are associated with linear models (tests etc.)
- Caution: assumptions are not always verified.

Example: Linear models for predicting wheat grain protein contents

($N=43$, $P=1$, 2 or 3)



Two linear models

```
Mod.1<-lm(TAB$TeneurPro~TAB$SPAD, data=TAB)
```

```
Mod.2<-lm(TAB$TeneurPro~TAB$INN, data=TAB)
```

```
summary(Mod.1)
```

```
summary(Mod.2)
```

Model 1

Call:

```
lm(formula = TAB$TeneurPro ~ TAB$SPAD, data = TAB)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.2791	-0.7759	-0.0530	0.5362	4.1580

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.6901	2.3146	2.890	0.00613 **
TAB\$SPAD	0.1129	0.0506	2.232	0.03113 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.31 on 41 degrees of freedom

Multiple R-squared: 0.1084, Adjusted R-squared: 0.08661

F-statistic: 4.982 on 1 and 41 DF, p-value: 0.03113

Model 2

Call:

```
lm(formula = TAB$TeneurPro ~ TAB$INN, data = TAB)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.7011	-0.5999	-0.1737	0.4900	2.7126

Coefficients:

	Estimate	Std. Error.	t value	Pr(> t)
(Intercept)	6.085	1.040	5.852	7.02e-07 ***
TAB\$INN	6.895	1.232	5.598	1.61e-06 ***

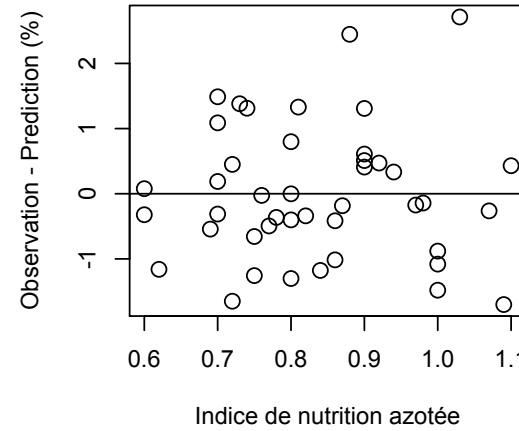
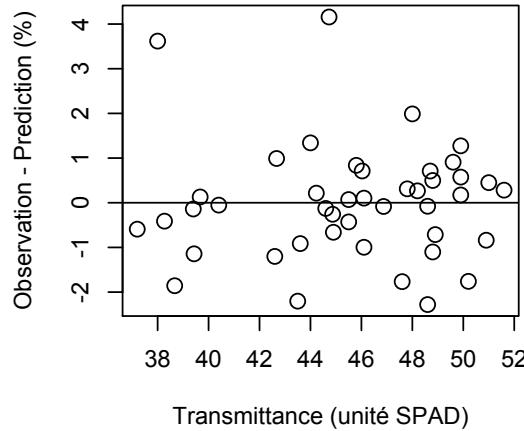
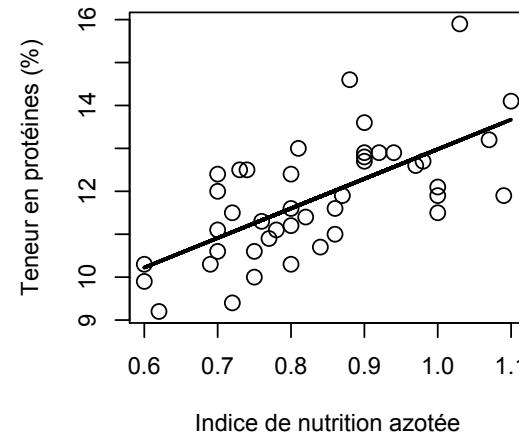
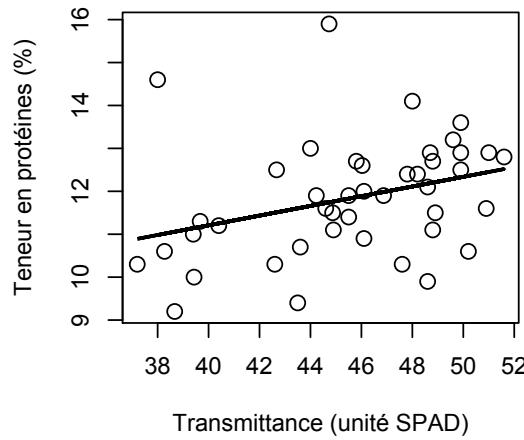
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.045 on 41 degrees of freedom

Multiple R-squared: 0.4332, Adjusted R-squared: 0.4194

F-statistic: 31.33 on 1 and 41 DF, p-value: 1.614e-06

Models 1 and 2



Limits of the linear model

$$y = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p + \varepsilon$$

$$\mathbb{E}(\varepsilon) = 0$$

$$\mathbb{E}(y) = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p$$

The average response is a linear combination
of inputs and parameters.

Limits of the linear model

- Not well adapted to deal with discrete response variables (e.g. 0/1, counting...).
- Not very flexible in terms of response function

Outline

- Some reminders
- GLM: why and what
- Logistic and binomial models
- Poisson models

Why GLM?

GLM: a standard approach to deal with categorical responses

- Binary data: 0/1, Yes/No
- Proportion data: K positive cases out of N
- Count data: $0, 1, 2, \dots, K$ occurrences

GLM: a standard approach to deal with categorical responses

- Able to handle categorical data
- Include interpretable parameters
- Deal with some particular types of nonlinear responses
- More general than the linear model (LM is a special case of GLM)

Two components of a GLM

Deterministic component:

Express the expected values of the observations as a function of the inputs

Random component:

Describe the variability of the observations

Deterministic component based on a link function

The link function g is defined such as:

$$g[E(Y|X_1, \dots, X_p)] = \theta_0 + \theta_1 X_1 + \dots + \theta_p X_p$$

$$E(Y|X_1, \dots, X_p) = g^{-1}(\theta_0 + \theta_1 X_1 + \dots + \theta_p X_p)$$

Deterministic component based on a link function

The link function g is defined such as:

$$g[E(Y|X_1, \dots, X_p)] = \theta_0 + \theta_1 X_1 + \dots + \theta_p X_p$$

$$E(Y|X_1, \dots, X_p) = g^{-1}(\theta_0 + \theta_1 X_1 + \dots + \theta_p X_p)$$

Examples of link functions commonly used:

$$g(y) = \log(y)$$

$$g(y) = \text{logit}(y) = \log(y/(1-y))$$

Deterministic component based on a link function

The link function g is defined such as:

$$g[E(Y|X_1, \dots, X_p)] = \theta_0 + \theta_1 X_1 + \dots + \theta_p X_p$$

$$E(Y|X_1, \dots, X_p) = g^{-1}(\theta_0 + \theta_1 X_1 + \dots + \theta_p X_p)$$

Examples of link functions commonly used:

$$g(y) = \log(y)$$

$$g(y) = \text{logit}(y) = \log(y/(1-y))$$

$$E(Y|X_1, \dots, X_p)$$

$$= \exp(\theta_0 + \theta_1 X_1 + \dots + \theta_p X_p)$$

Deterministic component based on a link function

The link function g is defined such as:

$$g[E(Y|X_1, \dots, X_p)] = \theta_0 + \theta_1 X_1 + \dots + \theta_p X_p$$

$$E(Y|X_1, \dots, X_p) = g^{-1}(\theta_0 + \theta_1 X_1 + \dots + \theta_p X_p)$$

Examples of link functions commonly used:

$$g(y) = \log(y)$$

$$g(y) = \text{logit}(y) = \log(y/(1-y))$$

$$\begin{aligned} & E(Y|X_1, \dots, X_p) \\ &= \frac{\exp(\theta_0 + \theta_1 X_1 + \dots + \theta_p X_p)}{1 + \exp(\theta_0 + \theta_1 X_1 + \dots + \theta_p X_p)} \end{aligned}$$

Random component based on the exponential family

- Describe the distribution of the data conditionnally to the inputs (i.e., distribution of $Y|X$)
- Distributions commonly used:
 - Bernouilli
 - Binomial
 - Poisson
 - (Gaussian)

Random component based on the exponential family

- Describe the distribution of the data conditionnally to the inputs (i.e., distribution of $Y|X$)
- Distributions commonly used:
 - Bernouilli: binary observations (0/1)
 - Binomial
 - Poisson
 - (Gaussian)

$$Y \sim B(\pi)$$

Probability of $Y=1$



Random component based on the exponential family

- Describe the distribution of the data conditionnally to the inputs (i.e., distribution of $Y|X$)
- Distributions commonly used:
 - Bernouilli
 - Binomial: number of « success » out of N
 - Poisson
 - (Gaussian)

$$Y \sim \text{Bin}(N, \pi)$$

Probability of « success »

Random component based on the exponential family

- Describe the distribution of the data conditionnally to the inputs (i.e., distribution of $Y|X$)
- Distributions commonly used:
 - Bernouilli
 - Binomial
 - Poisson: count data
 - (Gaussian)

$$Y \sim P(\lambda)$$



Mean number of occurrences

Distribution	Mean	Variance
Gaussian	μ	σ^2
Bernouilli	π	$\pi(1 - \pi)$
Binomial	$N\pi$	$N\pi(1 - \pi)$
Poisson	λ	λ

Popular GLMs

Model	Link function	Distribution
Linear Gaussian	identity	Gaussian
Logistic regression Binomial model	logit	Bernouilli or Binomial
Poisson log linear	log	Poisson

Example 1: Effect of insecticide in mortality of caterpillars

	$\text{Log}_2 (\text{Dose insecticide})$					
Sexe	0	1	2	3	4	5
M	1	4	9	13	18	20
F	0	2	6	10	12	16

Collett, 1991

Number of deaths out
of 20 individuals

Probability of death = $f(\text{sex}, \text{dose})$

Call:

```
glm(formula = propMort ~ ldose * sexe, family = binomial, weights = Freq)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.39849	-0.32094	-0.07592	0.38220	1.10375

Coefficients:

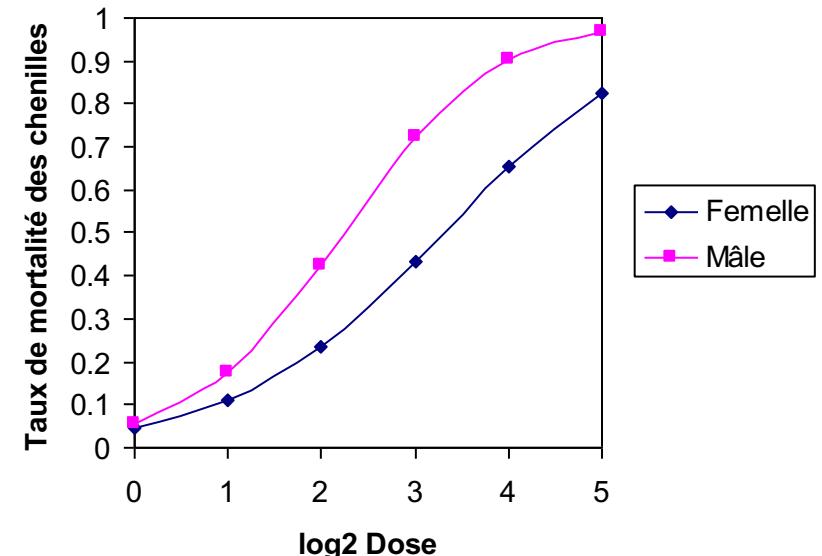
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.9935	0.5527	-5.416	6.09e-08 ***
ldose	0.9060	0.1671	5.422	5.89e-08 ***
sexeM	0.1750	0.7783	0.225	0.822
ldose:sexeM	0.3529	0.2700	1.307	0.191

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

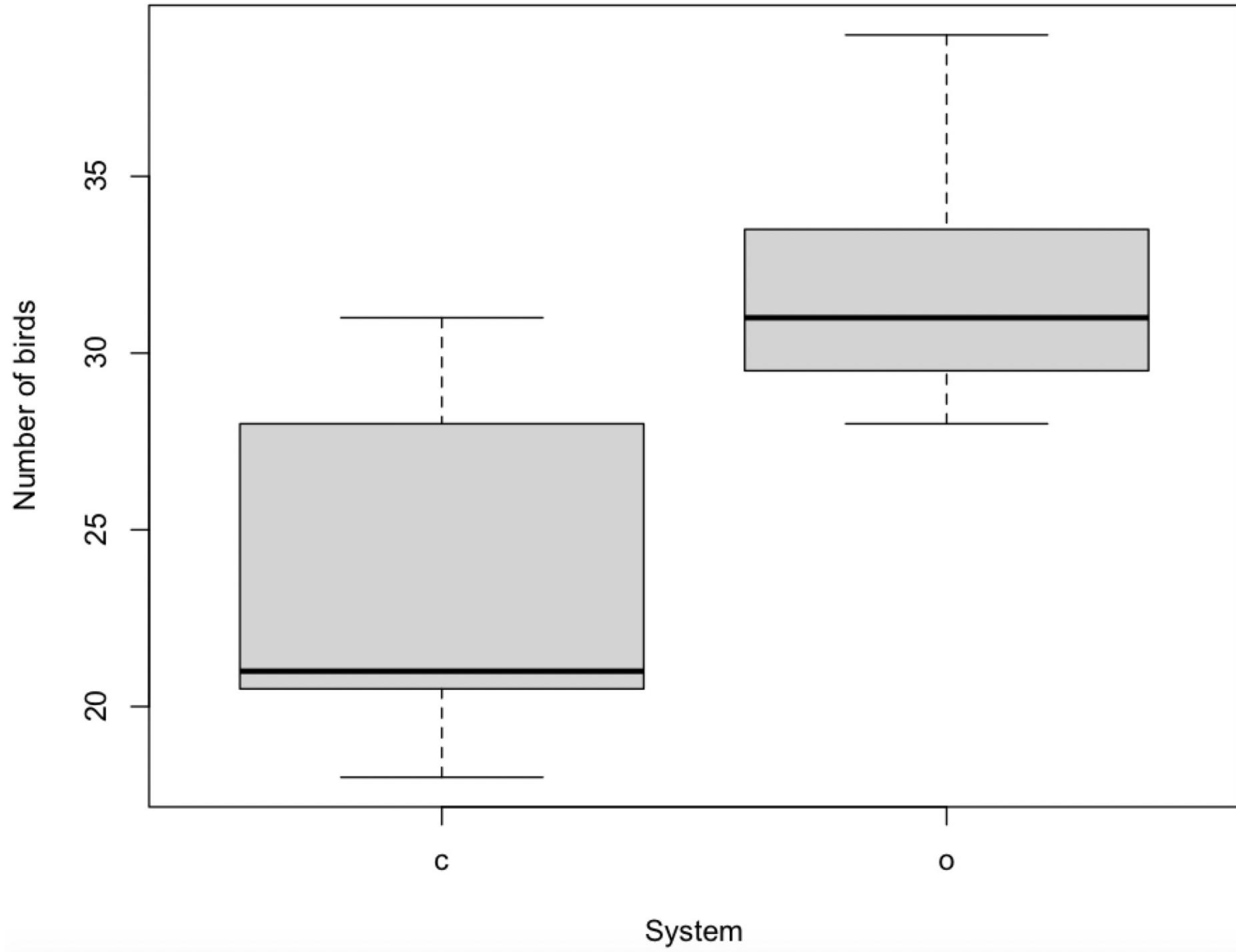
Null deviance: 124.8756 on 11 degrees of freedom
Residual deviance: 4.9937 on 8 degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4



Example 2: Number of birds in Organic vs. Conventional farming

	System	Birds	
1	o	30	
2	o	29	
3	o	35	Count data
4	o	32	
5	o	28	
6	o	31	
7	o	39	
8	c	20	
9	c	25	
10	c	31	
11	c	31	
12	c	18	
13	c	21	
14	c	21	



GLM: Poisson log linear model

Call:

