Basic Statistical Analysis in Life and Environmental Sciences

Rodrigo Labouriau

Department of Mathematics, Aarhus University

Module 5, Day 8 - Closing - 2024

(Overview, generalised linear models, random components, other types of models)

1

[1] Copyright © 2024 by Rodrigo Labouriau.

This material is only for internal use in the course. Please, do not circulate and do not record.

General Remark

This material is only for internal use in the course.

Please, do not circulate and **do not record**.

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Outline

- Binomial Models
  - Binomial classification models
  - Binomial (logistic) regression models
- Poisson Models
  - Poisson classification models
  - Poisson regression models
- Normal models
  - Normal models with classification structure
  - Normal regression models
- Generalized Linear Models
- A simple model with random components
- Some Other Models
  - Non-parametric regression
  - A Poisson model with random components ...

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Review

Overview of models

- Binomial - Proportions
  - One- way and two-ways classification
    Lecture 3: Seed germination under different watering levels and cancer prevalence
  - Logistic regression
    Lecture 4: Leave abscission of Radamachera
- Poisson - Counts
  - One- way and two-ways classification
    Deaths by horse kicks
  - Linear and non-linear regression
    CFU of *Penicillium verrucosum*
- Normal (Gaussian) - Continuous varying responses
  - One- way and two-ways classification
    Lecture 7: Seed weights of *Dolichos biflorus*
  - Linear and non-linear regression
    Lecture 7: Maize response to P

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Overview of models

- Binomial - Proportions
  - One- way and two-ways classification

    Lecture 3: Seed germination under different watering levels and cancer prevalence
  - Logistic regression
    Lecture 4: Leave abscission of Radamachera
- Poisson - Counts
  - One- way and two-ways classification
    Deaths by horse kicks
  - Linear and non-linear regression
    CFU of *Penicillium verrucosum*
- Normal (Gaussian) - Continuous varying responses
  - One- way and two-ways classification
    Lecture 7: Seed weights of *Dolichos biflorus*
  - Linear and non-linear regression
    Lecture 7: Maize response to P

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Binomial one-way classification models

Seed germination example: The data

Watering Level				
1	2	3	4	5
22	41	66	82	79
25	46	72	73	68
27	59	51	73	74
23	38	78	84	70

Number of germinated seeds, out of 100 seeds,
For $w = 1, \ldots, 5$ (indexing the watering levels) and $r = 1, \ldots, 4$ (indexing the repetitions)
Saturated model: $Y_{wr} \sim Bi(100, \rho_{wr})$
Full model: $Y_{wr} \sim Bi(100, \rho_w)$
Null model: $Y_{wr} \sim Bi(100, \rho)$

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Testing homogeneity

Large Model = Saturated Model					
	Watering Level				
Repetition	1	2	3	4	5
1	ρ_{11}	ρ_{21}	ρ_{31}	ρ_{41}	ρ_{51}
2	ρ_{12}	ρ_{22}	ρ_{32}	ρ_{42}	ρ_{52}
3	ρ_{13}	ρ_{23}	ρ_{33}	ρ_{43}	ρ_{53}
4	ρ_{14}	ρ_{24}	ρ_{34}	ρ_{44}	ρ_{54}

Reduced Model = One-way					
	Watering Level				
Repetition	1	2	3	4	5
1	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
2	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
3	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
4	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5

Probabilities of germination in each box under the saturated (large) and the one-way (reduced) models.

- Idea: Compare the saturated model with the one-way binomial model using the likelihood ratio test
- Equivalent to test the null hypothesis

H_0 : "The probability parameters associated with observations with the same level of the classification variable are all equal"

- The log-likelihood ratio statistic for this test is given by $\Lambda = 2\left\{l_L - l_S\right\}$, where l_L and l_S are the log-likelihood functions of the one-way binomial model and the log-likelihood of the saturated, both evaluated at their maxima, respectively.
- Λ is approximately chi-square distributed with $20 - 5 = 15$ degrees of freedom,

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Binomial one-way classification models - Calculations in R

Identifying the cause of the lack of homogeneity for the seed germination data

```
   Box Water Germ fitted.v    res.dev
1    1     1   22   0.2425 -0.5307697
2    2     1   25   0.2425  0.1743868
3    3     1   27   0.2425  0.6338446
4    4     1   23   0.2425 -0.2934009
5    5     2   41   0.4600 -1.0067849
6    6     2   46   0.4600  0.0000000
7    7     2   59   0.4600  2.6049858   <--
8    8     2   38   0.4600 -1.6156868
9    9     3   66   0.6675 -0.1589032
10  10     3   72   0.6675  1.1308793
11  11     3   51   0.6675 -3.2480607   <--
12  12     3   78   0.6675  2.4749070   <--
13  13     4   82   0.7800  0.9887183
14  14     4   73   0.7800 -1.1774922
15  15     4   73   0.7800 -1.1774922
16  16     4   84   0.7800  1.5032032
17  17     5   79   0.7275  1.4421817
18  18     5   68   0.7275 -1.0491850
19  19     5   74   0.7275  0.2821216
20  20     5   70   0.7275 -0.6114889
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

The likelihood ratio test

Seed germination example: Testing the effect of watering

Large Model					
	Watering Level				
Repetition	1	2	3	4	5
1	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
2	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
3	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
4	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5

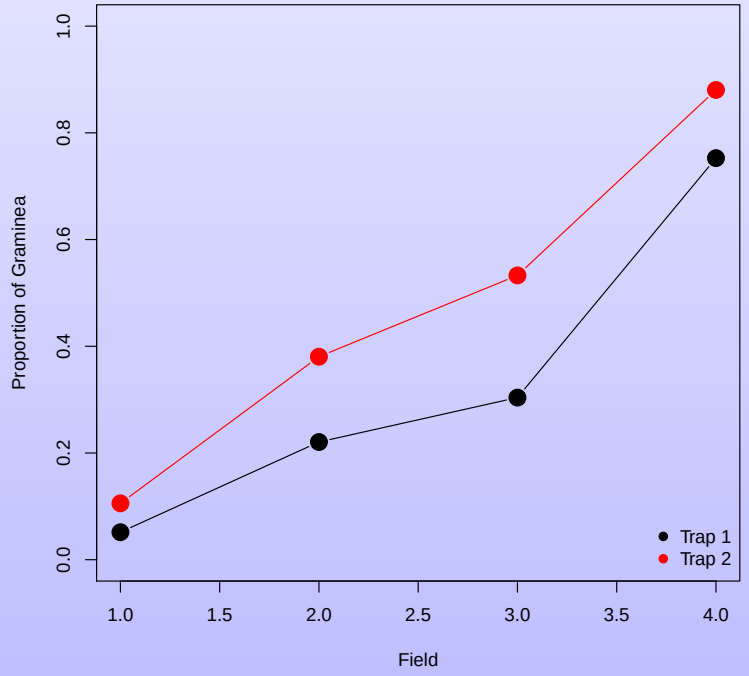
Reduced Model					
	Watering Level				
Repetition	1	2	3	4	5
1	ρ	ρ	ρ	ρ	ρ
2	ρ	ρ	ρ	ρ	ρ
3	ρ	ρ	ρ	ρ	ρ
4	ρ	ρ	ρ	ρ	ρ

Probabilities of germination in each box under the large and the reduced models.

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Binomial two-ways classification models

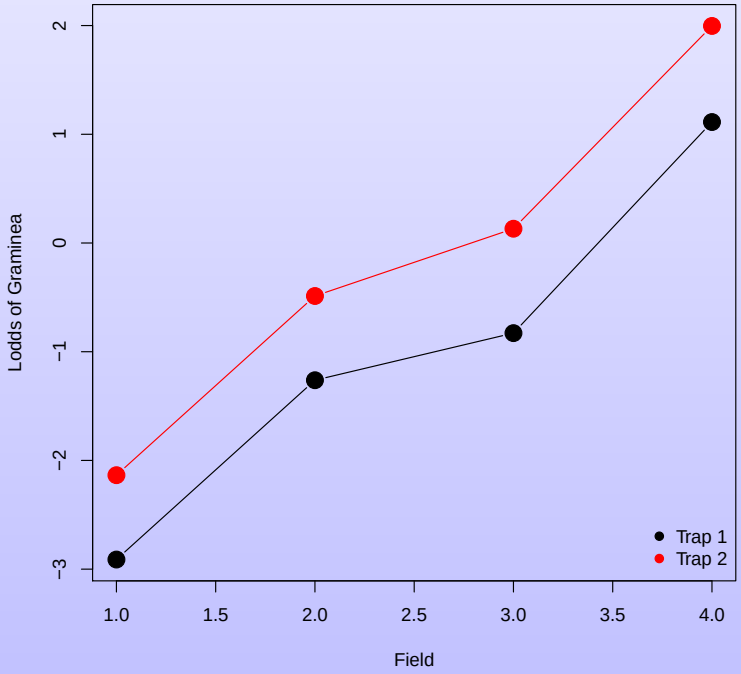
Example: vegetation composition in four fields



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Binomial two-ways classification models

Example: vegetation composition in four fields



The two profiles are parallel

$\Longrightarrow$ the differences of the lodds of (any) pair of fields is the same for the two traps

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Binomial two-ways classification models

Example: vegetation composition in four fields

- I will consider a range of models
- **Saturated** Model: One lodd per observation
- **Effect-Modification** (Full) Model:
  One lodd per combination of field and trap
- **Additive** Model: The lodds are described as the sum of
  a quantity depending on the Field and a quantity depending on the trap
- **No-effect of Field** Model: The lodds are the same for all the
  observations coming from the same trap
- **No-effect of Trap** Model: The lodds are the same for all the
  observations coming from the same field
- **Null** Model: The lodds are the same for all the observations

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Overview:

Binomial models, two-ways classification models

- Two classification variables, say T and S
  Y_{tsi} the ith repetition of observations classified as t and s
- $Y_{111}, Y_{112}, \ldots$ independent
- $Y_{tsi} \sim Bi(n_{tsi}, p_{tsi})$, for $t, s = 1, 2, \ldots$
  with several possibilities for p_{tsi} (yielding different models)
- Some possibilities are:
  - $logit(p_{tsi}) = (T * S)_{tsi}$ (the saturated model)
  - $logit(p_{tsi}) = (T * S)_{ts}$ (interaction model)
  - $logit(p_{tsi}) = T_t + S_s$ (additive model)
  - $logit(p_{tsi}) = S_s$ (no effect of T)
  - $logit(p_{tsi}) = T_s$ (no effect of S)
  - $logit(p_{tsi}) = k$ (null model)

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

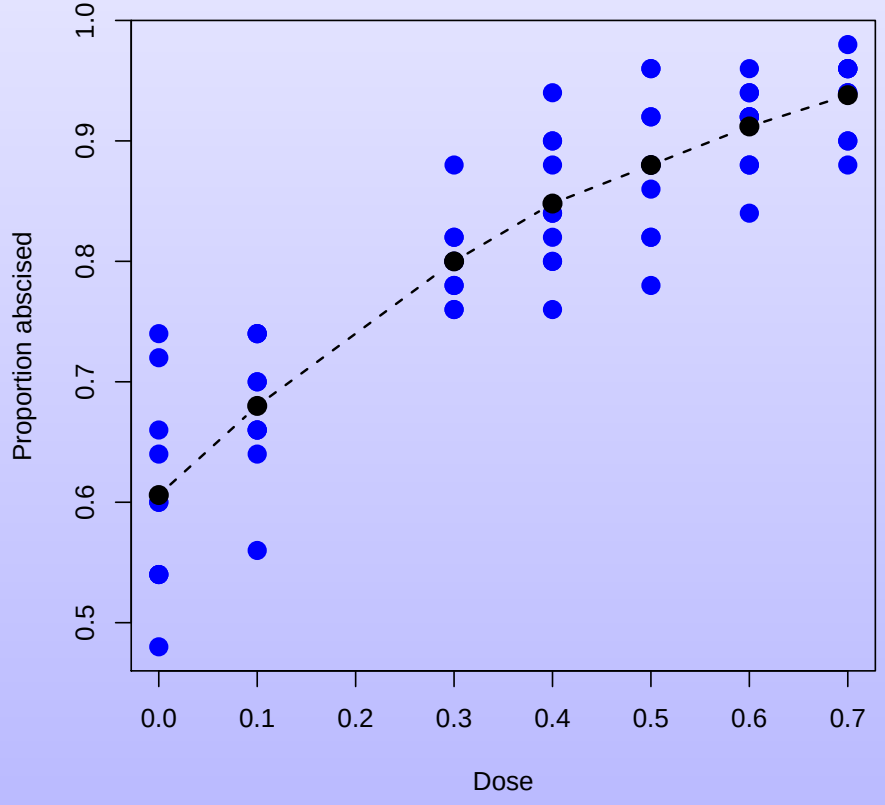
Overview of models

- Binomial - Proportions
  - One- way and two-ways classification
    Lecture 3: Seed germination under different watering levels and cancer prevalence
  - Logistic regression
    Lecture 4: Leave abscission of Radamachera
- Poisson - Counts
  - One- way and two-ways classification
    Deaths by horse kicks
  - Linear and non-linear regression
    CFU of *Penicillium verrucosum*
- Normal (Gaussian) - Continuous varying responses
  - One- way and two-ways classification
    Lecture 7: Seed weights of *Dolichos biflorus*
  - Linear and non-linear regression
    Lecture 7: Maize response to P

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Review, Binomial regression

Leave abscission of Radamachera



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Review, Binomial regression models

Logistic regression

- Regression model: assume that the probabilities of abscision are a (continuous) function of the dose of abscisic acid
- Y is number of plants with more than 50 % of abscised leaves
  out of the 50 plants in each batch
- d is the dose (mg/plant)
- $Y \sim Bi(50, p_d)$
- We assume

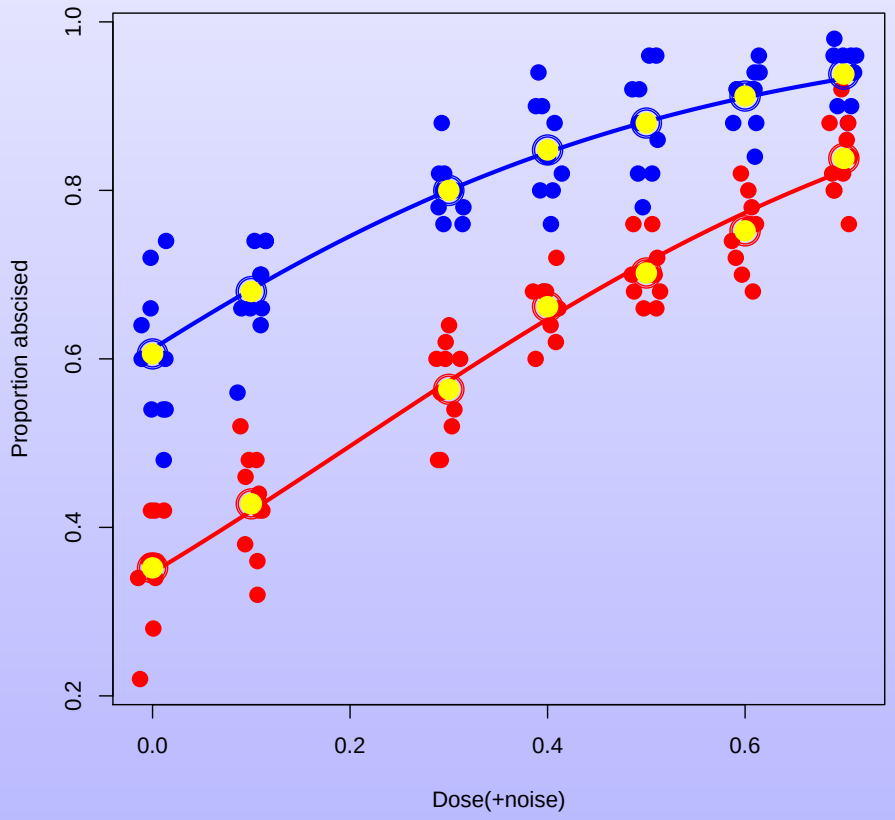
$$\log\left(\frac{p_d}{1-p_d}\right) = \alpha + \beta d$$

- α and β are parameters in the model.

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Review, Binomial regression models

Leave abscision of Radamachera, with two varieties



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Overview of models

- Binomial - Proportions
  - One- way and two-ways classification
    Lecture 3: Seed germination under different watering levels and cancer prevalence
  - Logistic regression
    Lecture 4: Leave abscission of Radamachera
- Poisson - Counts
  - One- way and two-ways classification
    Deaths by horse kicks
  - Linear and non-linear regression
    CFU of *Penicillium verrucosum*
- Normal (Gaussian) - Continuous varying responses
  - One- way and two-ways classification
    Lecture 7: Seed weights of *Dolichos biflorus*
  - Linear and non-linear regression
    Lecture 7: Maize response to P

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Overview, Poisson classification models

Example: Horse-kicks

- The data are registers of Prussian military persons killed by kicks of horses.
- Ten corps observed (separately) during 20 years: 1875-1894

  (4 less representative corps were eliminated)
- The table below (next slide) displays the data
- The frequencies of number of deaths per year are:

Deaths					
0	1	2	3	4	≥ 5
109	65	22	3	1	0

- We are facing a rare event!
  (122 occurrences in 20 years 6.1 / year 0.61 per corp year)

  We will try to use the Poisson distribution!

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Overview, Poisson classification models

One-way Poisson model

- We start by analysing the total number of deaths per year

  We sum, for each year, the number of deaths occurred in each corp.

- The question is whether the number of deaths per year varies.
- Y_{year} number of deaths occurred in this year
- $Y_{year} \sim$ Poisson
- Two possible models:
  - Common intensity model: $Y_{year} \sim Po(\lambda)$
  - Saturated model: $Y_{year} \sim Po(\lambda_{year})$

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

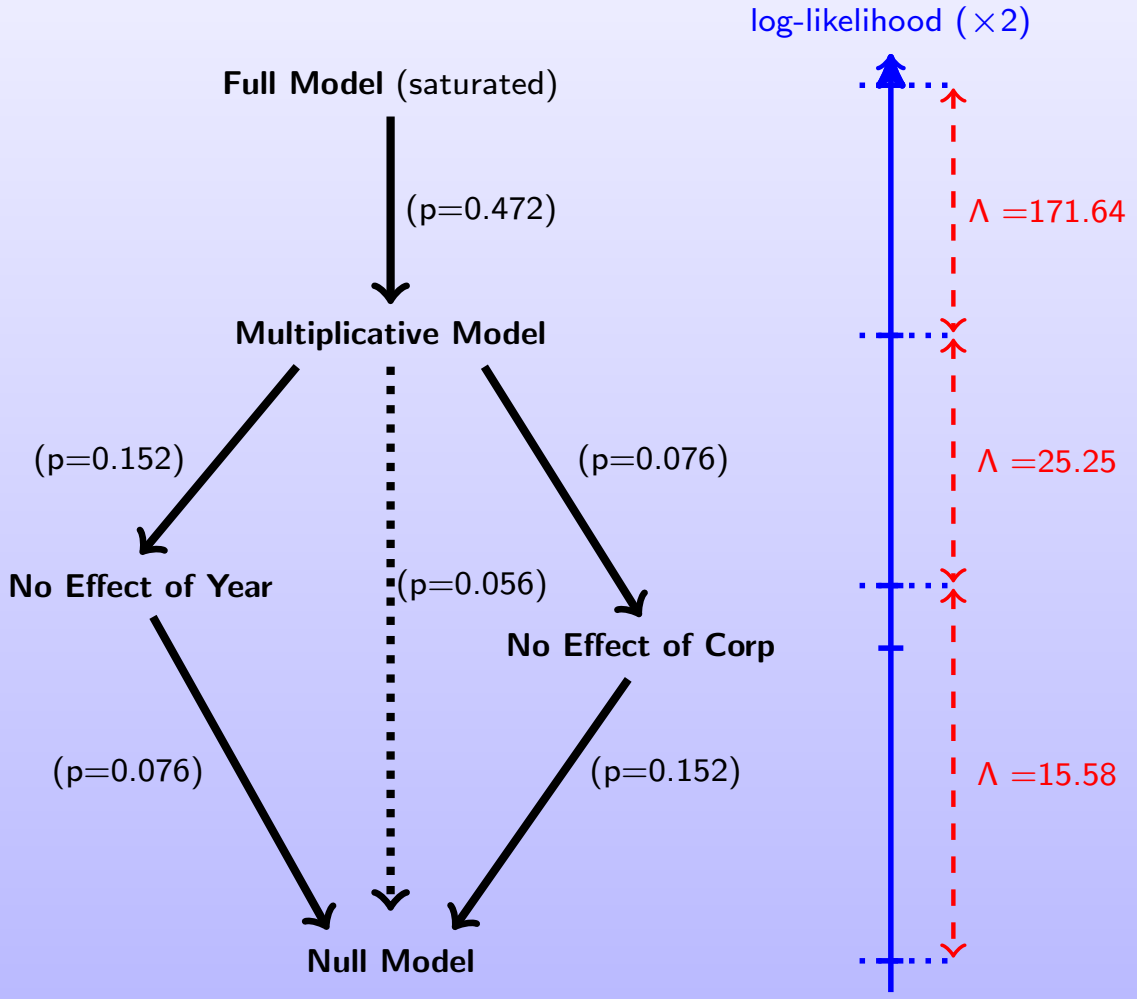
Overview, Poisson classification models

The complete data of deaths by horse kicks, two ways classification

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	Total
1875	0	0	0	0	1	1	0	0	1	0	3
1876	0	0	1	0	0	0	0	0	1	1	3
1877	0	0	0	0	1	0	0	1	2	0	4
1878	2	1	1	0	0	0	0	1	1	0	6
1879	0	1	1	2	0	1	0	0	1	0	6
1880	2	1	1	1	0	0	2	1	3	0	11
1881	0	2	1	0	1	0	1	0	0	0	5
1882	0	0	0	0	0	1	1	2	4	1	9
1883	1	2	0	1	1	0	1	0	0	0	6
1884	1	0	0	0	1	0	0	2	1	1	6
1885	0	0	0	0	0	0	2	0	0	1	3
1886	0	0	1	1	0	0	1	0	3	0	6
1887	2	1	0	0	2	1	1	0	2	0	9
1888	1	0	0	1	0	0	0	0	1	0	3
1889	1	1	0	1	0	0	1	2	0	2	8
1890	0	2	0	1	2	0	2	1	2	2	12
1891	0	1	1	1	1	1	0	3	1	0	9
1892	2	0	1	1	0	1	1	0	1	0	7
1893	0	0	0	1	2	0	0	1	0	0	4
1894	0	0	0	0	0	1	0	1	0	0	2

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Poisson two ways classification model



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

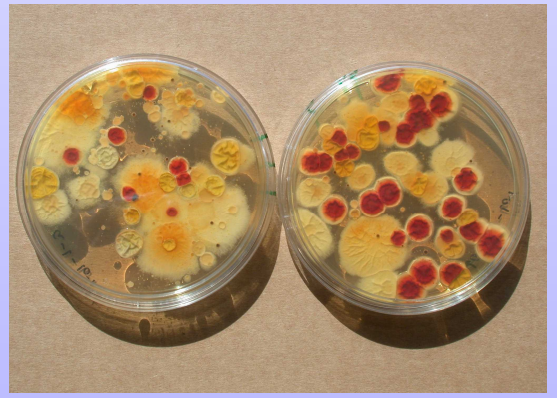
Overview of models

- Binomial - Proportions
  - One- way and two-ways classification
    Lecture 3: Seed germination under different watering levels and cancer prevalence
  - Logistic regression
    Lecture 4: Leave abscission of Radamachera
- Poisson - Counts
  - One- way and two-ways classification
    Deaths by horse kicks
  - Linear and non-linear regression
    CFU of *Penicillium verrucosum*
- Normal (Gaussian) - Continuous varying responses
  - One- way and two-ways classification
    Lecture 7: Seed weights of *Dolichos biflorus*
  - Linear and non-linear regression
    Lecture 7: Maize response to P

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Overview, Poisson regression models

Example: Penicillium in soil



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Overview, Poisson regression models

Example: Penicillium in soil

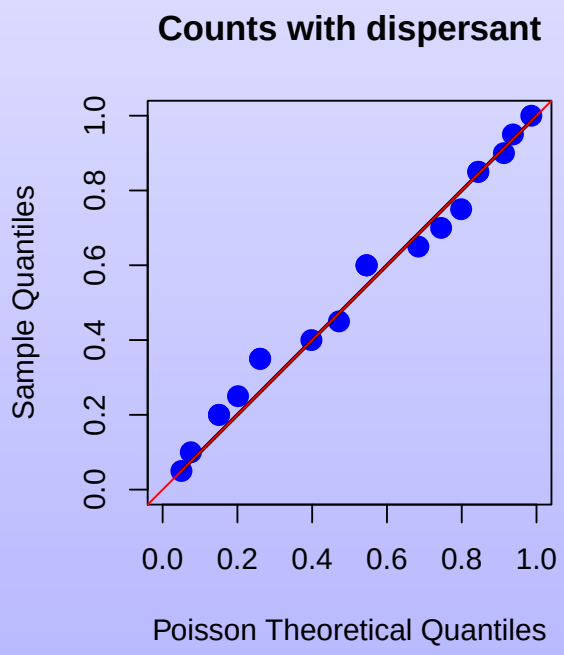
- We performed the following experiment:
  - Make a suspension of the soil;
  - Take successive dilutions of the suspension;
  - Plate the dilutions in Petri dishes and count the number of colonies that appeared after an incubation time.
- This technique is called the plating method (Fisher, 1922).
- Knowing the amount of soil added, estimate the number of CFU / g soil
- Better method:
  Use several amounts of soil and assume that the expected number of CFU is proportional to the amount of soil added

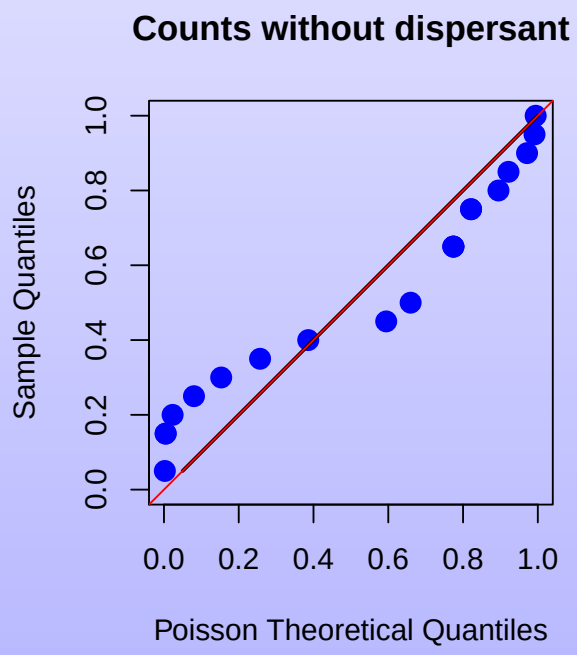
 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Basic assumptions of a Poisson process:

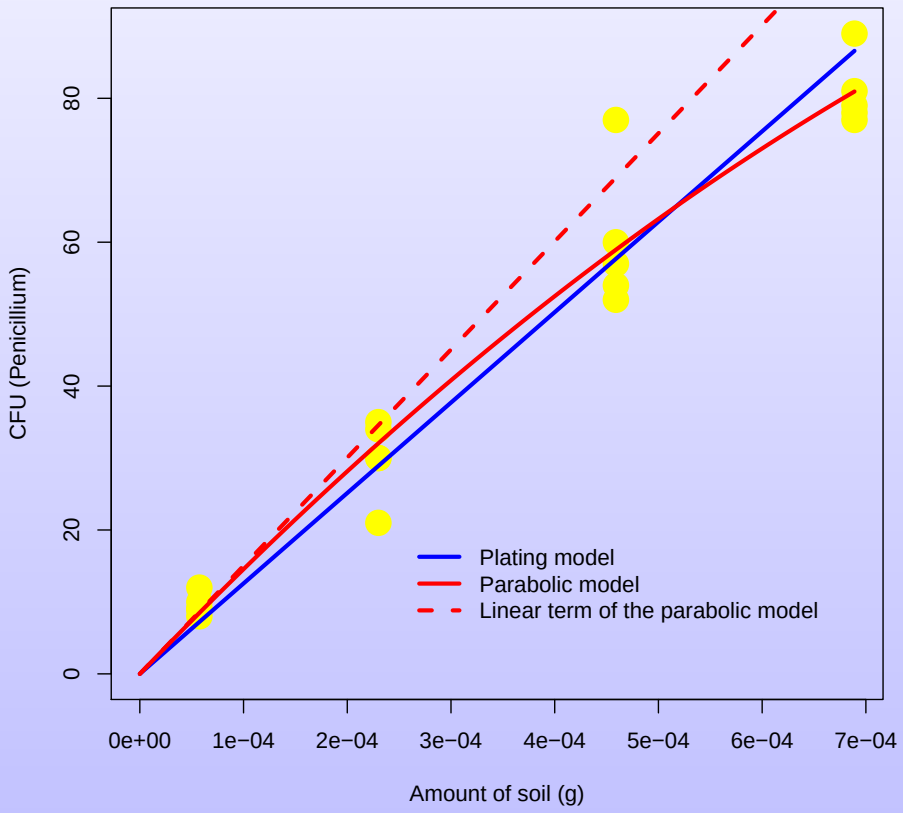
- Homogeneous distribution of the CFUs in the suspension.
- The number of CFUs in two disjoint portions of the suspension are independent
- The CFUs are not clustered together.

Under these assumptions the counts should be Poisson distributed!





 Copyright © 2023 by Rodrigo Labouriau, please do not circulate



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Example: Penicillium in soil

Saturated Model $E(Y_{g,d}) = \lambda_{g,d}$

(p=0.695)

Full Model (free curve) $E(Y_{g,d}) = \lambda_g$

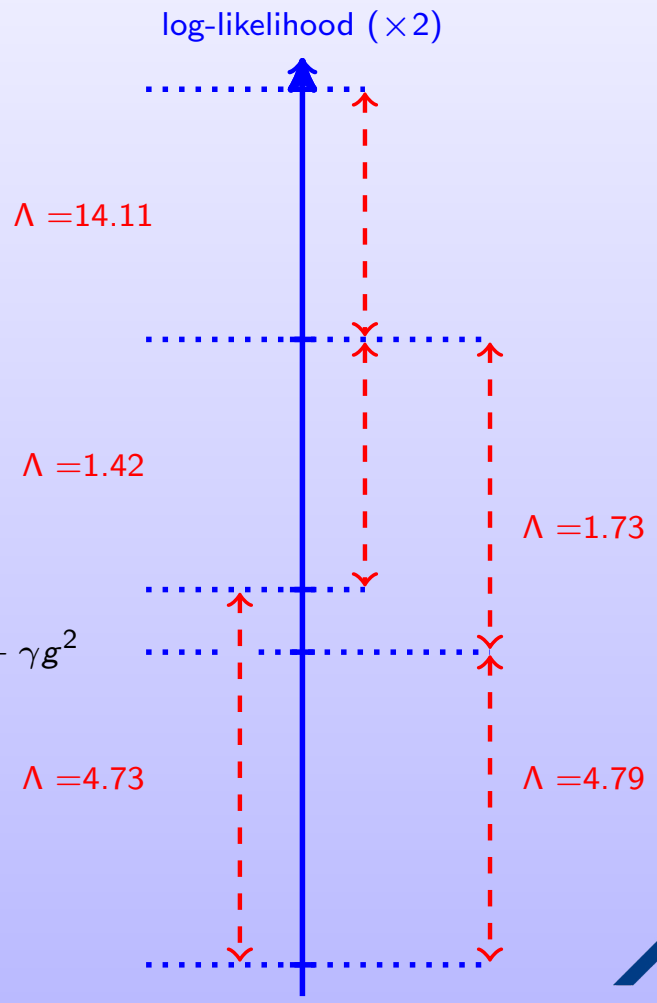
(p=0.491) (p=0.476)

Linear model $E(Y_{g,d}) = \alpha + \beta g$

Quadratic model (competition) $E(Y_{g,d}) = \beta g + \gamma g^2$

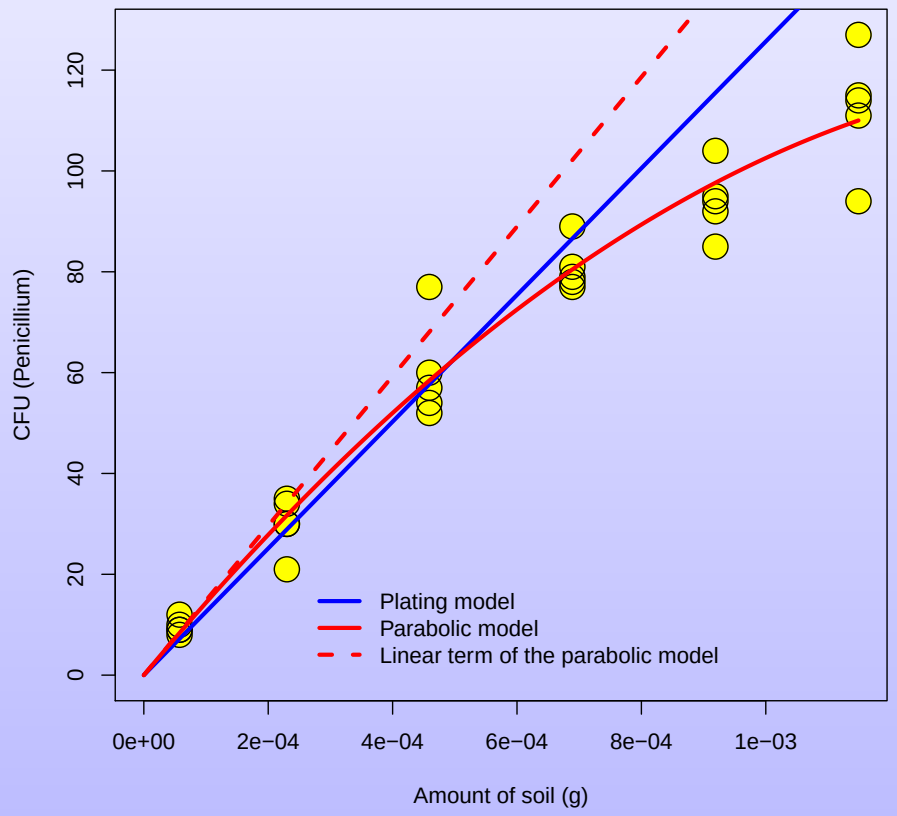
(p=0.028) (p=0.029)

Plating Model $E(Y_{g,d}) = \beta g$



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Extended experiment



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Example: Penicillium in soil - Extended experiment

Saturated Model $E(Y_{g,d}) = \lambda_{g,d}$

(p=0.710)

Full Model (free curve) $E(Y_{g,d}) = \lambda_g$

(p=0.652)

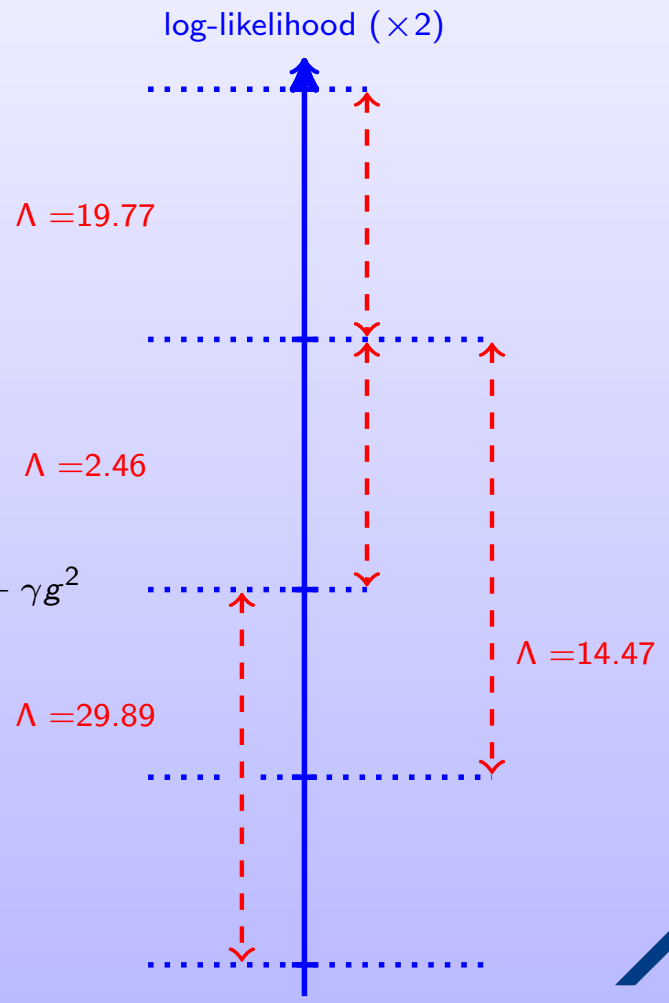
(p=0.006)

Quadratic model (competition) $E(Y_{g,d}) = \beta g + \gamma g^2$

(p<0.001)

Linear model $E(Y_{g,d}) = \alpha + \beta g$

Plating Model $E(Y_{g,d}) = \beta g$



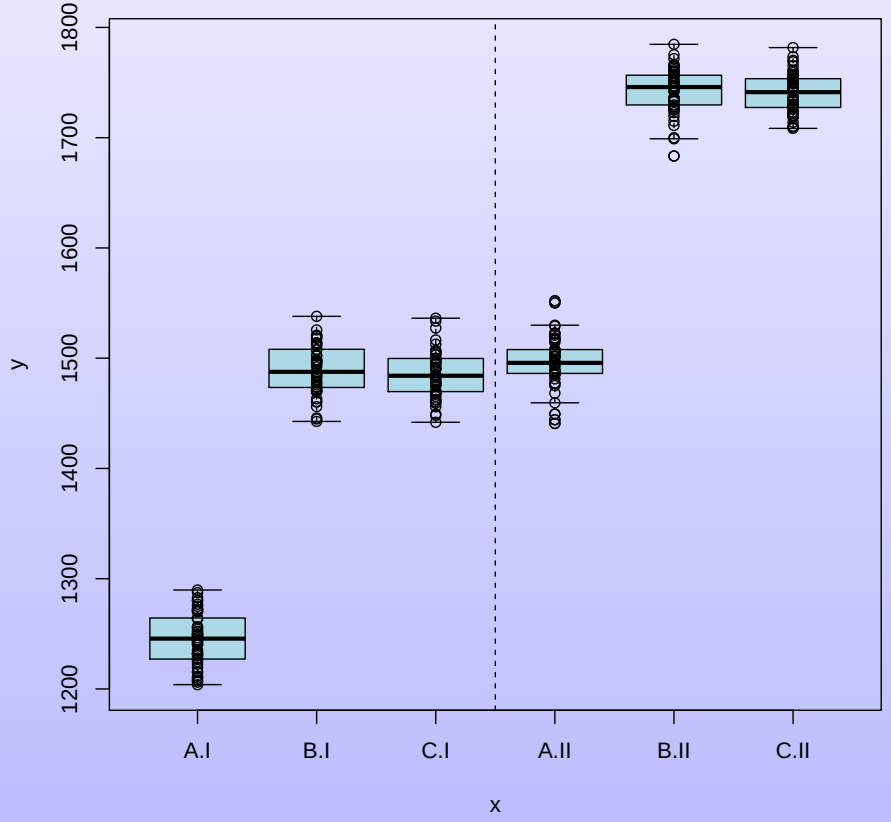
 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Overview of models

- Binomial - Proportions
  - One- way and two-ways classification
    Lecture 3: Seed germination under different watering levels and cancer prevalence
  - Logistic regression
    Lecture 4: Leave abscission of Radamachera
- Poisson - Counts
  - One- way and two-ways classification
    Deaths by horse kicks
  - Linear and non-linear regression
    CFU of *Penicillium verrucosum*
- Normal (Gaussian) - Continuous varying responses
  - One- way and two-ways classification
    Lecture 7: Seed weights of *Dolichos biflorus*
  - Linear and non-linear regression
    Lecture 7: Maize response to P

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Two-ways ANOVA - comparing three varieties in two fields



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Two-ways ANOVA - comparing three varieties in two fields - