STATISTICAL MODELS, AN INTRODUCTION

David Makowski
INRA

Paris, France, 2015

Outline

- What is a statistical model?
- Why are statistical models useful?
- Types of statistical models, and R functions
- Generalized linear models
- Main steps for developping a statistical model

What is a statistical model?

- A statistical model is a special type of mathematical model
- It includes both observable and unobservable quantities
  - Unobservable quantities are the model parameters and the model hidden variables
- Some of these quantities are defined as random variables
- A statistical model allows one to solve specific problems based on a dataset

Linear regression model

Output/Response variable
(observable)

$$Y = \alpha_0 + \alpha_1 X + \varepsilon$$

Input variable (observable)

Linear regression model

$$Y = \alpha_0 + \alpha_1 X + \varepsilon$$

Parameters (unobservable)

Linear regression model

$$Y = \alpha_0 + \alpha_1 X + \varepsilon$$

Residual error (unobservable)

Why are statistical models useful?

- Statistical models can be used to solve different problems:
  - Estimate the value of a parameter of interest
  - Compare the value of a parameter with another value
  - Predict the value of a variable
  - Analyze uncertainty in parameter estimation and model prediction

Examples of questions

- Is **pesticide A more efficient than pesticide B** to control a given disease?
- **What are the main factors** (soil, climate, practices) influcencing the incidence and severity of a given disease?
- Is it possible **to predict the level of weed infestation** in a crop in function of cropping practices, soil, and climate?
- **How many insects** may survive a heat treatment of wood materials?

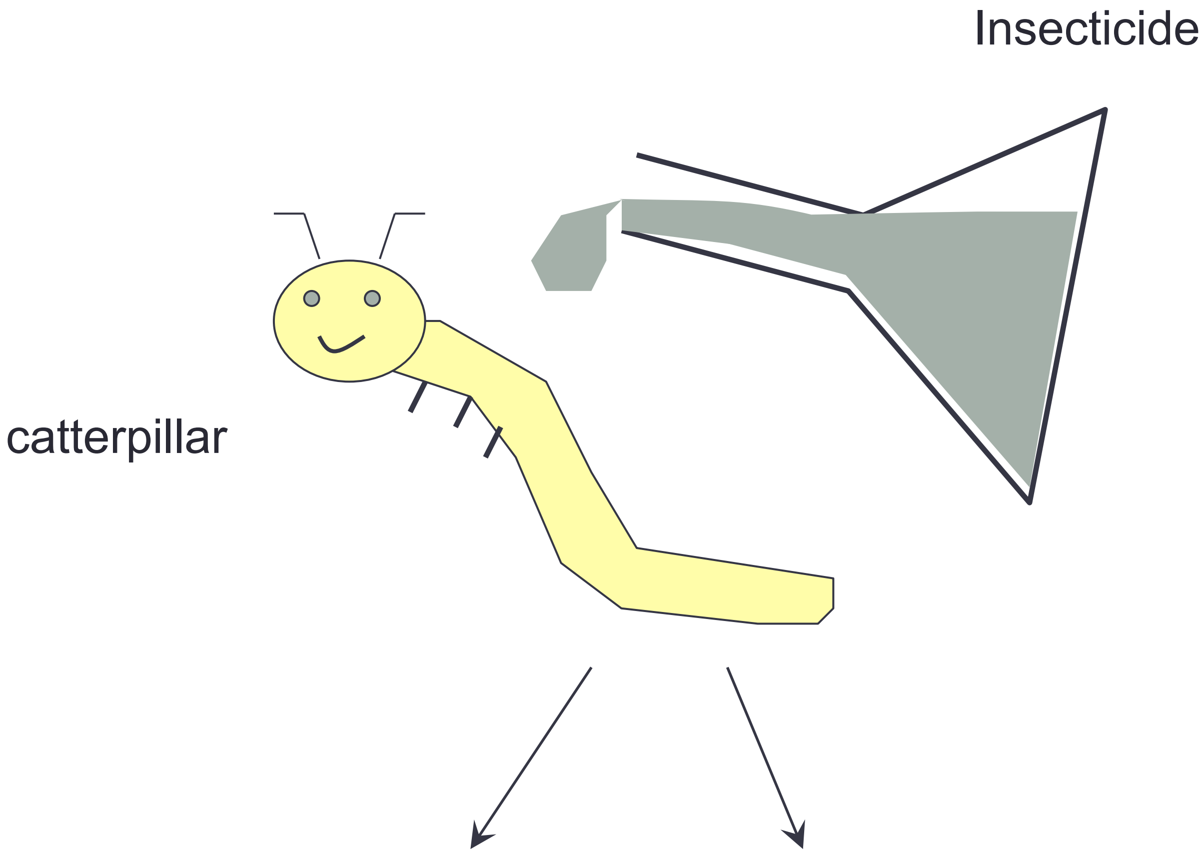
Types of statistical models, and R functions

- A great diversity of statistical models exist
- Different types of statistical models can be defined based on:
  - the type of data used for their development
  - the model equation
  - the assumptions made on the residual error
  - the method used for parameter estimation

Types of data frequently used in plant health studies

Type of data	Example
Continuous	Crop yield, yield loss, disease incidence
Binary	Presence/Absence of a disease
Categorical with more than 2 levels	Low, medium, high severity
Count	Number of insects, number of weed plants
Repeated measurements	Disease incidence measured for different tretaments applied on the same plots

Example 1: Binary data



$Y = 1$ **if the caterpillar dies** $Y = 0$ **otherwise**

Example 1 (continued)

Number of dead caterpillars

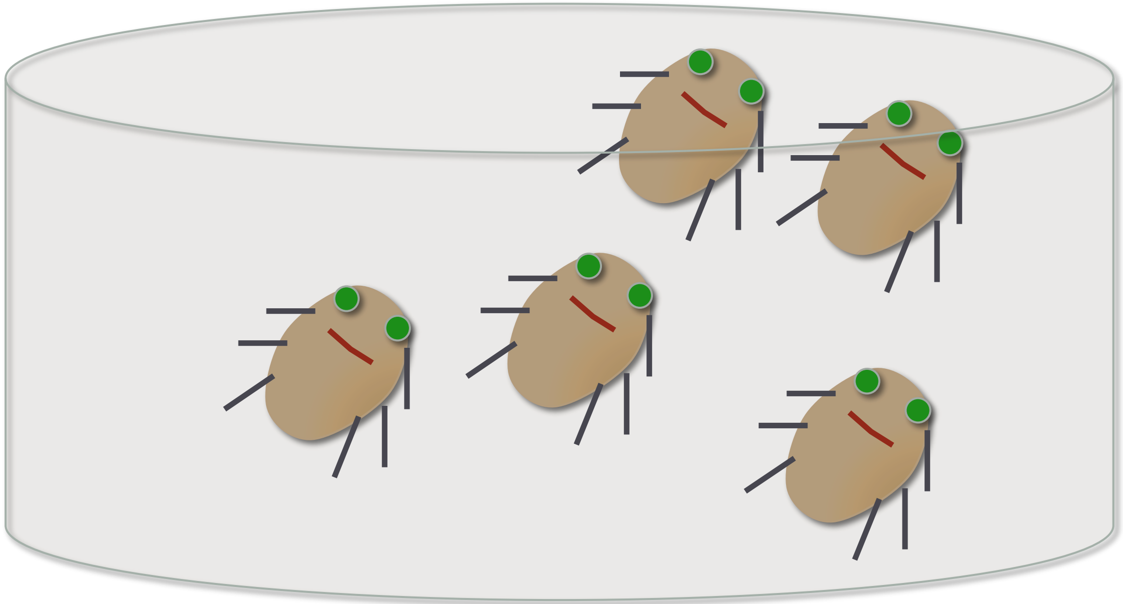
(group size = 20 caterpillars per dose and per sex)

	Insecticide dose					
Sexe	0	1	2	3	4	5
M	1	4	9	13	18	20
F	0	2	6	10	12	16

Collett, 1991

Example 2 (Count data)

Heat treatment of an infested piece of wood



Y = 0, 1, 2, 3, 4, 5… **surviving insects**

Example 2 (continued)

- We consider the experiment of Myers et al. (2009) on the effect of heat treatment on the insect species *Agrilus planipennis*, a pest of ash (a tree species)

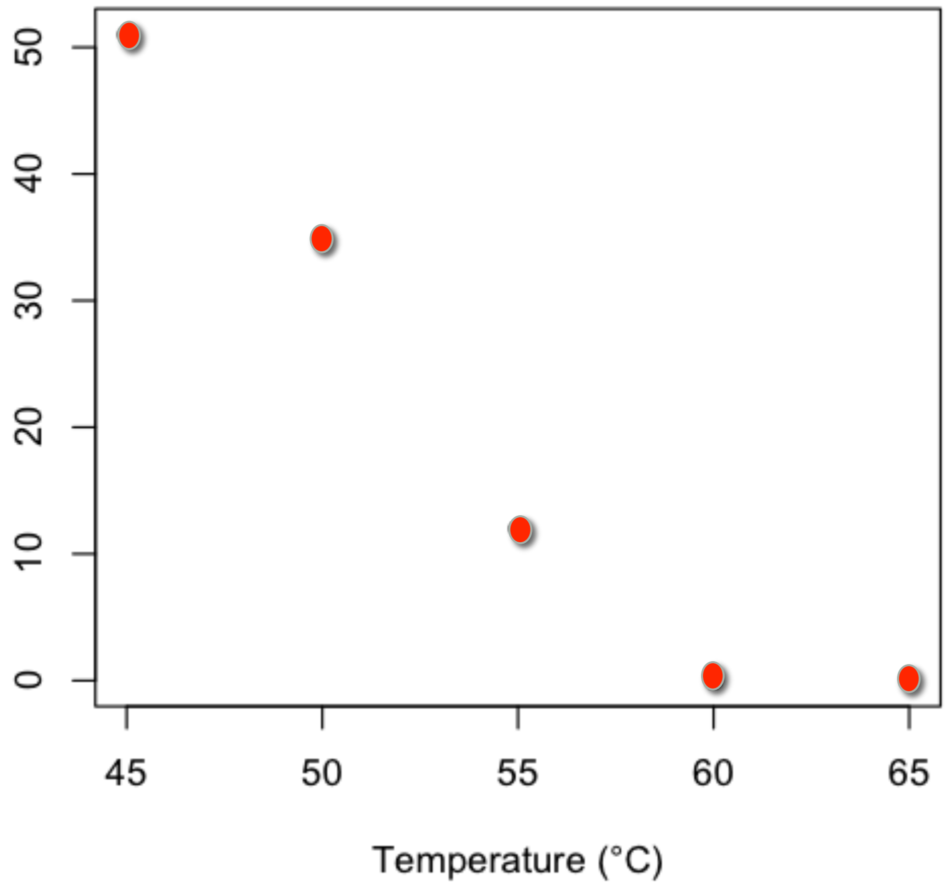
- Ash wood were treated at five different temperatures (45, 50, 55, 60, 65°C) during 30 min

- The number of surviving insects were counted after each heat treatments

Temperature	Nb of insects (for 1 m2 of wood)
45	51
50	35
55	12
60	0
65	0

Example 2: Count data (continued)

Number of surviving insects (for 1 m²)



Models frequently used in plant health studies

Type of data	Model name	R functions
Continuous	Linear model	lm, glm
Continuous	Nonlinear model	nls
Binary	Binomial logit	glm
Categorical with more than two levels	Multinomial logit	mlogit
Count	Poisson log-linear	glm
Repeated measurements	Mixed-effect model	lme, nlme, lmer, glmer

Models less frequently used in plant health studies, but sometimes useful!

Model name	Interest	R packages
Quantile regression	No assumption on the probability distribution of the error term	quantreg
Bayesian models	- More flexible - Useful for combining different types of information - Powerful for uncertainty analysis	R2WINBUGS BRUGS

Models frequently used in plant health studies

Type of data	Model name	R functions
Continuous	Linear model	lm, glm
Continuous	Nonlinear model	nls
Binary	Binomial logit	glm
Categorical with more than two levels	Multinomial logit	mlogit
Count	Poisson log-linear	glm
Repeated measurements	Mixed-effect model	lme, nlme, lmer, glmer

Models frequently used in plant health studies

Type of data	Model name	R functions
Continuous	Linear model	lm, glm
Continuous	Nonlinear model	nls
Binary	Binomial logit	glm
Categorical with more than two levels	Multinomial logit	mlogit
Count	Poisson log-linear	glm
Repeated measurements	Mixed-effect model	lme, nlme, lmer, glmer

Generalized linear models

Outline

- What is a statistical model?
- Why are statistical models useful?
- Types of statistical models, and R functions
- **Generalized linear models**
- Main steps for developping a statistical model
- Conclusions

Generalized linear models

- Useful for analyzing binary and count data
- Deal with some nonlinear relationship between output and inputs
- More general than linear models (linear models are special cases)

Linear model

$$Y = \alpha_0 + \alpha_1 X + \varepsilon$$

Linear model

$$Y = \alpha_0 + \alpha_1 X + \varepsilon$$

Stochastic part

deterministic part

Linear model

$$Y = \alpha_0 + \alpha_1 X + \varepsilon$$

Stochastic part

deterministic part

$$E(Y|X) = \alpha_0 + \alpha_1 X$$

Generalized linear models

Deterministic part: describe the expected value of the data conditionally to the input variables (« mean response »)

Stochastic part: describe the variability of the data conditionally to the input variables

Generalized linear models

Deterministic part

It is defined by a link function *g* such as *:*

$$g\,[\,E(\,Y\,|\,X_1,\,\ldots,\,X_p\,)\,] = \alpha_0 + \alpha_1\,X_1 + \ldots + \alpha_p\,X_p$$

g relates the expected value of the output to the input variables

Linear models are special cases with *g*=identity

Generalized linear models

Stochastic part

Different probability distributions can be used, especially:

- Binomial distribution
- Poisson distribution
- Gaussian (normal) distribution

Important types of generalized linear models

Type	Deterministic part	Stochastic part	R function
Binomial logit	logit link	Binomial distribution	glm(Y~X, family=binomial(link = "logit"))
Poisson log linear	log link	Poisson distribution	glm(Y~X, poisson(link = "log"))
Gaussian linear	Identity link	Gaussian distribution	glm(Y~X, gaussian(link = "identity"))

Main steps for developping statistical models

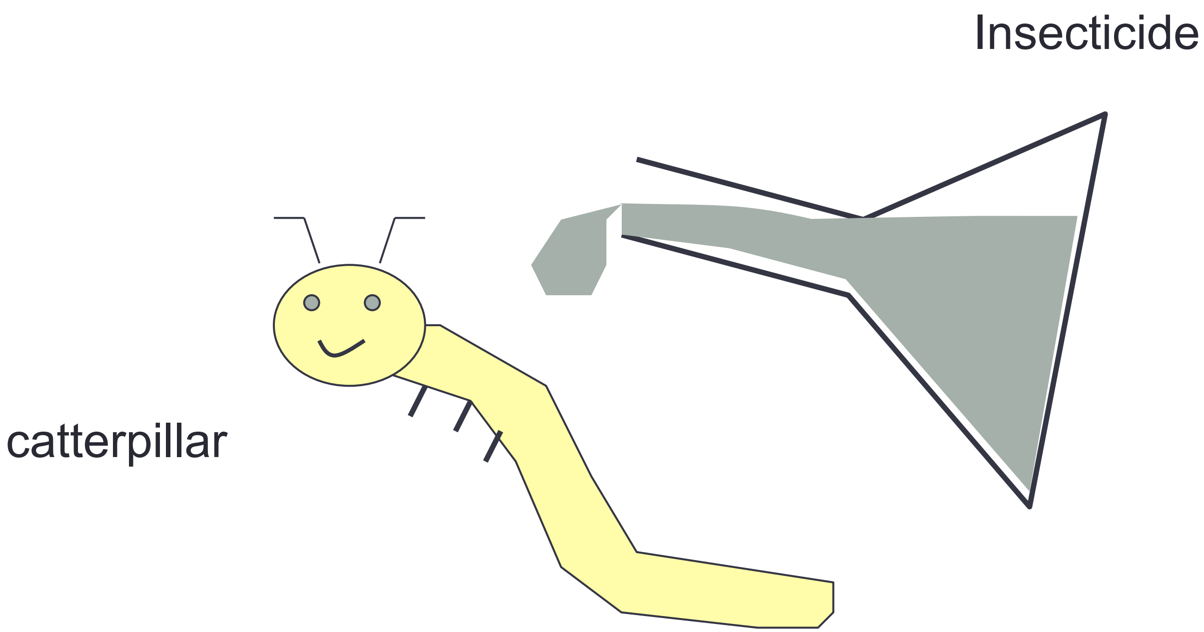
1. Define your objective
2. Look at your data
3. Define output and input variables of one or several models
4. Define model equations relating the output to the inputs
5. Estimate the model parameters
6. Evaluate the model(s)
7. Answer the question

1. Define your objective

- General objective (context of the study)

- Specific objective:
  - List tested hypotheses
  - List predicted variables

Example 1

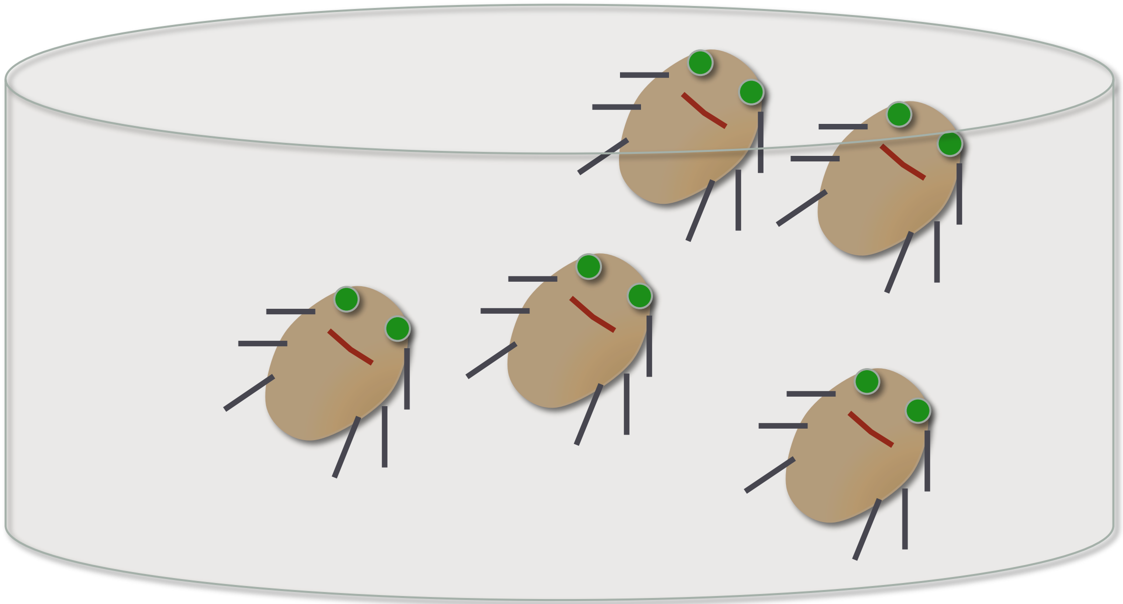


Hypothesis :

« The effectiveness of the insecticide depends on the insecticide dose and on the catterpillar sex »

Example 2

Heat treatment of an infested piece of wood



« Prediction of the number of surviving insects after a heat treatment at 56°C during 30min (official heat treatment) »

2. Look at your data

- Tables
- Figures
- Summary statistics (min, median, mean, max, quartiles)

Example 1

Number of dead caterpillars

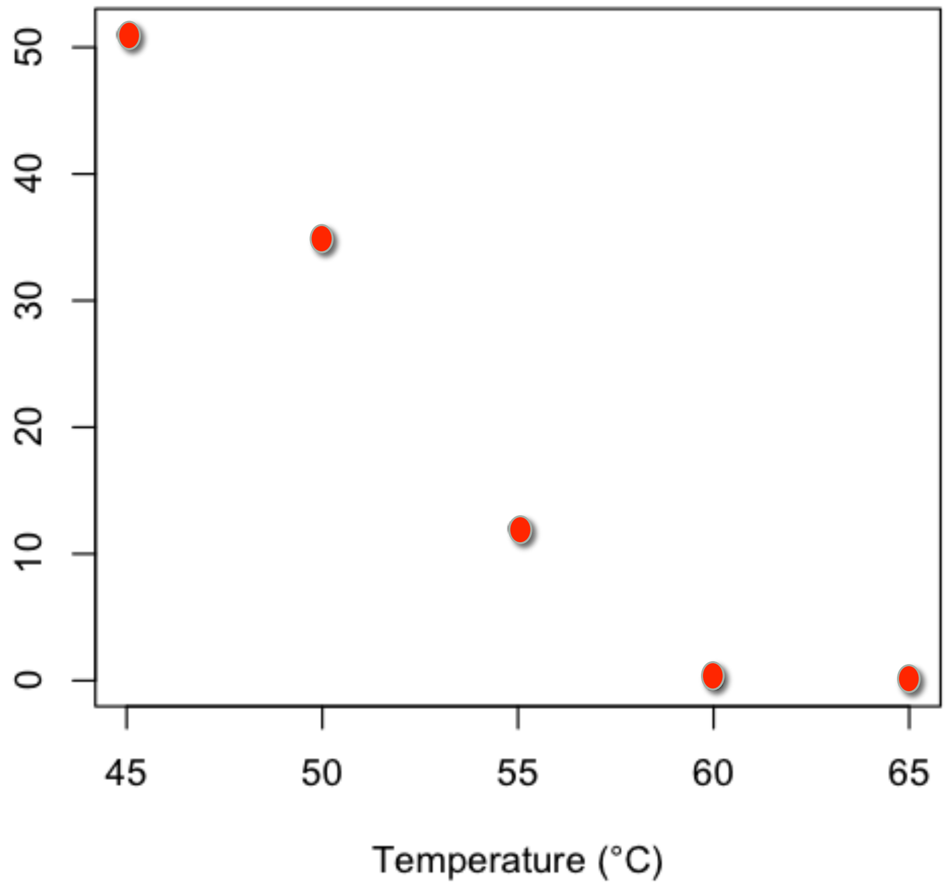
(group size = 20 caterpillars per dose and per sex)

	Insecticide dose					
Sexe	0	1	2	3	4	5
M	1	4	9	13	18	20
F	0	2	6	10	12	16

Collett, 1991

Example 2

Number of surviving insects (for 1 m^2)



3. Define inputs and outputs for one or several models

- Outputs
- Inputs
- Units
- Better to define several models of different levels of complexity

Example 1

- Model output: Mortality rate of the caterpillars
- Model inputs:
  - None
  - Number of dose
  - Number of dose and Sex
  - Number of dose, Sex, interaction between Number of dose and Sex

Example 2

- Model output: Number of surviving insects in $1m^2$ of wood
- Model inputs:
    - None
    - Temperature of the heat treatment

4. Define the model equations

- One or several equations can be defined to relate outputs to inputs
- Identify the model parameters that need to be estimated

Example 1

Stochastic part

$$Y \sim Binomial\left(20, \pi\right)$$

Y: number of killed caterpillar in a group of 20 when the dose x was applied
π: probability that one caterpillar is killed

Deterministic part

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

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

θ_0 and θ_1 are two parameters that need to be estimated

Example 1

Variants for the deterministic part

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

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

Binary variable
(0 for female, 1 for male)

$$\log\left\{\frac{\pi}{1-\pi}\right\} = \text{logit}(\pi) = \theta_0 + \theta_1 x + \theta_2 Sex$$

$$\log\left\{\frac{\pi}{1-\pi}\right\} = \text{logit}(\pi) = \theta_0 + \theta_1 x + \theta_2 Sex + \theta_3 Sex * x$$

Example 2

Stochastic part

$$Y \sim Poisson(\lambda)$$

Y: number of surviving insects when a heat treatment at temperature *x* is applied
λ: expected value of the number of surviving insects

Deterministic part

$$\log(\lambda) = \theta_0 + \theta_1 x$$

$$\lambda = \exp(\theta_0 + \theta_1 x)$$

θ_0 and θ_1 are two parameters that need to be estimated

Example 2

Variants for the deterministic part

$$\log(\lambda) = \theta_0$$

$$\log(\lambda) = \theta_0 + \theta_1 x$$

Main steps for developping statistical models

1. Define your objective
2. Look at your data
3. Define output and input variables of one or several models
4. Define model equations relating the output to the inputs
5. **Estimate the model parameters**
6. Evaluate the model(s)
7. Answer the question

A popular estimation method:

Maximum likelihood

Principle: find the parameter values maximizing the probability of the data conditionally to the model parameters

$$\mathrm{Prob}\left(\mathrm{y}_1,\ldots,\mathrm{y}_\mathrm{N}\,\middle|\,\theta_0,\theta_1,\ldots,\theta_p\right)$$

Other estimation methods:

- Quasi likelihood
- Bayesian methods

In practice…

Estimation using statistical software:

SAS: genmod

R: glm

glm(y~x, family=binomial(link= 'logit'), data…)

glm(y~x, family=poisson(link= 'log'), data…)

glm(y~x, family=gaussian(link= 'identity'), data)

Example 1

$$\log\left\{\frac{\pi}{1-\pi}\right\} = \text{logit}(\pi) = \theta_0 + \theta_1 x + \theta_2 Sex + \theta_3 Sex * x$$

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

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

Example 1

$$\log\left\{\frac{\pi}{1-\pi}\right\} = \text{logit}(\pi) = \theta_0 + \theta_1 x + \theta_2 Sex + \theta_3 Sex * x$$

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

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

Example 1

$$\log\left\{\frac{\pi}{1-\pi}\right\} = \text{logit}(\pi) = \theta_0 + \theta_1 x + \theta_2 Sex + \theta_3 Sex * x$$

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

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

Example 1

$$\log\left\{\frac{\pi}{1-\pi}\right\} = \text{logit}(\pi) = \theta_0 + \theta_1 x + \theta_2 Sex + \theta_3 Sex * x$$

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

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

Example 1

$$\log\left\{\frac{\pi}{1-\pi}\right\} = \mathrm{logit}(\pi) = \theta_0 + \theta_1 x + \theta_2 Sex + \theta_3 Sex * x$$

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

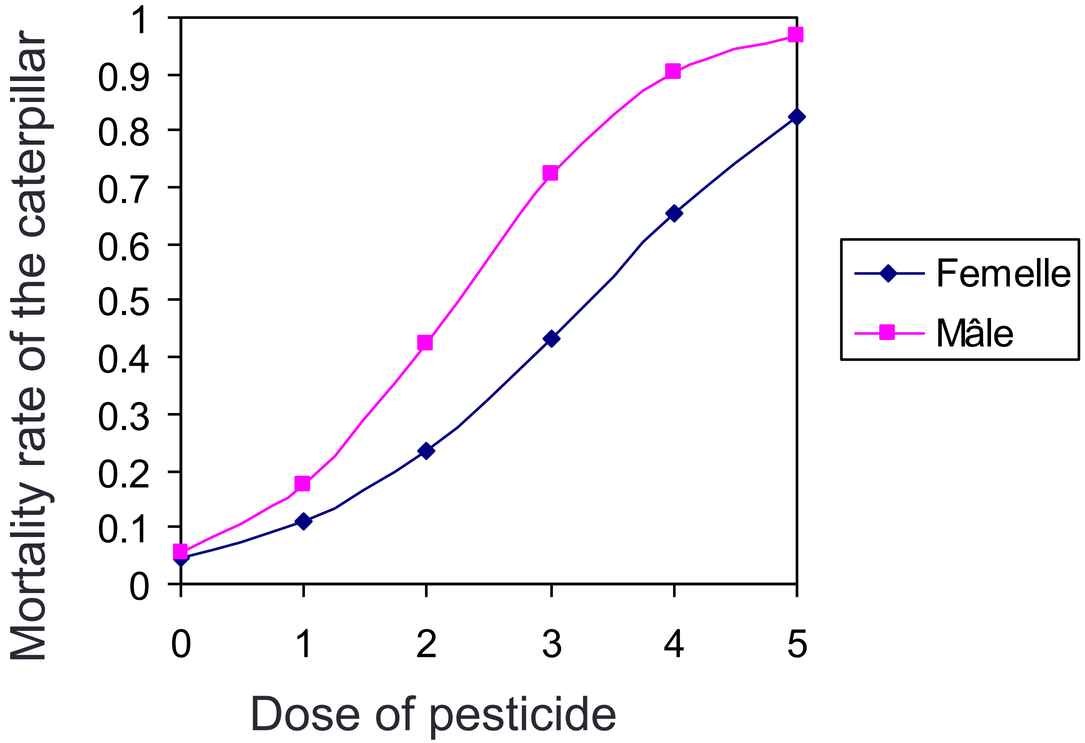
Coefficients:

	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

Example 1

$$\log\left\{\frac{\pi}{1-\pi}\right\} = \mathrm{logit}(\pi) = \theta_0 + \theta_1 x + \theta_2 Sex + \theta_3 Sex * x$$



Example 2

$$\log(\lambda) = \theta_0 + \theta_1 x$$

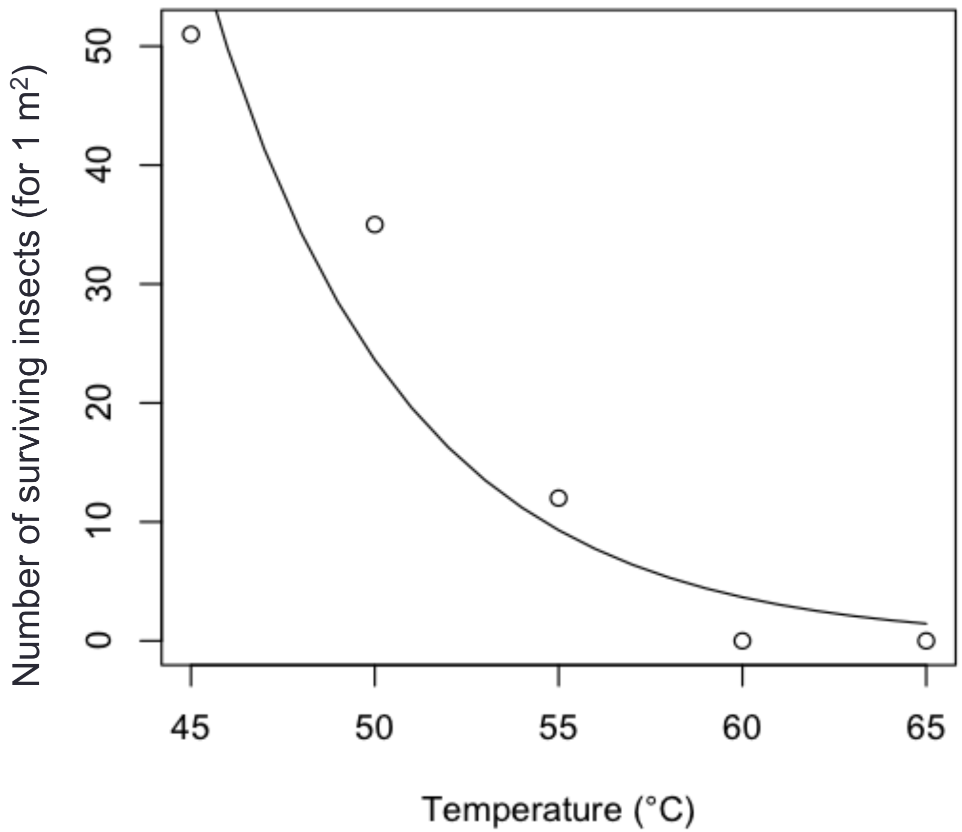
glm(Y~X, family=poisson)

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

Example 2



Number of surviving insects (for 1 m²)
0
10
20
30
40
50
45
50
55
60
65
Temperature (°C)

Main steps for developping statistical models

1. Define your objective
2. Look at your data
3. Define output and input variables of one or several models
4. Define model equations relating the output to the inputs
5. Estimate the model parameters
6. **Evaluate the model(s)**
7. Answer the question

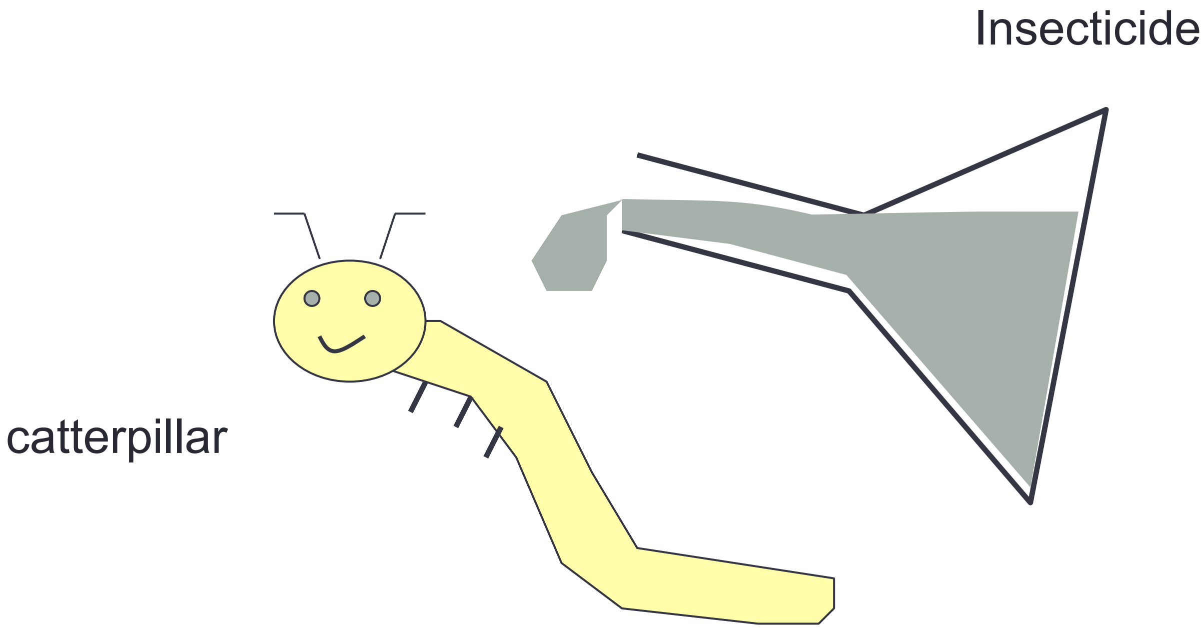
Evaluation methods

- Analysis of the residuals (Obs. – Fitted values)
- Statistical tests
- Confidence intervals
- Selection criteria (AIC, BIC etc.)
- Assessment of prediction/classification errors (MSEP, ROC analysis)

Main steps for developping statistical models

1. Define your objective
2. Look at your data
3. Define output and input variables of one or several models
4. Define model equations relating the output to the inputs
5. Estimate the model parameters
6. Evaluate the model(s)
7. **Answer the question**

Example 1



Hypothesis :

« The effectiveness of the insecticide depends on the insecticide dose and on the catterpillar sex »

Example 1

$$\log\left\{\frac{\pi}{1-\pi}\right\} = \text{logit}(\pi) = \theta_0 + \theta_1 x + \theta_2 Sex + \theta_3 Sex * x$$

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

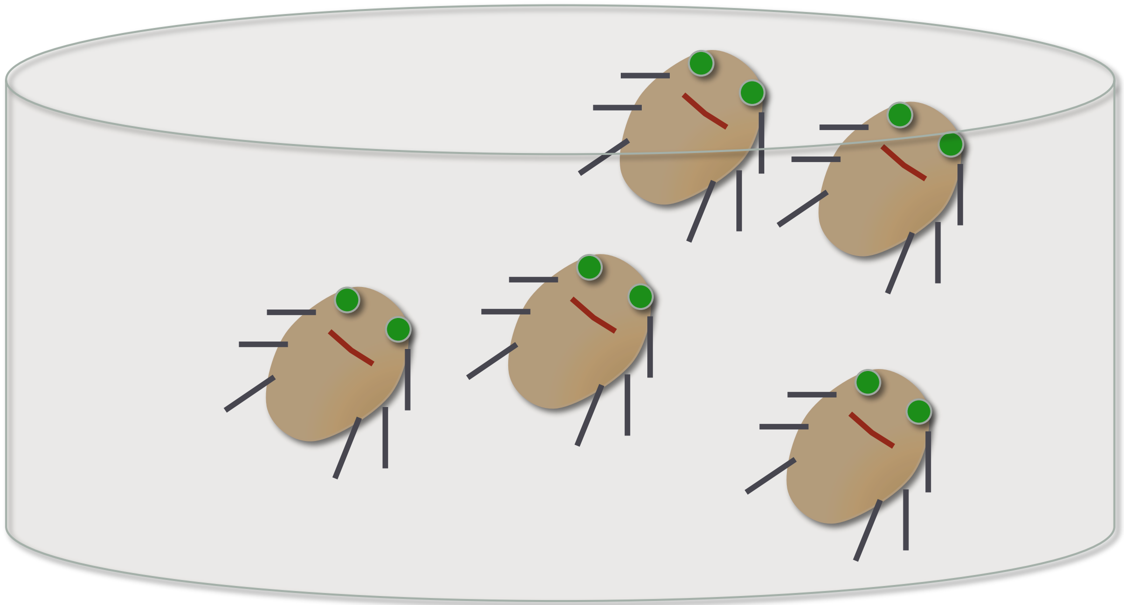
Coefficients:

	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

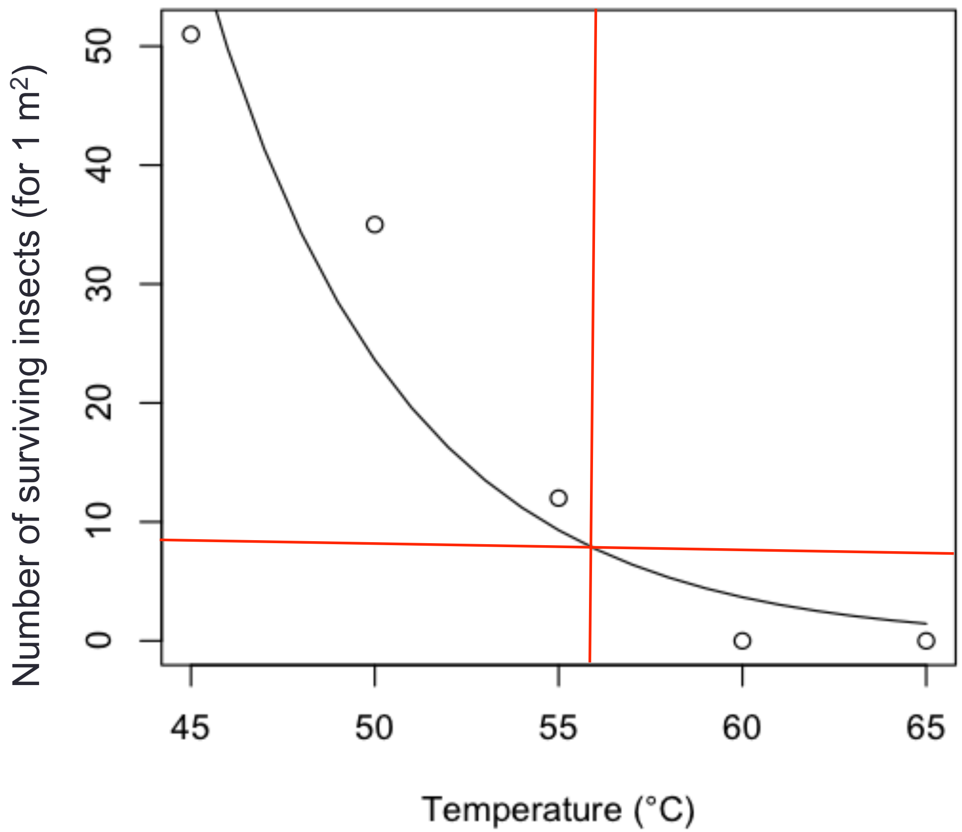
Example 2

Heat treatment of an infested piece of wood



« Prediction of the number of surviving insects after a heat treatment at 56°C during 30min (official heat treatment) »

Example 2



7.69 insects per m^2
95% conf. int. [5.12, 11.56]

Main steps for developping statistical models

1. Define your objective
2. Look at your data
3. Define output and input variables of one or several models
4. Define model equations relating the output to the inputs
5. Estimate the model parameters
6. Evaluate the model(s)
7. Answer the question

Models frequently used in plant health studies

Type of data	Model name	R functions
Continuous	Linear model	lm, glm
Continuous	Nonlinear model	nls
Binary	Binomial logit	glm
Categorical with more than two levels	Multinomial logit	mlogit
Count	Poisson log-linear	glm
Repeated measurements	Mixed-effect model	lme, nlme, lmer, glmer

Models less frequently used in plant health studies, but sometimes useful!

Model name	Interest	R packages
Quantile regression	No assumption on the probability distribution of the error term	quantreg
Bayesian models	- More flexible - Useful for combining different types of information - Powerful for uncertainty analysis	R2WINBUGS BRUGS