```
glm(formula = Birds ~ System, family = poisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2541	-0.5973	-0.2675	0.4498	1.3973

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.17208	0.07738	40.992	< 2e-16 ***
Systemo	0.29365	0.10224	2.872	0.00408 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 17.9192 on 13 degrees of freedom
Residual deviance: 9.5801 on 12 degrees of freedom
AIC: 85.679

Number of Fisher Scoring iterations: 4

$$\begin{aligned}\text{log(mean bird number)} &= \text{Intercept} + \text{Systemo}^*X \\ &= 3.17 \quad \text{for Conventional} \\ &= 3.17 + 0.29 = 3.46 \quad \text{for Organic}\end{aligned}$$

Call:

```
glm(formula = Birds ~ System, family = poisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2541	-0.5973	-0.2675	0.4498	1.3973

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.17208	0.07738	40.992	< 2e-16 ***
Systemo	0.29365	0.10224	2.872	0.00408 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 17.9192 on 13 degrees of freedom
Residual deviance: 9.5801 on 12 degrees of freedom
AIC: 85.679

Number of Fisher Scoring iterations: 4

$$\begin{aligned}\text{mean bird number} &= \exp(\text{Intercept} + \text{Systemo}^*X) \\ &= 23.8 \quad \text{for Conventional} \\ &= 31.8 \quad \text{for Organic}\end{aligned}$$

Call:

```
glm(formula = Birds ~ System, family = poisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2541	-0.5973	-0.2675	0.4498	1.3973

Coefficients:

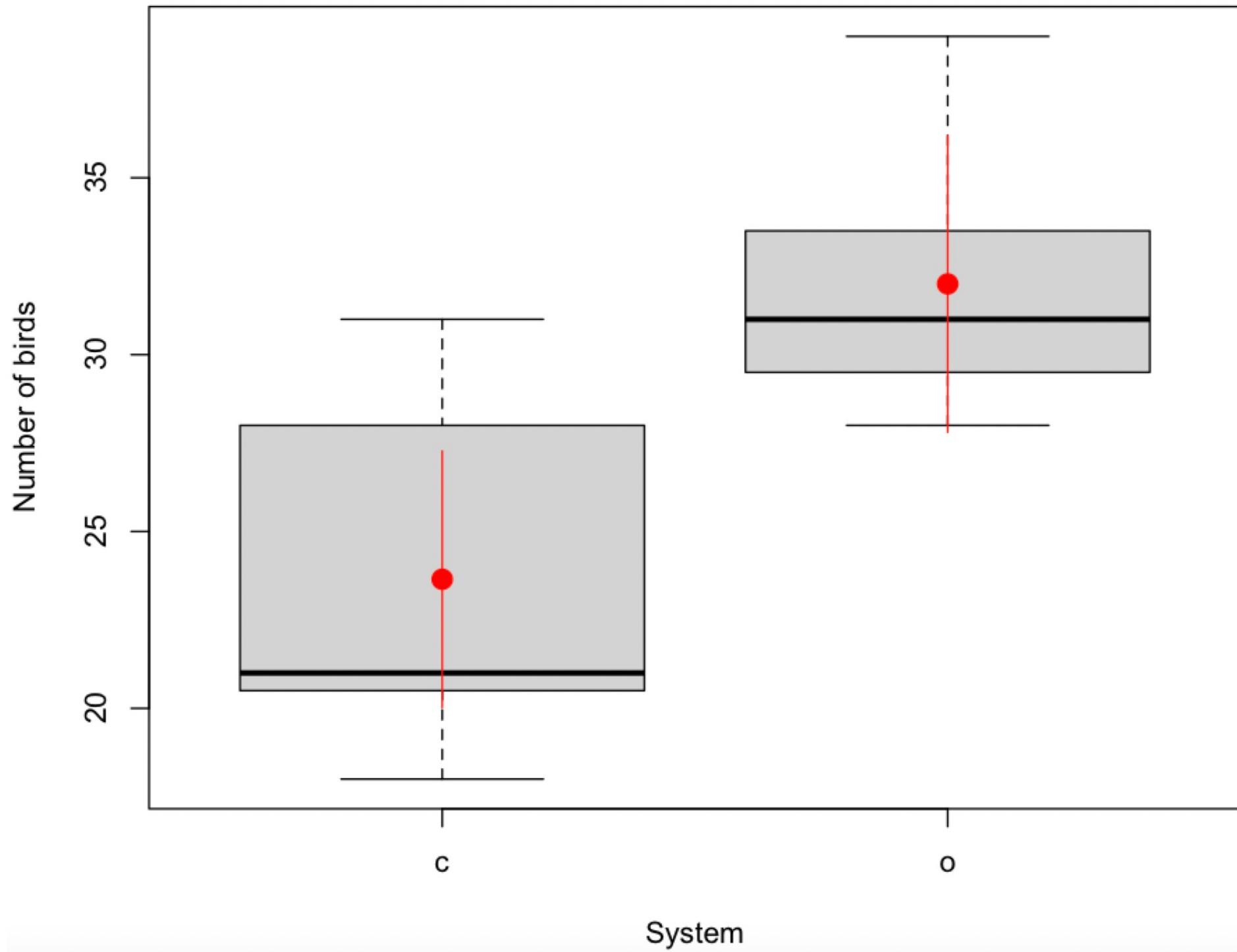
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.17208	0.07738	40.992	< 2e-16 ***
Systemo	0.29365	0.10224	2.872	0.00408 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 17.9192 on 13 degrees of freedom
Residual deviance: 9.5801 on 12 degrees of freedom
AIC: 85.679

Number of Fisher Scoring iterations: 4



Example 3: modelling the effect of temperature on the survival of an invasive species

Data:

- Pieces of wood with heat treatment
- Five different temperatures tested
- Observations: Number of insects surviving heat treatment

Modèle Poisson log linéaire

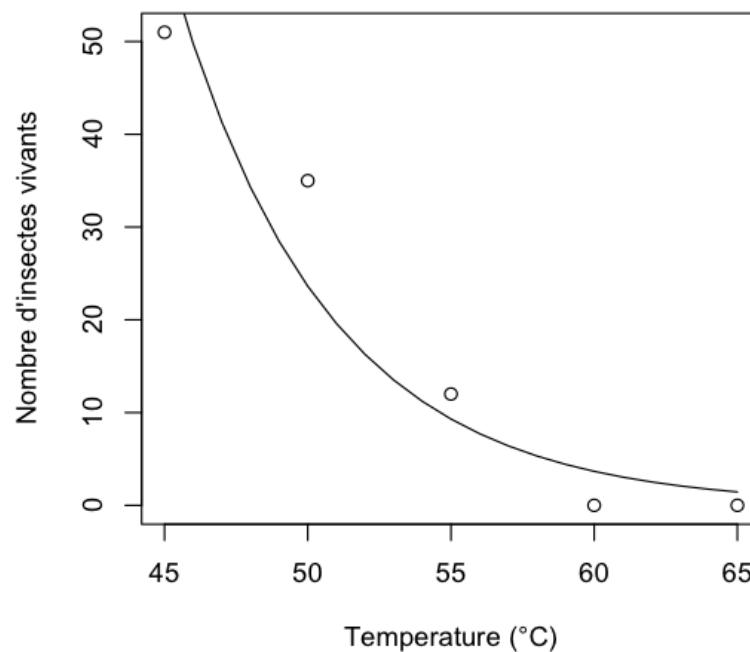
$$\text{mean insect number} = \exp(\text{Intercept} + \text{Slope} * \text{time}) \\ = \exp(12.5 - 0.19 * \text{Time})$$

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	12.47853	1.06901	11.673	<2e-16 ***
X	-0.18633	0.02217	-8.406	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

AIC: 36.62



Example 4: modelling the number of fires across time

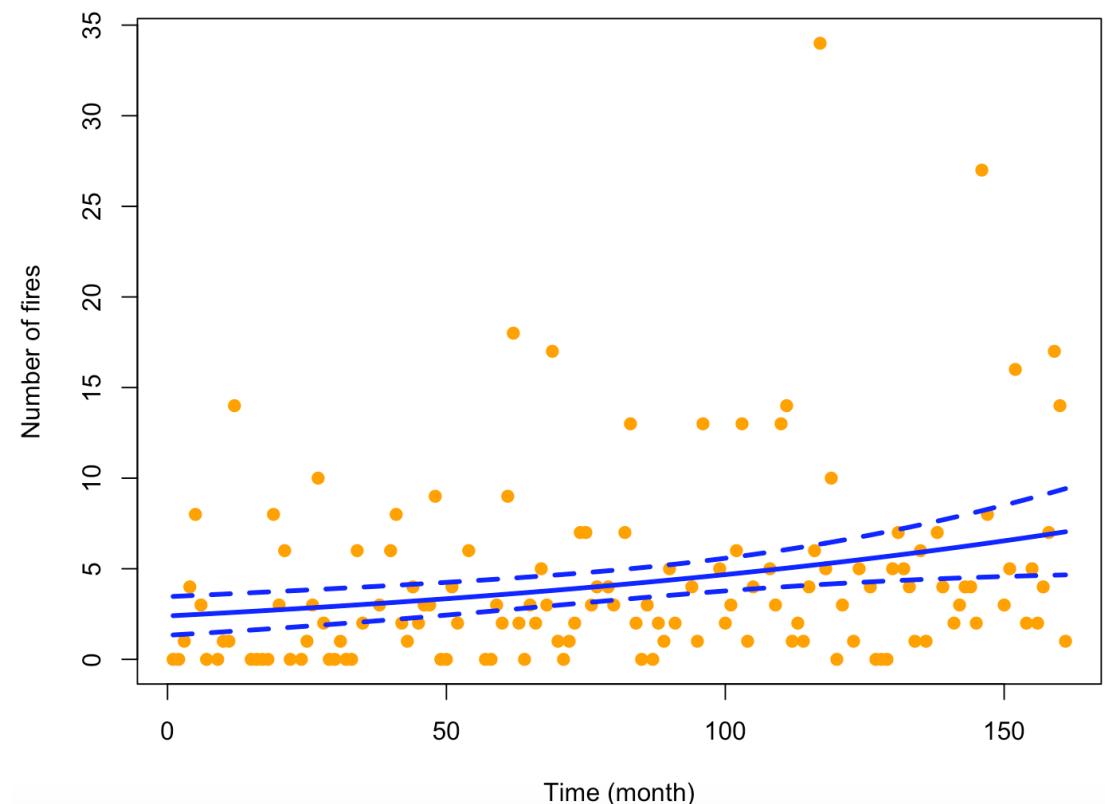
Data:

- One site in the US (county) during several years
- Observations: number of fires every month

$$\text{mean fire number} = \exp(\text{Intercept} + \text{Slope} * \text{time}) \\ = \exp(0.87 + 0.0067 * \text{Time})$$

Modèle quasi-Poisson log linéaire

```
Call:  
glm(formula = Y ~ X, family = "quasipoisson", data = DATA)  
  
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-3.3712 -2.1217 -0.7257  0.3497  8.3430  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 0.872939  0.226930  3.847 0.000184 ***  
X           0.006701  0.002154  3.110 0.002284 **  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for quasipoisson family taken to be 5.611049)  
  
Null deviance: 671.24 on 135 degrees of freedom  
Residual deviance: 615.46 on 134 degrees of freedom  
AIC: NA  
  
Number of Fisher Scoring iterations: 5
```



Outline

- Some reminders
- GLM: why and what
- Logistic and binomial models
- Poisson models

Outline

- Logistic and binomial models
 - Definition
 - Estimation
 - Test and deviance
 - Useful variants
 - Model selection
- Poisson models
 - Definition
 - Estimation
 - Test and deviance
 - Useful variants
 - Model selection

Outline

- **Logistic and binomial models**

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

- **Poisson models**

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

Definition

Model	Link function	Distribution
Linear Gaussian	identity	Gaussian
Logistic regression Binomial model	logit	Bernouilli or Binomial
Poisson log linear	log	Poisson

Definition (binary observations)

Deterministic component

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p$$

Random component

$$Y \sim B(\pi)$$

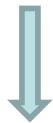
where Y is a binary observation (0 ou 1)

π is the probability of $Y=1$

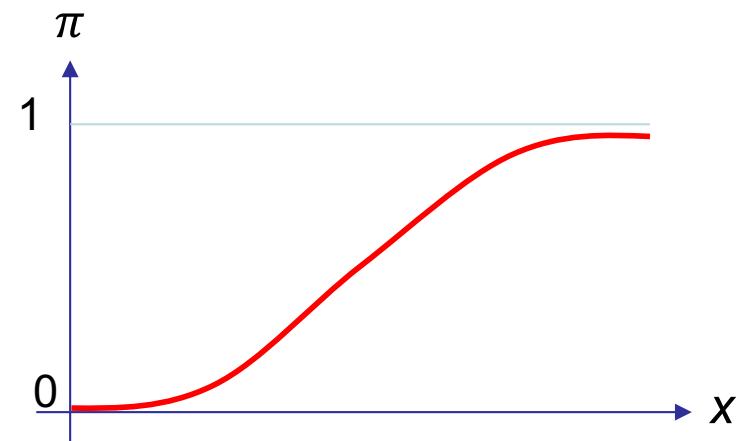
Definition (binary observations)

Deterministic component

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p$$



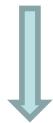
$$\pi = \frac{\exp(\theta_0 + \theta_1 x_1 + \dots + \theta_p x_p)}{1 + \exp(\theta_0 + \theta_1 x_1 + \dots + \theta_p x_p)}$$



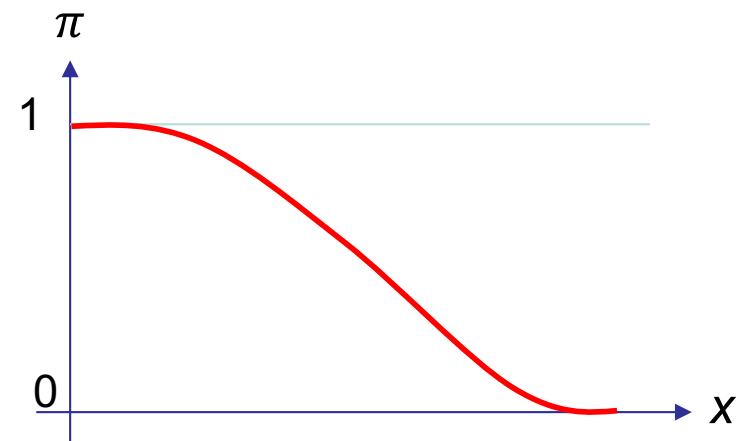
Definition (binary observations)

Deterministic component

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p$$



$$\pi = \frac{\exp(\theta_0 + \theta_1 x_1 + \dots + \theta_p x_p)}{1 + \exp(\theta_0 + \theta_1 x_1 + \dots + \theta_p x_p)}$$



Definition (binomial observations)

Deterministic component

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p$$

Random component

$$Y \sim Bin(N, \pi)$$

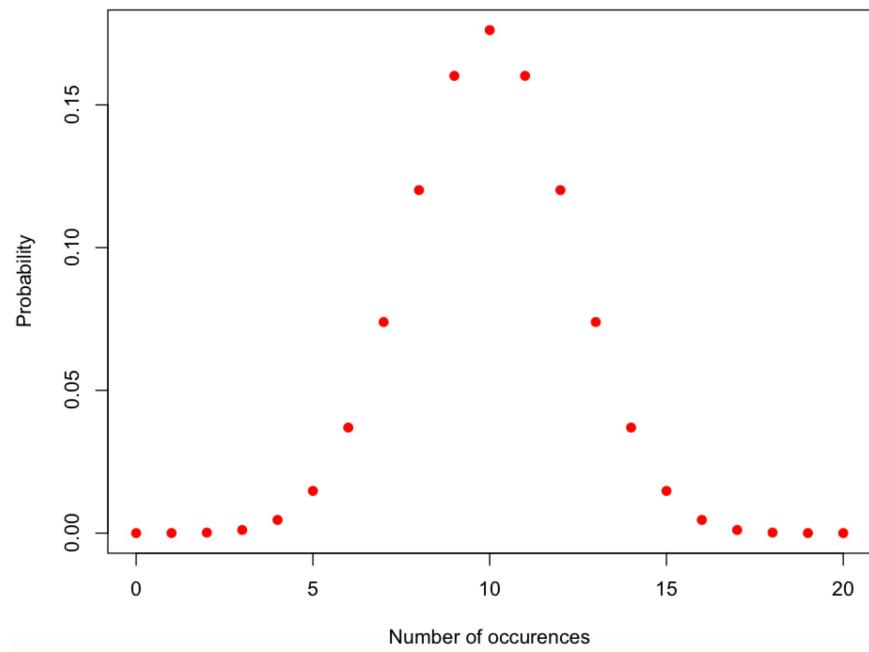
where Y is categorical variable ranging from 0 to N (0, 1, 2, ..., N)

π is the probability of success

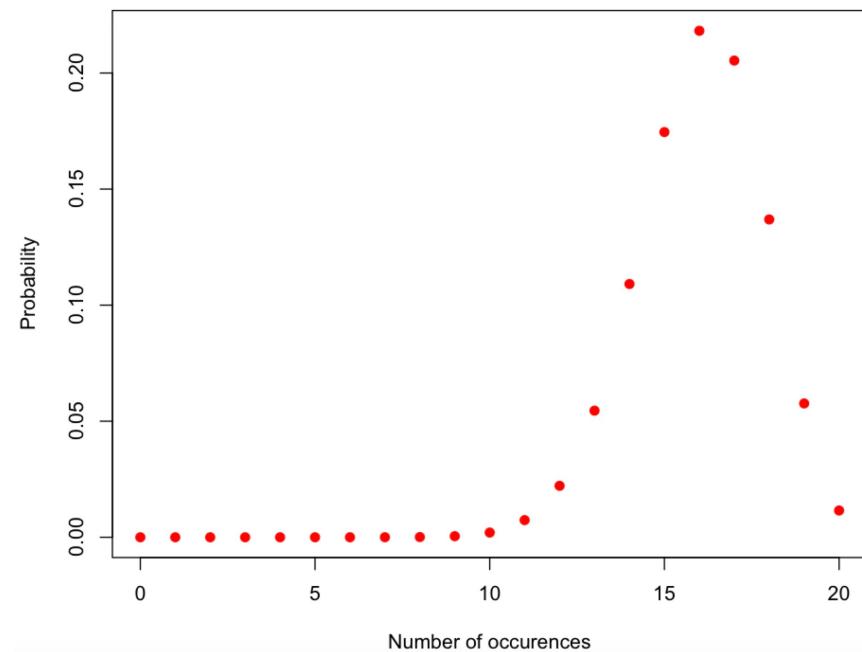
N is the sample size.

$$P(Y = m) = \binom{N}{m} \pi^m (1 - \pi)^{N-m}$$

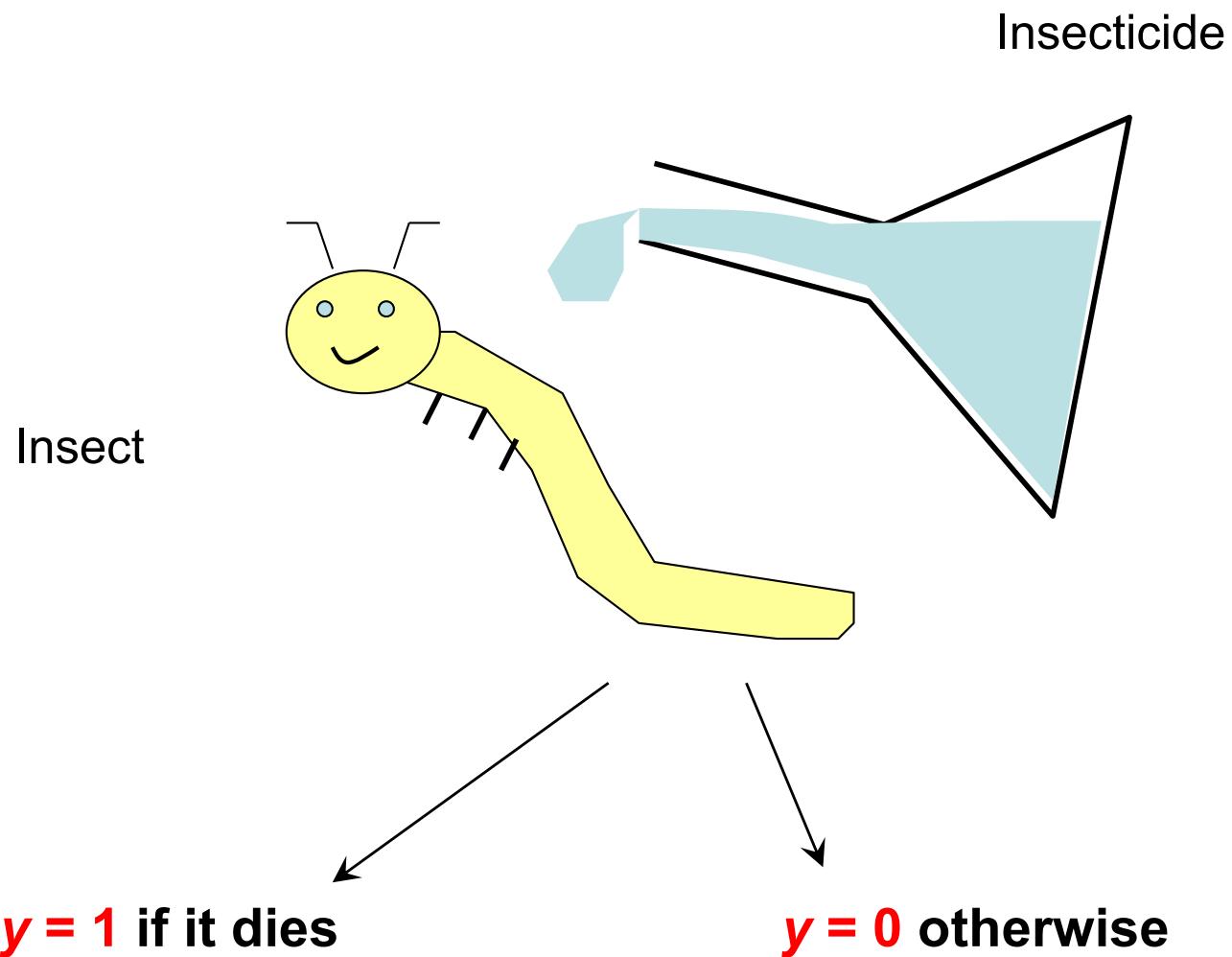
$\pi = 0.5, N = 20$



$\pi = 0.8, N = 20$



Example 1: Effect of insecticide in mortality of caterpillars



Example 1: Effect of insecticide in mortality of caterpillars

	$\text{Log}_2 (\text{Dose insecticide})$					
Sexe	0	1	2	3	4	5
M	1	4	9	13	18	20
F	0	2	6	10	12	16

Collett, 1991

Number of deaths out
of 20 individuals

Goal:

Assess the effect of insecticide on the death rate as a function of dose and sex

Models tested in example 1

Model 1.

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 x$$

Model 2.

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 sexe + \theta_2 x$$

Model 3.

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 sexe + \theta_2 x + \theta_3 sexe \times x$$

with x the value of \log_2 (dose of insecticide),
 $sexe=1$ for male
 $sexe=0$ for female.

Outline

- **Logistic and binomial models**

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

- **Poisson models**

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

Estimation of the parameters by maximum likelihood

Objective: Find the parameter values maximizing the probability of the data conditionnally to the parameters.

$$\text{Likelihood} = L(\theta_0, \theta_1, \dots, \theta_p) = \text{Prob}\left(y_1, \dots, y_N \mid \theta_0, \theta_1, \dots, \theta_p\right)$$

Likelihood for the bernouilli-logit model

$$\pi(x_{i,1}, \dots, x_{i,p}; \theta_0, \theta_1, \dots, \theta_p) = \frac{\exp(\theta_0 + \theta_1 x_{1,i} + \dots + \theta_p x_{p,i})}{1 + \exp(\theta_0 + \theta_1 x_{1,i} + \dots + \theta_p x_{p,i})}$$

$$\text{Likelihood} = \text{Prob}\left(y_1, \dots, y_N \mid \theta_0, \theta_1, \dots, \theta_p\right) = \\ \prod_{i=1}^N \pi(x_{i,1}, \dots, x_{i,p}; \theta_0, \theta_1, \dots, \theta_p)^{y_i} \left[1 - \pi(x_{i,1}, \dots, x_{i,p}; \theta_0, \theta_1, \dots, \theta_p)\right]^{1-y_i}$$

Likelihood for the bernouilli-logit model

$$\pi(x_{i,1}, \dots, x_{i,p}; \theta_0, \theta_1, \dots, \theta_p) = \frac{\exp(\theta_0 + \theta_1 x_{1,i} + \dots + \theta_p x_{p,i})}{1 + \exp(\theta_0 + \theta_1 x_{1,i} + \dots + \theta_p x_{p,i})}$$

$$\text{Likelihood} = \text{Prob}\left(y_1, \dots, y_N \mid \theta_0, \theta_1, \dots, \theta_p\right) =$$
$$\prod_{i=1}^N \pi(x_{i,1}, \dots, x_{i,p}; \theta_0, \dots, \theta_p)^{y_i} \left[1 - \pi(x_{i,1}, \dots, x_{i,p}; \theta_0, \dots, \theta_p)\right]^{1-y_i}$$

0/1 0/1

- The same type of likelihood function can be expressed for binomial models as well.
- The parameter values can be estimated by maximizing this function, using an iterative algorithm (Newton-Raphson).
- In most cases, it works well, but not always...

Models tested in example 1

Model 1.

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 x$$

Model 2.

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 sexe + \theta_2 x$$

Model 3.

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 sexe + \theta_2 x + \theta_3 sexe \times x$$

with x the value of \log_2 (dose of insecticide),
 $sexe=1$ for male
 $sexe=0$ for female.

```
glm(formula, family=binomial(link='logit'), data...)
```

Or

```
glm (formula, family=binomial, data...)
```

```
#####Data  
ldose=c(0:5, 0:5)  
sexe=factor(c(rep("M",6), rep("F", 6)))  
nombMort=c(1,4,9,13,18,20,0,2,6,10,12,16)  
nombViv=20-nombMort  
Reponse=cbind(nombMort,nombViv)
```

```
#####Option 1 (numbers)  
Modele=glm(Reponse~ldose*sexe,family=binomial)  
summary(Modele)
```

```
#####Option 2 (proportions)  
propMort=nombMort/20  
Freq=rep(20,12)
```

```
Modele=glm(propMort~ldose*sexe,family=binomial,weight=Freq)  
summary(Modele)
```

Call:

```
glm(formula = propMort ~ ldose * sexe, family = binomial, weights = Freq)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.39849	-0.32094	-0.07592	0.38220	1.10375

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.9935	0.5527	-5.416	6.09e-08 ***
ldose	0.9060	0.1671	5.422	5.89e-08 ***
sexeM	0.1750	0.7783	0.225	0.822
ldose:sexeM	0.3529	0.2700	1.307	0.191

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.8756 on 11 degrees of freedom
Residual deviance: 4.9937 on 8 degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4

```

Call:
glm(formula = propMort ~ ldose * sexe, family = binomial, weights = Freq)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-1.39849 -0.32094 -0.07592  0.38220  1.10375 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.9935    0.5527 -5.416 6.09e-08 ***
ldose        0.9060    0.1671  5.422 5.89e-08 ***
sexeM       0.1750    0.7783  0.225   0.822    
ldose:sexeM 0.3529    0.2700  1.307   0.191    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

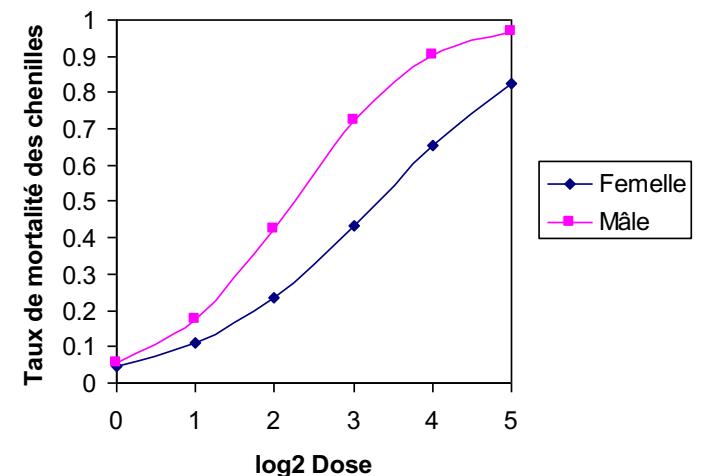
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.8756 on 11 degrees of freedom
Residual deviance: 4.9937 on 8 degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4

```

$$\pi(x_{i,1}, \dots, x_{i,p}; \theta_1, \dots, \theta_p) = \frac{\exp(\theta_0 + \theta_1 x_{1,i} + \dots + \theta_p x_{p,i})}{1 + \exp(\theta_0 + \theta_1 x_{1,i} + \dots + \theta_p x_{p,i})}$$



Odds and odds ratio

$$\log\left(\frac{\pi}{1-\pi}\right) = \theta_0 + \theta_1 x$$

$$\text{Odds} = \frac{\pi}{1-\pi} = \exp(\theta_0 + \theta_1 x)$$

$$\text{Odds ratio} = \frac{\exp(\theta_0 + \theta_1(x+1))}{\exp(\theta_0 + \theta_1 x)} = \frac{\exp(\theta_0 + \theta_1 x + \theta_1)}{\exp(\theta_0 + \theta_1 x)} = \exp(\theta_1)$$

$$\theta_1 = \log(\text{Odds ratio})$$

$\theta_1 > 0$ x increases the probability

$\theta_1 < 0$ x decreases the probability

$\theta_1 = 0$ No effect

Call:

```
glm(formula = propMort ~ ldose * sexe, family = binomial, weights = Freq)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.39849	-0.32094	-0.07592	0.38220	1.10375

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.9935	0.5527	-5.416	6.09e-08
ldose	0.9060	0.1671	5.422	5.89e-08
sexeM	0.1750	0.7783	0.225	0.822
ldose:sexeM	0.3529	0.2700	1.307	0.191

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.8756 on 11 degrees of freedom

Residual deviance: 4.9937 on 8 degrees of freedom

AIC: 43.104

Number of Fisher Scoring iterations: 4

- Positive effect of dose on the probability of death
- Positive effect of « male » on the probability of death
- Effect of dose is stronger for male

Outline

- **Logistic and binomial models**

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

- **Poisson models**

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

Call:

```
glm(formula = propMort ~ ldose * sexe, family = binomial, weights = Freq)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.39849	-0.32094	-0.07592	0.38220	1.10375

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.9935	0.5527	-5.416	6.09e-08 ***
ldose	0.9060	0.1671	5.422	5.89e-08 ***
sexeM	0.1750	0.7783	0.225	0.822
ldose:sexeM	0.3529	0.2700	1.307	0.191

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
	1			

?

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.8756 on 11 degrees of freedom
Residual deviance: 4.9937 on 8 degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4

Wald test

$$H_0: \theta = 0$$

$$z = \frac{\hat{\theta}}{\sqrt{\widehat{var}(\hat{\theta})}}$$

Parameter estimate Standard error

$$z^2 \sim \chi^2_{1ddl}$$

```
> 1-pchisq((-5.416)^2,df=1)
[1] 6.09471e-08
> 1-pchisq(5.422^2,df=1)
[1] 5.893588e-08
> 1-pchisq(0.225^2,df=1)
[1] 0.8219793
> 1-pchisq(1.307^2,df=1)
[1] 0.1912127
```

Call:

```
glm(formula = propMort ~ ldose * sexe, family = binomial, weights = Freq)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.39849	-0.32094	-0.07592	0.38220	1.10375

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.9935	0.5527	-5.416	6.09e-08 ***
ldose	0.9060	0.1671	5.422	5.89e-08 ***
sexeM	0.1750	0.7783	0.225	0.822
ldose:sexeM	0.3529	0.2700	1.307	0.191

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

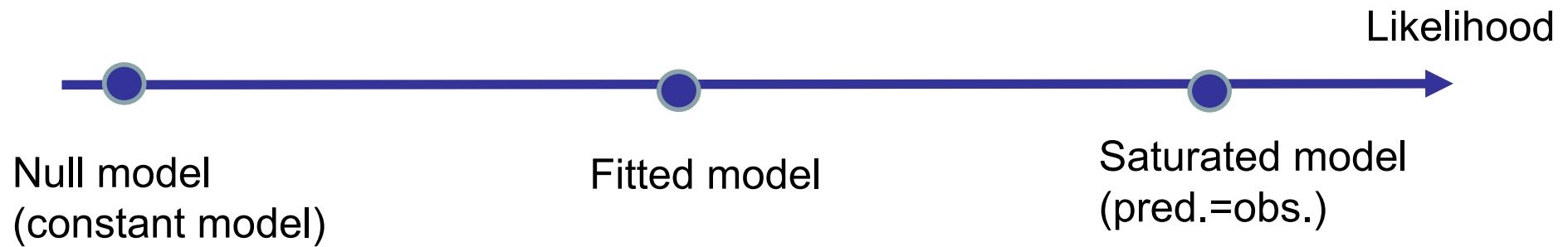
?

Null deviance: 124.8756 on 11 degrees of freedom
Residual deviance: 4.9937 on 8 degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4

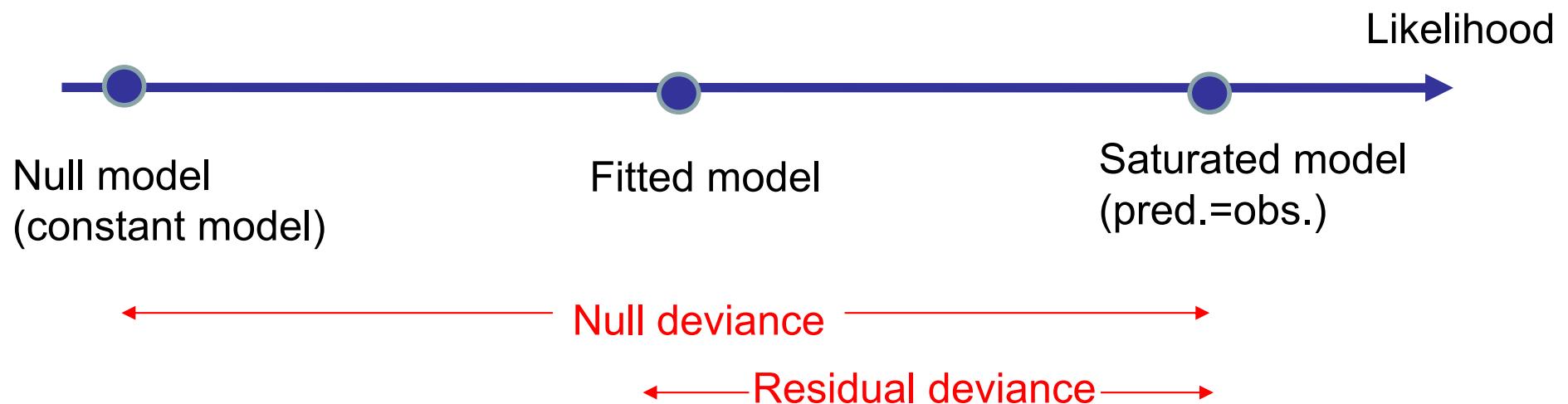
Deviance

- Measure the quality of a model compared to a « perfect » model (« saturated »)
- Generalization of the residual sum of squares
- Derived from the likelihood
- Key quantity to compare models
- Lower deviance, better fit



Null and residual deviances

- Measure the quality of a model compared to a « perfect » model (« saturated »)
- Generalization of the residual sum of squares
- Derived from the likelihood
- Key quantity to compare models
- Lower deviance, better fit



Null and residual deviances

- Measure the quality of a model compared to a « perfect » model (« saturated »)
- Generalization of the residual sum of squares
- Derived from the likelihood
- Key quantity to compare models
- Lower deviance, better fit

$$\text{Residual Deviance} = 2\{\log[\text{likelihood}(\text{Model}_{\text{saturated}})] - \log[\text{likelihood}(\text{Model}_{\text{fitted}})]\}$$

$$\text{Pseudo } R^2 = 1 - \frac{\text{Residual deviance}}{\text{Null deviance}}$$

Akaike Information Criterion (AIC)

$$-2 \log(Likelihood) + 2Number\ of\ parameters$$

Quality of fit – Penalty depending on the complexity

Call:

```
glm(formula = propMort ~ ldose * sexe, family = binomial, weights = Freq)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.39849	-0.32094	-0.07592	0.38220	1.10375

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.9935	0.5527	-5.416	6.09e-08 ***
ldose	0.9060	0.1671	5.422	5.89e-08 ***
sexeM	0.1750	0.7783	0.225	0.822
ldose:sexeM	0.3529	0.2700	1.307	0.191

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.8756 on 11 degrees of freedom
Residual deviance: 4.9937 on 8 degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4

Outline

- **Logistic and binomial models**

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

- **Poisson models**

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

Some useful variants

- Deterministic component
 - Polynomials (quadratic, cubic ...)
 - Other link functions (e.g., probit)
`glm(formula, family=binomial(link='probit'), data)`
 - Non-parametric response (R package mgcv)
- Random component
 - Add random parameters (see the course « Mixed models »)
 - Use Beta binomial instead of binomial (useful to deal with overdispersion)

Outline

- **Logistic and binomial models**

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

- **Poisson models**

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

Selection based on the likelihood ratio test

- Useful to compare two models, *Model 1* and *Model 2*, when *Model 1* is embedded in *Model 2*
- Rely on $T = \text{Deviance}(\text{Model 1}) - \text{Deviance}(\text{Model 2})$

$$T \sim \chi^2_{k \text{ ddl}}$$

where k is the number of additional parameters in *Model 2*

Models tested in example 1

Model 1.

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 x$$

Model 2.

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 \text{sex}\text{e} + \theta_2 x$$

Model 3.

$$\log \left\{ \frac{\pi}{1-\pi} \right\} = \theta_0 + \theta_1 \text{sex}\text{e} + \theta_2 x + \theta_3 \text{sex}\text{e} \times x$$

with x the value of \log_2 (dose of insecticide),
 $\text{sex}\text{e}=1$ for male
 $\text{sex}\text{e}=0$ for female.

```
#####Data  
ldose=c(0:5, 0:5)  
sexe=factor(c(rep("M",6), rep("F", 6)))  
nombMort=c(1,4,9,13,18,20,0,2,6,10,12,16)  
nombViv=20-nombMort  
Reponse=cbind(nombMort,nombViv)
```

```
Model_1=glm(Reponse~ldose,family=binomial)
```

```
Model_2=glm(Reponse~ldose+sexe,family=binomial)
```

```
Model_3=glm(Reponse~ldose*sexe,family=binomial)
```

```
anova(Model_1,Model_2, Model_3, test="LRT")
```

```
> anova(Model_1, Model_2, Model_3, test="LRT")
```

Analysis of Deviance Table

Model 1: Reponse ~ ldose

Model 2: Reponse ~ ldose + sexe

Model 3: Reponse ~ ldose * sexe

Resid.	Df	Resid.	Dev	Df	Deviance	Pr(>Chi)
--------	----	--------	-----	----	----------	----------

1	10	16.9840				
2	9	6.7571	1	10.2270	0.001384 **	
3	8	4.9937	1	1.7633	0.184209	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
> summary(Model_2)
```

Call:

```
glm(formula = Reponse ~ ldose + sexe, family = binomial)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.10540	-0.65343	-0.02225	0.48471	1.42944

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.4732	0.4685	-7.413	1.23e-13 ***
ldose	1.0642	0.1311	8.119	4.70e-16 ***
sexeM	1.1007	0.3558	3.093	0.00198 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.8756 on 11 degrees of freedom

Residual deviance: 6.7571 on 9 degrees of freedom

AIC: 42.867

Number of Fisher Scoring iterations: 4

Selection based on AIC

```
> AIC(Model_1,Model_2,Model_3)
      df      AIC
Model_1  2 51.09443
Model_2  3 42.86747
Model_3  4 43.10413
```

Outline

- Logistic and binomial models

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

- Poisson models

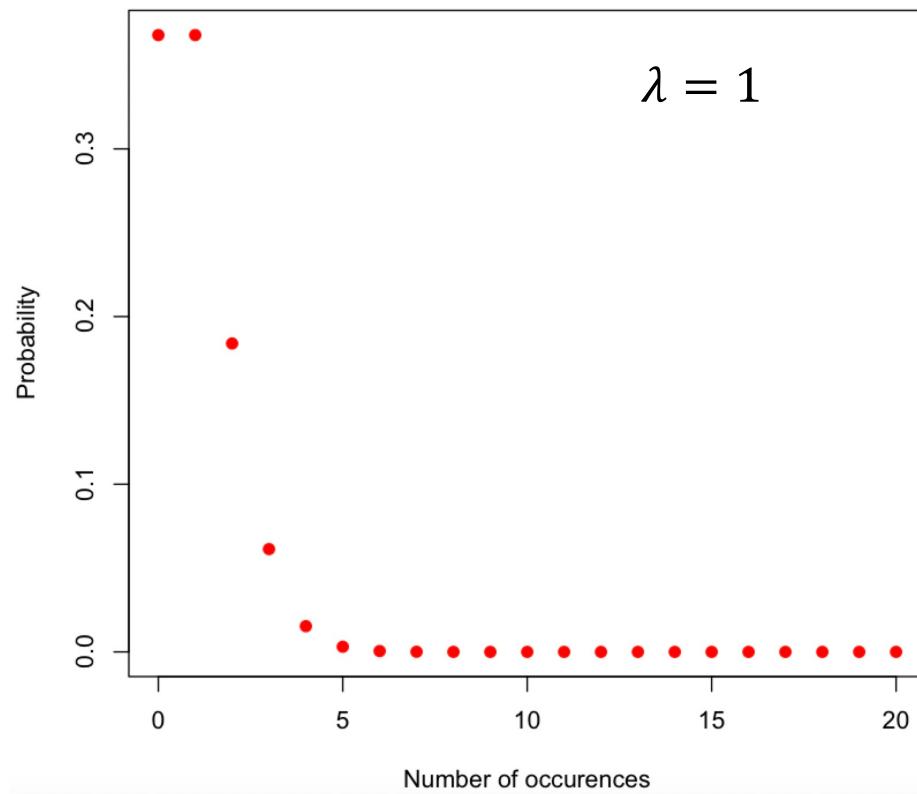
- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

Definition

Model	Link function	Distribution
Linear Gaussian	identity	Gaussian
Logistic regression Binomial model	logit	Bernouilli or Binomial
Poisson log linear	log	Poisson

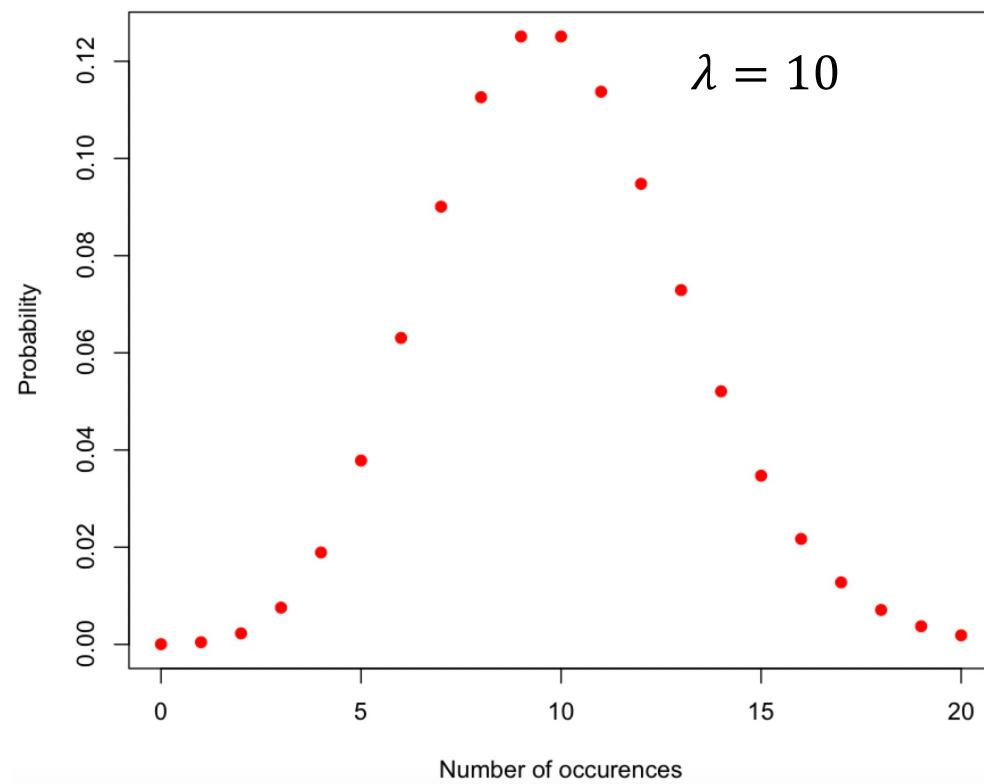
Distribution	Mean	Variance
Gaussian	μ	σ^2
Bernouilli	π	$\pi(1 - \pi)$
Binomial	$N\pi$	$N\pi(1 - \pi)$
Poisson	λ	λ

$$p(y = m | \lambda) = \exp(-\lambda) \lambda^m / m!$$



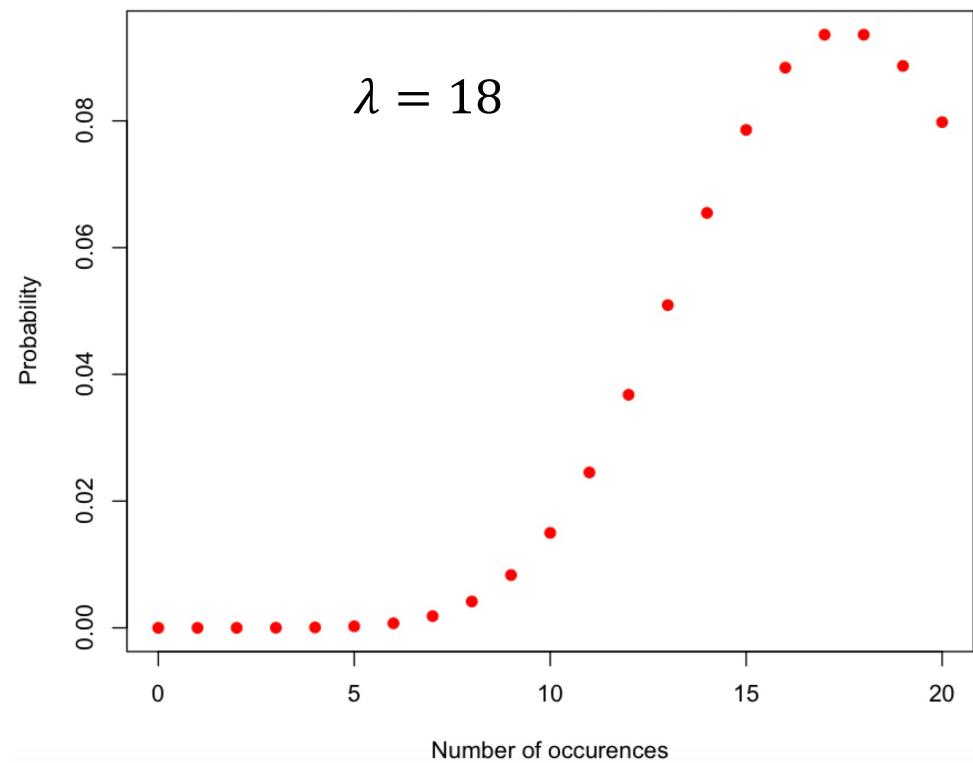
Distribution	Mean	Variance
Gaussian	μ	σ^2
Bernouilli	π	$\pi(1 - \pi)$
Binomial	$N\pi$	$N\pi(1 - \pi)$
Poisson	λ	λ

$$p(y = m | \lambda) = \exp(-\lambda) \lambda^m / m!$$



Distribution	Mean	Variance
Gaussian	μ	σ^2
Bernouilli	π	$\pi(1 - \pi)$
Binomial	$N\pi$	$N\pi(1 - \pi)$
Poisson	λ	λ

$$p(y = m | \lambda) = \exp(-\lambda) \lambda^m / m!$$



Model Poisson log linear

Deterministic component

$$g[E(y)] = \log[E(y)] = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p$$

$$E(y) = \exp(\theta_0 + \theta_1 x_1 + \dots + \theta_p x_p)$$

Random component

$$y \sim Poisson(\lambda)$$

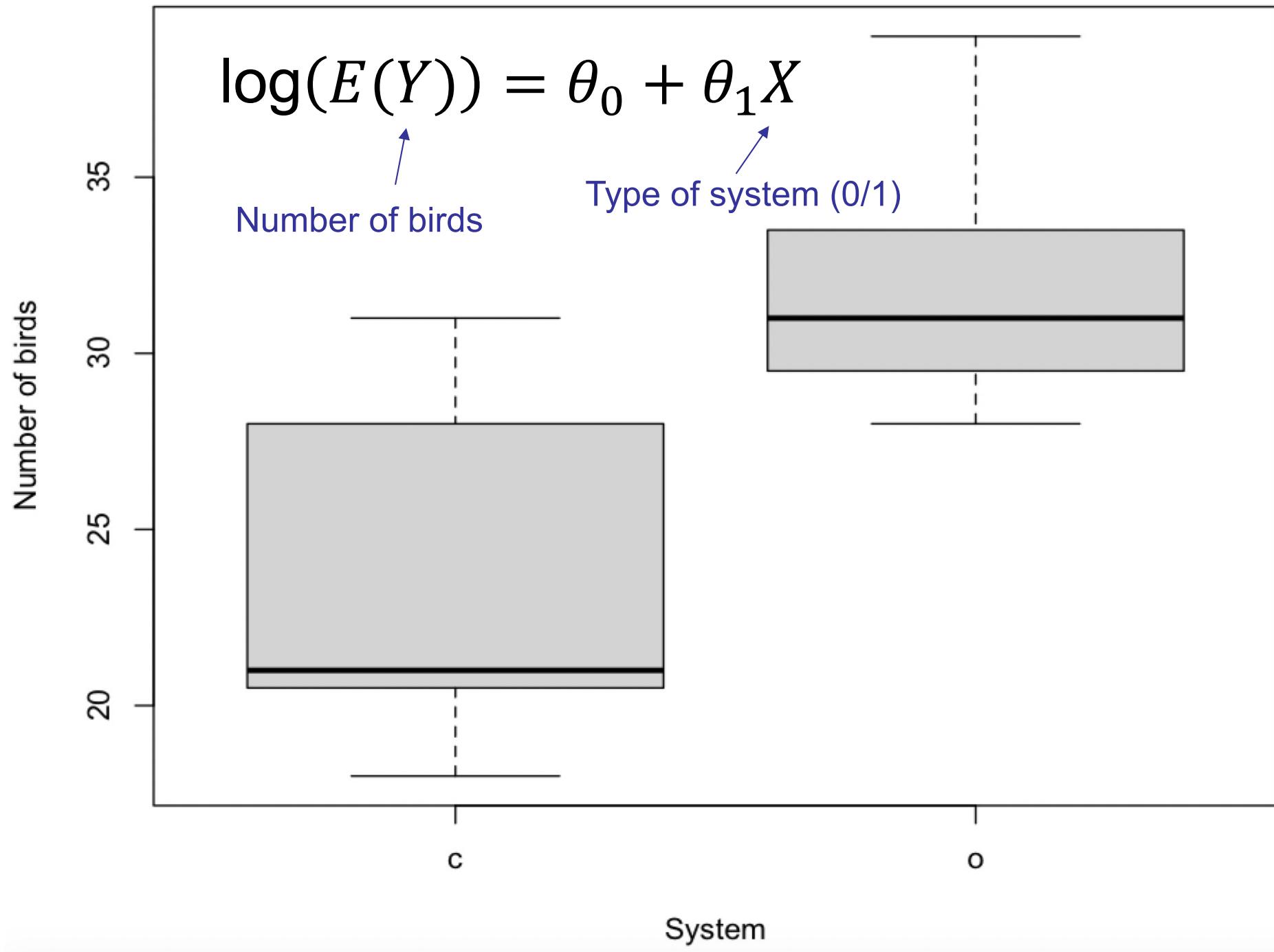
where y is a number of occurrences (ex: number of birds, number of insects, number of fires etc.)

$$E(y) = \lambda$$

$$p(y = m | \lambda) = \exp(-\lambda) \lambda^m / m!$$

Example 2: Number of birds in Organic vs. Conventional farming

	System	Birds	
1	o	30	
2	o	29	
3	o	35	Count data
4	o	32	
5	o	28	
6	o	31	
7	o	39	
8	c	20	
9	c	25	
10	c	31	
11	c	31	
12	c	18	
13	c	21	
14	c	21	

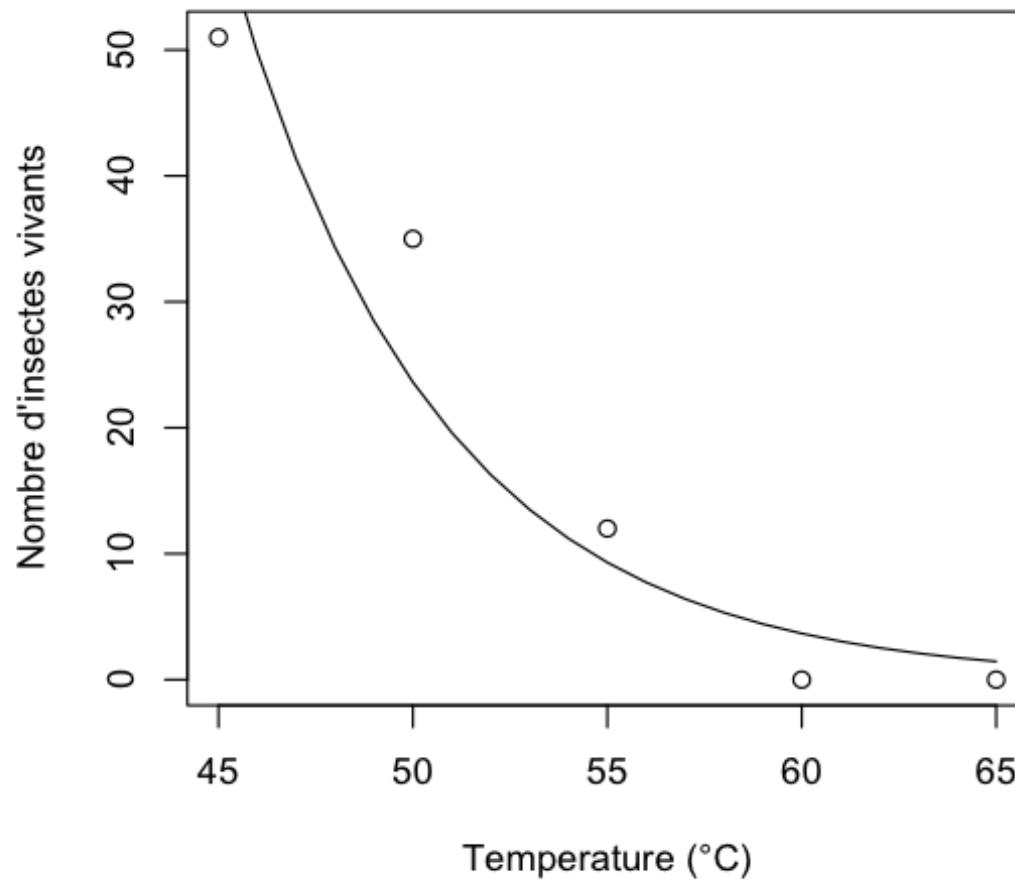


Example 3: modelling the effect of temperature on the survival of an invasive species

- We are interested in an insect pest called *Agrilus planipennis* that attacks ash wood.
- Data is available from the experiment conducted by Myers et al (2009)
Pieces of ash wood were exposed to five different temperatures for 30 min.
- The number of live insects measured on the pieces of wood after the heat treatment was equal to 51, 35, 12, 0, 0 per m² of wood for temperatures 45, 50, 55, 60, 65°C respectively

$$\log(E(Y)) = \theta_0 + \theta_1 X$$

Number of insects Temperature



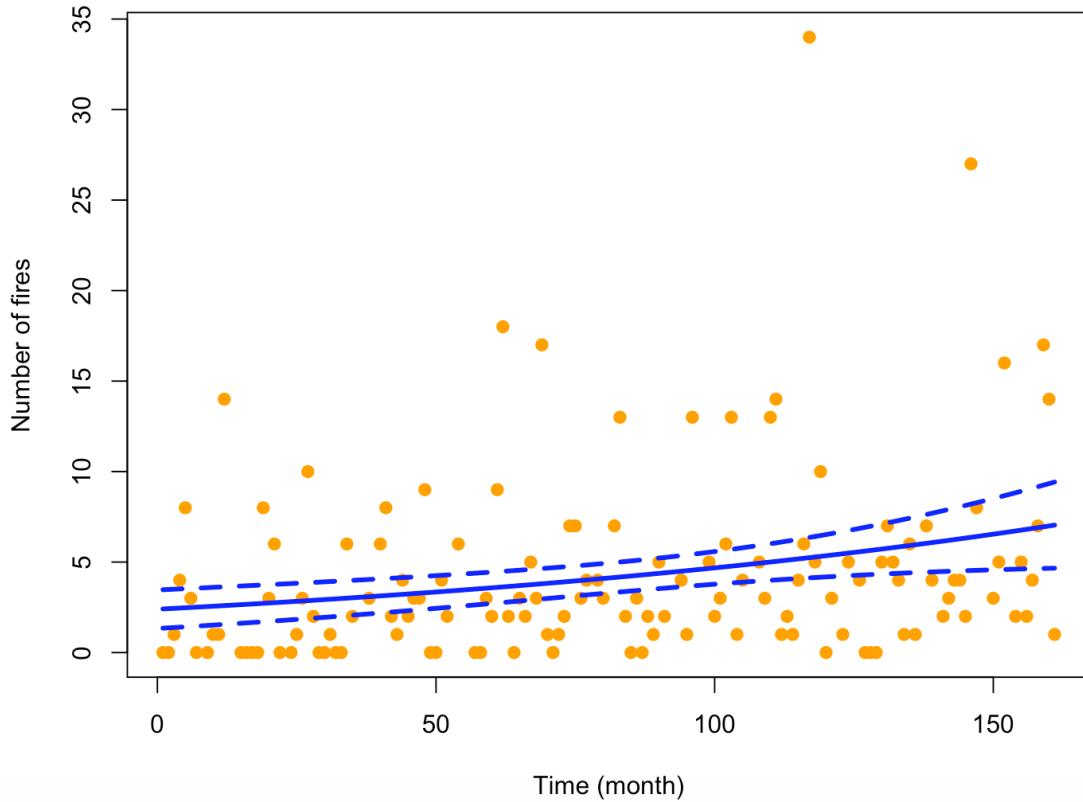
Example 4: modelling the number of fires across time

Data:

- One site in the US (county) during several years
- Observations: number of fires every month

$$\log(E(Y)) = \theta_0 + \theta_1 X$$

Number of fires Time (in months)



Outline

- Logistic and binomial models

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

- Poisson models

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

Example 2: Number of birds in Organic vs. Conventional farming

```
Birds=c(30, 29, 35, 32, 28, 31, 39, 20, 25, 31, 31, 18, 21, 21)
System=c("o", "o", "o", "o", "o", "o","o","c", "c", "c", "c", "c","c")
data.frame(System, Birds)
```

```
Mod<-glm(Birds~System, family=poisson)
summary(Mod)
```

```
predict(Mod, type="response", se.fit=T)
```

```
par(mfrow=c(1,1))
```

```
plot(Birds~as.factor(System), xlab="System", ylab="Number of birds")
```

```
points(c(1,2),c(23.65, 32), pch=19, cex=1.5, col="red")
```

```
lines(c(1,1),c(23.65-1.96*1.85, 23.65+1.96*1.85), col="red")
lines(c(2,2),c(32-1.96*2.14, 32+1.96*2.14), col="red")
```

```
> summary(Mod)
```

Call:
glm(formula = Birds ~ System, family = poisson)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.2541	-0.5973	-0.2675	0.4498	1.3973

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.17208	0.07738	40.992	< 2e-16 ***
Systemo	0.29365	0.10224	2.872	0.00408 **

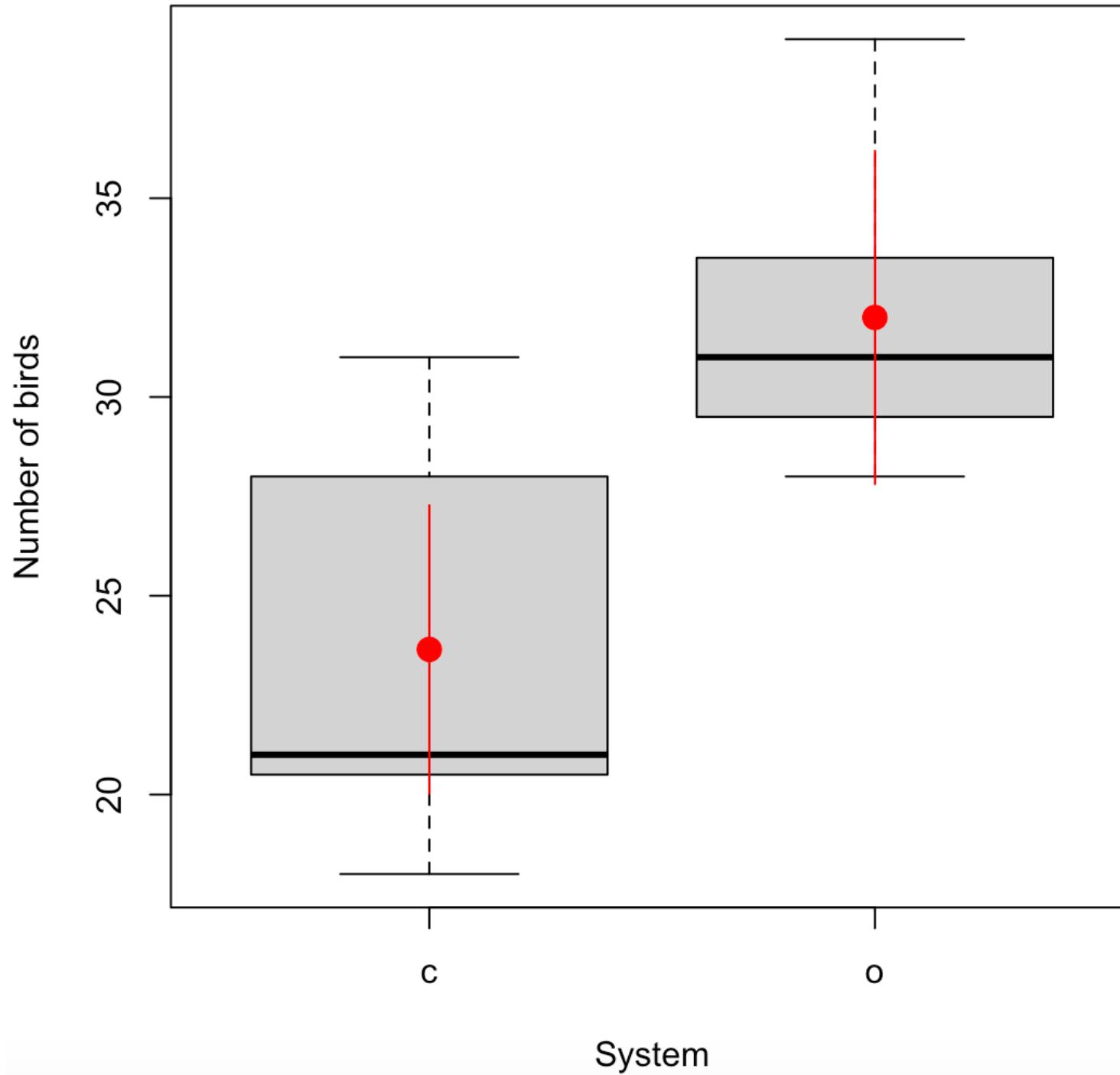
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 17.9192 on 13 degrees of freedom
Residual deviance: 9.5801 on 12 degrees of freedom
AIC: 85.679

Number of Fisher Scoring iterations: 4

```
>  
> predict(Mod, type="response", se.fit=T)  
$fit  
1 2 3 4 5 6 7 8 9  
32.00000 32.00000 32.00000 32.00000 32.00000 32.00000 32.00000 23.85714 23.85714  
10 11 12 13 14  
23.85714 23.85714 23.85714 23.85714 23.85714  
  
$se.fit  
1 2 3 4 5 6 7 8 9  
2.138090 2.138090 2.138090 2.138090 2.138090 2.138090 2.138090 1.846121 1.846121  
10 11 12 13 14  
1.846121 1.846121 1.846121 1.846121 1.846121
```



Example 3: modelling the effect of temperature on the survival of an invasive species

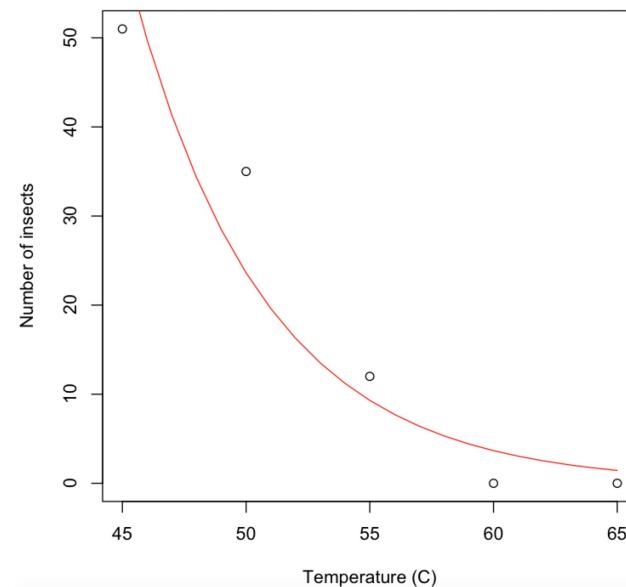
```
X = c(45, 50, 55, 60, 65)
```

```
Y = c(51, 35, 12, 0, 0)
```

```
Model <- glm(Y~X, family=poisson)  
summary(Model)
```

```
plot(X,Y, xlab="Temperature (C)",ylab="Number of insects")  
lines(45:65,exp(coef(Model)[1]+coef(Model)[2]*45:65), col="red")
```

```
Call:  
glm(formula = Y ~ X, family = poisson)  
  
Deviance Residuals:  
    1      2      3      4      5  
-1.1886  2.1833  0.8455 -2.7074 -1.6992  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) 12.47853   1.06901 11.673  <2e-16 ***  
X           -0.18633   0.02217 -8.406  <2e-16 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
(Dispersion parameter for poisson family taken to be 1)  
  
Null deviance: 126.355 on 4 degrees of freedom  
Residual deviance: 17.112 on 3 degrees of freedom  
AIC: 36.62  
  
Number of Fisher Scoring iterations: 5
```



Example 4: modelling the number of fires across time

```
> summary(Mod)
```

Call:

```
glm(formula = Y ~ X, family = "poisson", data = DATA)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.3712	-2.1217	-0.7257	0.3497	8.3430

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.8729387	0.0958008	9.112	< 2e-16 ***
X	0.0067010	0.0009095	7.367	1.74e-13 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

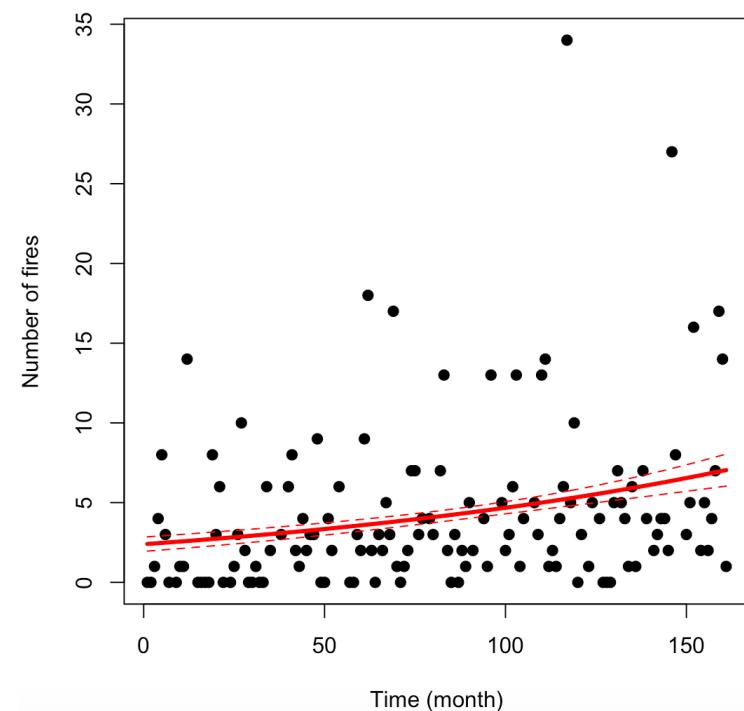
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 671.24 on 135 degrees of freedom

Residual deviance: 615.46 on 134 degrees of freedom

AIC: 972.87

Number of Fisher Scoring iterations: 5



Outline

- Logistic and binomial models

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

- Poisson models

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

Test and deviance

Similar to the binomial model

- Wald test
- Null deviance, residual deviance
- Likelihood ratio test
- AIC

Special attention should be given to overdispersion

Overdispersion

What is it?

- Poisson models assume that the mean and the variance are equal
- This assumption is often wrong: Variance > Mean

How can we detect it?

- Compare the residual deviance to the ddl
- *Residual deviance >> ddl* indicates overdispersion

What can we do?

- Adapt the model by adding an extra-parameter
- Change the probability distribution

Poisson  Negative binomial

Distribution	Mean	Variance
Gaussian	μ	σ^2
Bernouilli	π	$\pi(1 - \pi)$
Binomial	$N\pi$	$N\pi(1 - \pi)$
Poisson	λ	λ
Negative binomial	$r (1 - \pi)/\pi$	$r (1 - \pi)/\pi^2$

Example 4: modelling the number of fires across time

```
> summary(Mod)
```

Call:

```
glm(formula = Y ~ X, family = "poisson", data = DATA)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.3712	-2.1217	-0.7257	0.3497	8.3430

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.8729387	0.0958008	9.112	< 2e-16 ***
X	0.0067010	0.0009095	7.367	1.74e-13 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

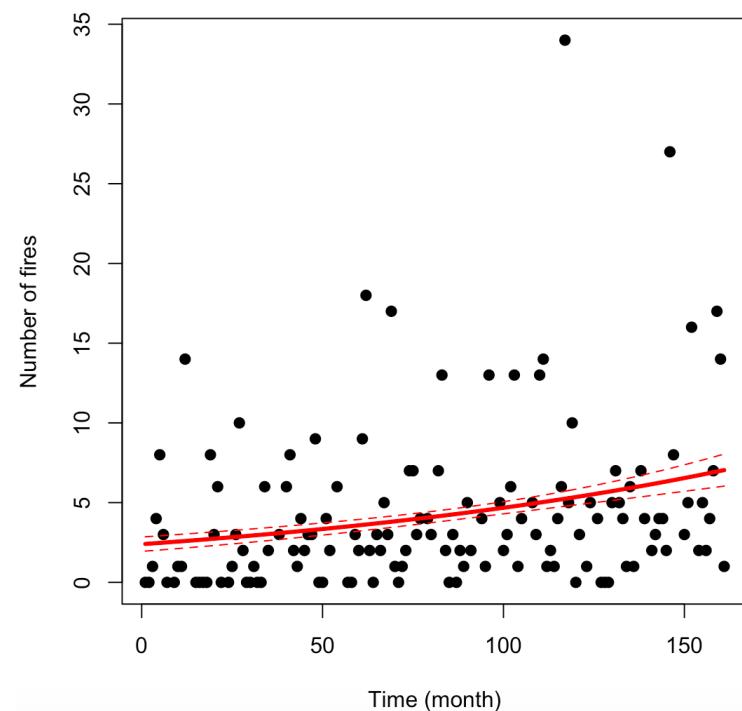
Null deviance: 671.24 on 135 degrees of freedom

Residual deviance: 615.46 on 134 degrees of freedom

AIC: 972.87

Number of Fisher Scoring iterations: 5

$$\frac{615}{134} = 4,5$$



Quasi-poisson

Variance = K x Mean

```
> summary(Mod)
```

Call:
glm(formula = Y ~ X, family = "quasipoisson", data = DATA)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.3712	-2.1217	-0.7257	0.3497	8.3430

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.872939	0.226930	3.847	0.000184 ***
X	0.006701	0.002154	3.110	0.002284 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

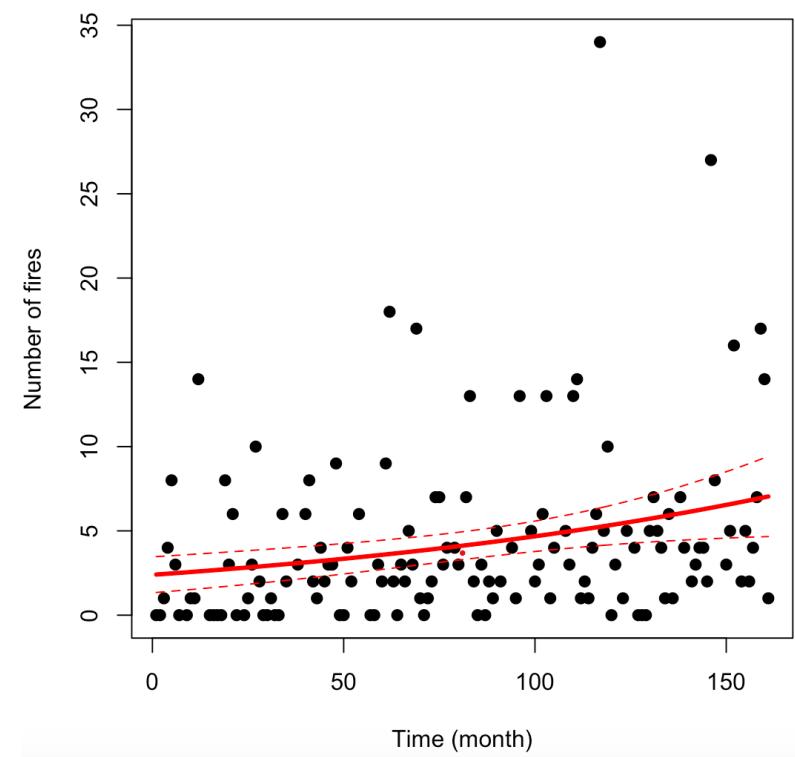
(Dispersion parameter for quasipoisson family taken to be 5.611049)

Null deviance: 671.24 on 135 degrees of freedom

Residual deviance: 615.46 on 134 degrees of freedom

AIC: NA

Number of Fisher Scoring iterations: 5



Distribution	Mean	Variance
Gaussian	μ	σ^2
Bernouilli	π	$\pi(1 - \pi)$
Binomial	$N\pi$	$N\pi(1 - \pi)$
Poisson	λ	λ

Distribution	Mean	Variance
Gaussian	μ	σ^2
Bernouilli	π	$\pi(1 - \pi)$
Binomial	$N\pi$	$N\pi(1 - \pi)$
Poisson	λ	λ
Negative binomial	$r (1 - \pi)/\pi$	$r (1 - \pi)/\pi^2$

Negative binomial

```
library(MASS)
```

```
Model_nb<-glm.nb(Y~X, data=DATA)
```

```
summary(Model_nb)
```

```
> summary(Model_nb)
```

Call:

```
glm.nb(formula = Y ~ X, data = DATA, init.theta = 1.102557845,  
       link = log)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0048	-1.0392	-0.3590	0.1711	2.6435

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.858679	0.190632	4.504	6.66e-06 ***
X	<u>0.006864</u>	0.001988	3.453	0.000555 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.1026) family taken to be 1)

Null deviance: 165.29 on 135 degrees of freedom

Residual deviance: 153.57 on 134 degrees of freedom

AIC: 693.88

Number of Fisher Scoring iterations: 1

Theta: 1.103

Std. Err.: 0.181

2 x log-likelihood: -687.884

Outline

- Logistic and binomial models

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

- Poisson models

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

Some useful variants

- Deterministic component
 - Polynomials
 - Non-parametric response (mgcv)
 - Offset
- Random component
 - Negative binomial
 - Zero-inflated
 - Additional random effects (mixed models)

Offset

- Useful when the count data are collected at different scales
- Can be easily included in a model using `glm()`

Example

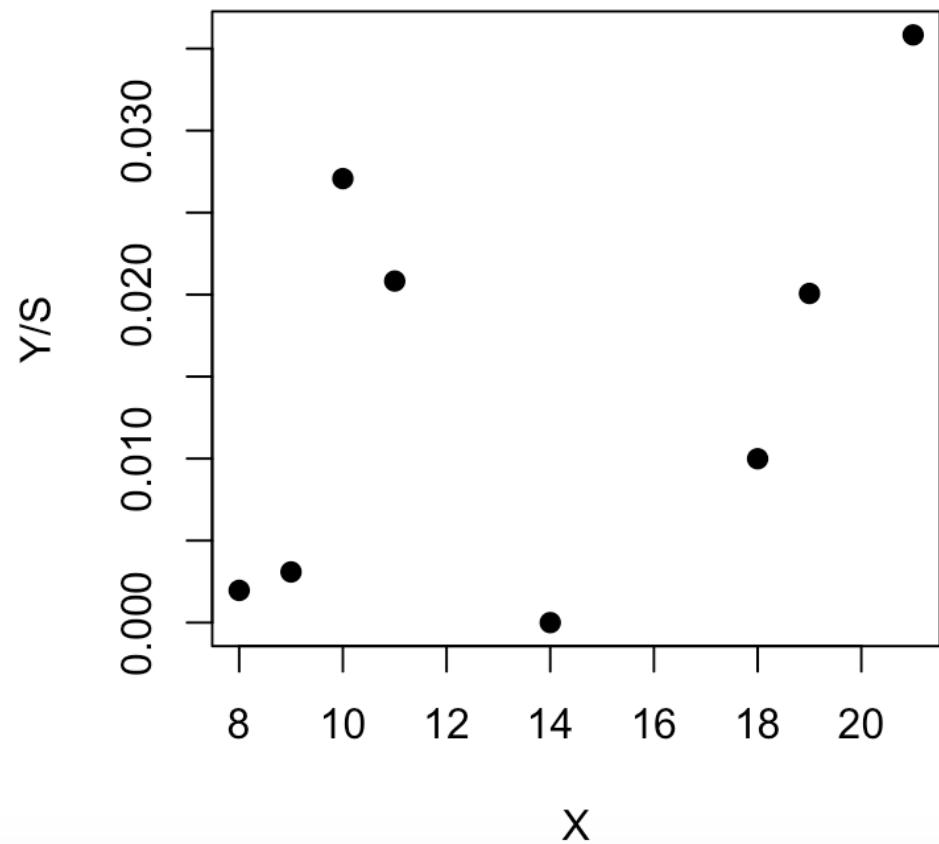
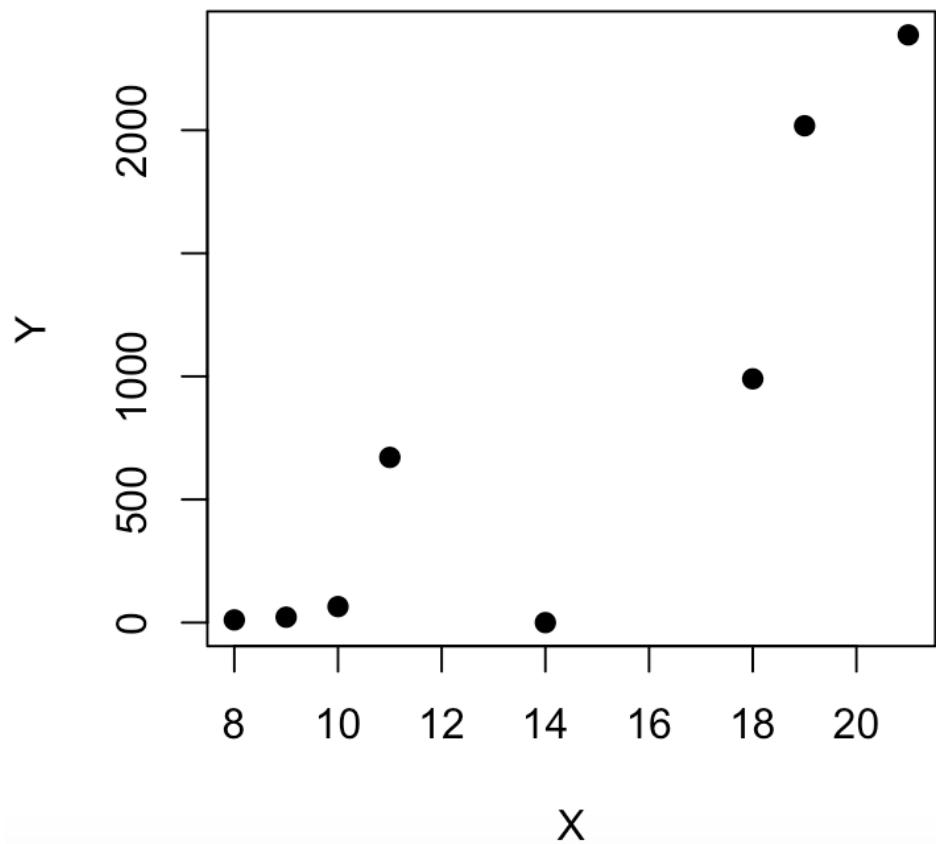
We want to estimate the effect of temperature on the number of covid19 cases in cities of different sizes.

We have N observations of covid19-case numbers (Y) in N cities of different sizes S and temperatures X

$$\log[E(Y)] = \log(S) + \theta_0 + \theta_1 X$$
$$E(Y) = S \times \exp(\theta_0 + \theta_1 X)$$

```
S=c(1100, 5602, 100551, 99090, 32218, 2401, 7123,66612)  
X=c(14, 8, 19, 18, 11, 10, 9, 21)  
Y=c(0, 11, 2018, 990, 671, 65, 22, 2387)
```

```
par(mfrow=c(1,2))  
plot(X,Y, pch=19)  
plot(X,Y/S, pch=19)
```



```
Mod_1=glm(Y~X, family=poisson)  
summary(Mod_1)
```

```
Mod_2=glm(Y~X+ offset(log(S)), family=poisson)  
summary(Mod_2)
```

```
Mod_3=glm(Y~X, family=quasipoisson)  
summary(Mod_3)
```

```
Mod_4=glm(Y~X+offset(log(S)), family=quasipoisson)  
summary(Mod_4)
```

$$\log[E(Y)] = \log(S) + \theta_0 + \theta_1 X$$

```
> summary(Mod_1)

Call:
glm(formula = Y ~ X, family = poisson)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-28.1721 -9.6856 -5.9444 -0.0256 28.2420 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 2.228884  0.075293 29.60   <2e-16 ***  
X           0.268188  0.003991 67.19   <2e-16 ***  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 9023.5 on 7 degrees of freedom
Residual deviance: 2018.7 on 6 degrees of freedom
AIC: 2074

Number of Fisher Scoring iterations: 6
```

```
> Mod_3=glm(Y~X, family=quasipoisson)
> summary(Mod_3)

Call:
glm(formula = Y ~ X, family = quasipoisson)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-28.1721 -9.6856 -5.9444 -0.0256 28.2420 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.22888  1.42066  1.569   0.1677    
X           0.26819  0.07531  3.561   0.0119 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 356.0194)

Null deviance: 9023.5 on 7 degrees of freedom
Residual deviance: 2018.7 on 6 degrees of freedom
AIC: NA
```

```
> Mod_2=glm(Y~X+ offset(log(S)), family=poisson)
> summary(Mod_2)

Call:
glm(formula = Y ~ X + offset(log(S)), family = poisson)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-23.050 -5.722 -3.762  9.556  16.632 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.591646  0.091655 -61.01   <2e-16 ***  
X           0.091033  0.004882  18.65   <2e-16 ***  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
```

```
Null deviance: 1641.5 on 7 degrees of freedom
Residual deviance: 1224.8 on 6 degrees of freedom
AIC: 1280.1
```

Number of Fisher Scoring iterations: 5

```
> Mod_4=glm(Y~X+offset(log(S)), family=quasipoisson)
> summary(Mod_4)

Call:
glm(formula = Y ~ X + offset(log(S)), family = quasipoisson)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-23.050 -5.722 -3.762  9.556  16.632 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.59165  1.32085 -4.233  0.00548 **  
X           0.09103  0.07035  1.294  0.24325    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 207.6794)
```

```
Null deviance: 1641.5 on 7 degrees of freedom
Residual deviance: 1224.8 on 6 degrees of freedom
AIC: NA
```

Number of Fisher Scoring iterations: 5

Outline

- Logistic and binomial models

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

- Poisson models

- Definition
- Estimation
- Test and deviance
- Useful variants
- Model selection

Model selection

- Deviance
- Likelihood ratio test
- AIC
- Check overdispersion

Summary

- Specify the objective of your analysis
- Look at you data
 - With yes/no data, try Bernouilli models
 - With proportions, try Binomial models
 - With count data, try Poisson models
- Define several models including 0, 1, 2, ..., inputs
- Fit the models
- Test the different variants
- Check overdispersion
- Interpret results and conclude

Exercise

Objective:

Assess whether the probability of presence of the bird *Tringa totanus* is higher in grassland with high vs. low grass height

```
> DATA
```

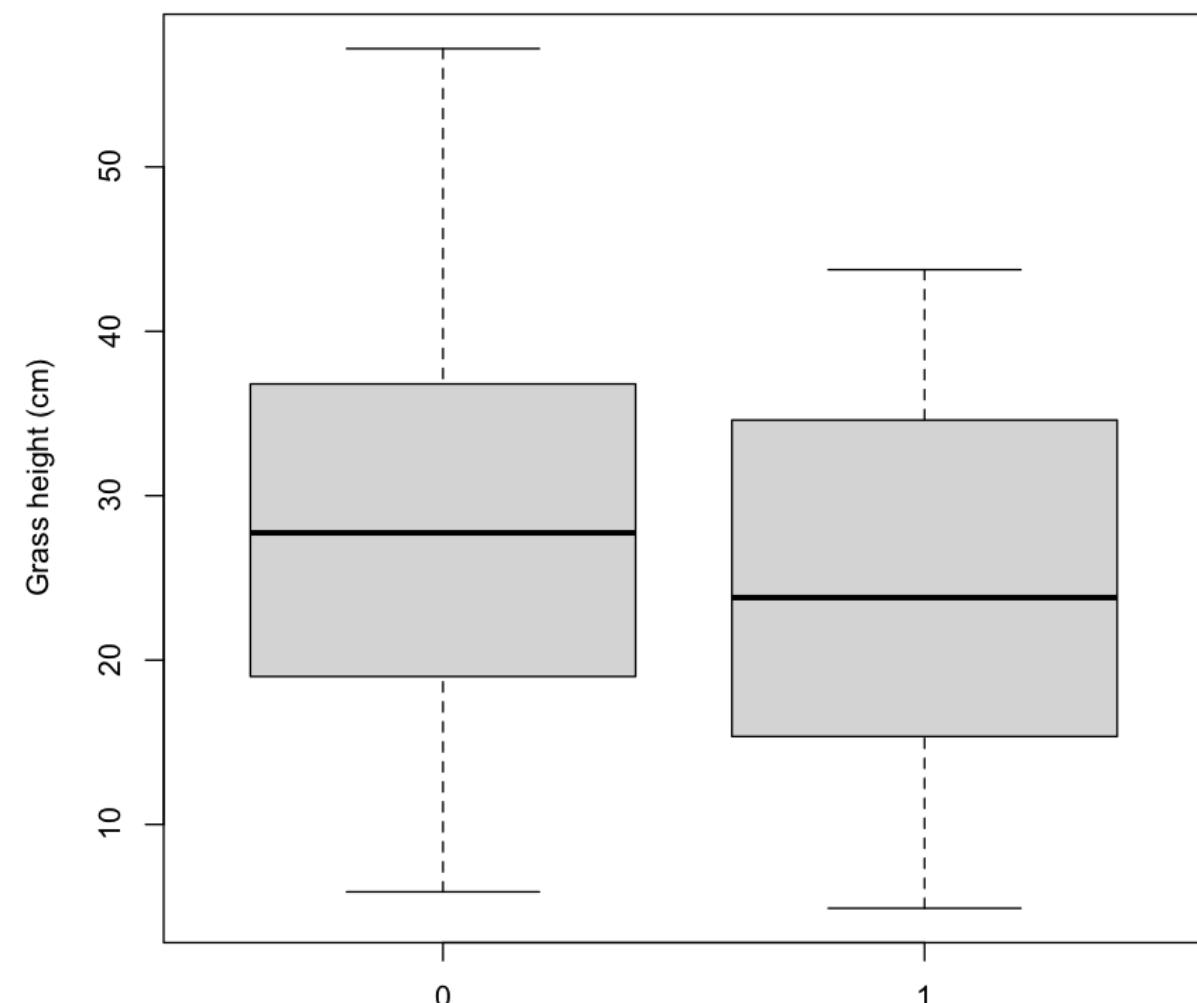
	Presence	HauteurHerbe
1	0	48.80
2	0	15.70
3	0	15.40
4	0	22.80
5	1	30.40
6	1	34.30
7	0	31.30
8	1	12.80
9	1	4.90
10	0	26.30
11	0	10.40
12	0	42.90
13	0	24.20
14	0	16.70
15	0	27.70
16	0	33.30

```
> dim(DATA)
```

```
[1] 424 2
```

```
> summary(DATA)
```

Presence	HauteurHerbe
Min. :0.0000	Min. : 4.90
1st Qu.:0.0000	1st Qu.:18.70
Median :0.0000	Median :27.00
Mean :0.1203	Mean :27.89
3rd Qu.:0.0000	3rd Qu.:36.48
Max. :1.0000	Max. :57.20



Questions

1. Define one or several models
2. How to estimate the model parameters?

```
> Mod<-glm(Presence~HauteurHerbe, data=DATA, family="binomial")
> summary(Mod)
```

Call:
glm(formula = Presence ~ HauteurHerbe, family = "binomial", data = DATA)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.6797	-0.5541	-0.4730	-0.4102	2.2831

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)							
(Intercept)	-1.16333	0.38712	-3.005	0.00265 **							
HauteurHerbe	-0.03123	0.01423	-2.195	0.02813 *							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 311.63 on 423 degrees of freedom
Residual deviance: 306.59 on 422 degrees of freedom
AIC: 310.59

Number of Fisher Scoring iterations: 5

```
> Mod0<-glm(Presence~1, data=DATA, family="binomial")
> summary(Mod0)
```

Call:
glm(formula = Presence ~ 1, family = "binomial", data = DATA)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.5063	-0.5063	-0.5063	-0.5063	2.0581

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)							
(Intercept)	-1.9898	0.1493	-13.33	<2e-16 ***							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

(Dispersion parameter for binomial family taken to be 1)

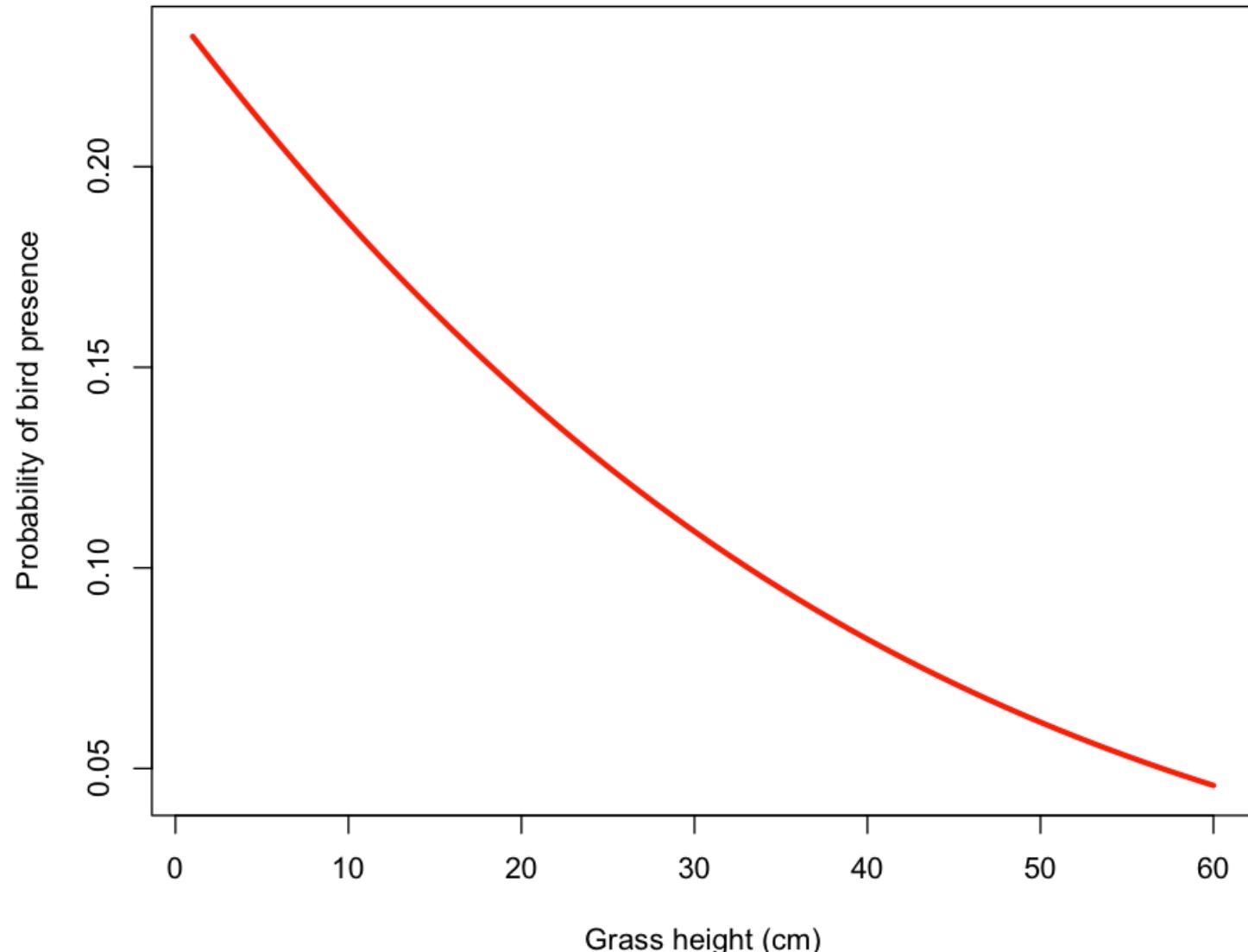
Null deviance: 311.63 on 423 degrees of freedom
Residual deviance: 311.63 on 423 degrees of freedom
AIC: 313.63

Number of Fisher Scoring iterations: 4

Questions

1. Define one or several models
2. How to estimate the model parameters?
3. Is the effect of grass height significant?
4. Is its effect positive or negative on bird presence probability?

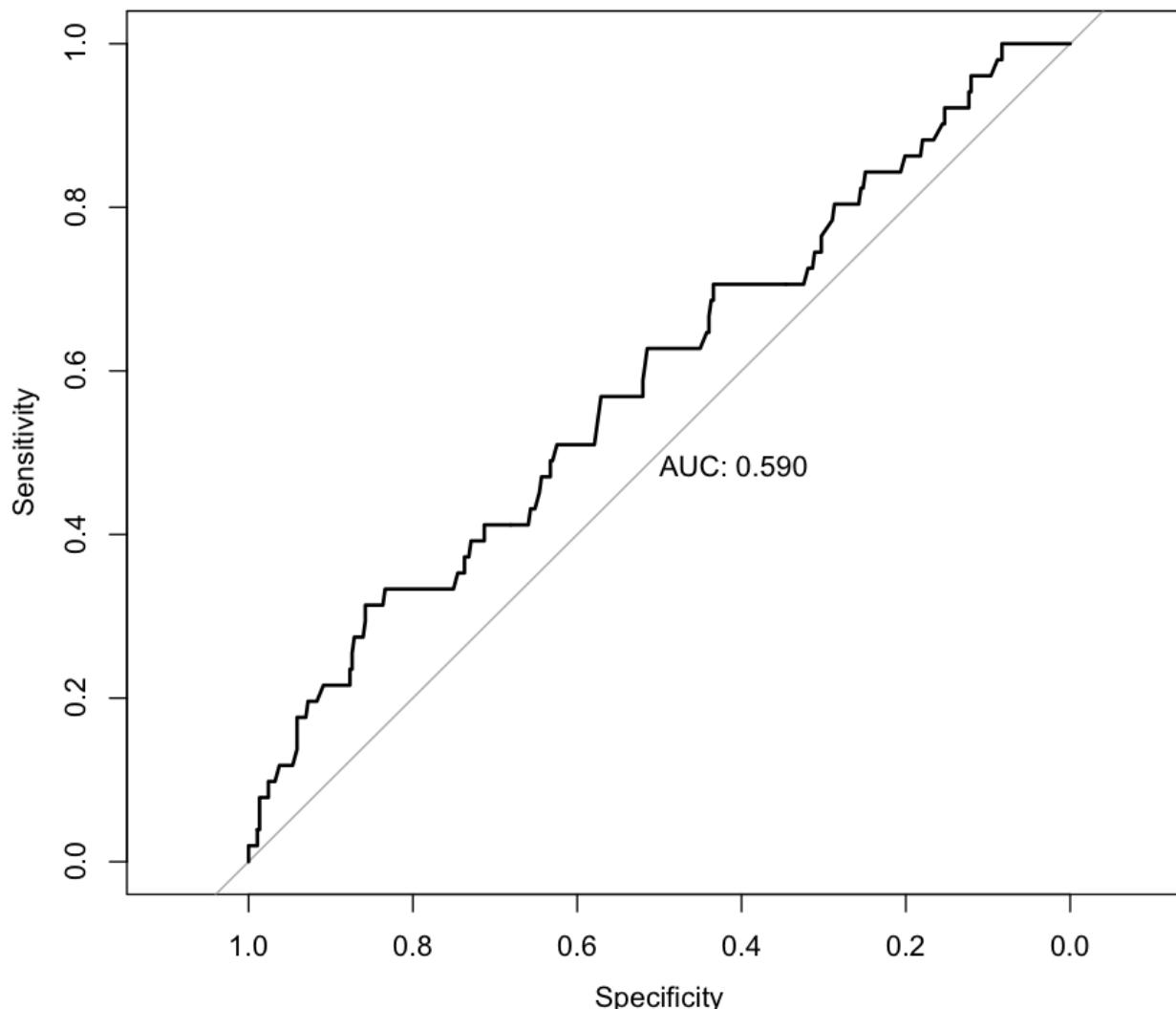
```
Proba<-predict(Mod, newdata=data.frame(HauteurHerbe=1:60),type="response")
plot(1:60, Proba, xlab="Grass height (cm)", ylab="Probability of bird presence", type="l", lwd=3, col="red")
```



Questions

1. Define one or several models
2. How to estimate the model parameters?
3. Is the effect of grass height significant?
4. Is its effect positive or negative on bird presence probability?
5. What is the probability of bird presence for a grass height of 30cm?

```
library(pROC)
Pred=predict(Mod, type="response")
roc1=roc(DATA$Presence~Pred)
plot(roc1, print.auc=T)
```



Questions

1. Define one or several models
2. How to estimate the model parameters?
3. Is the effect of grass height significant?
4. Is its effect positive or negative on bird presence probability?
5. What is the probability of bird presence for a grass height of 30cm?
6. Is the model reliable for predicting bird presence?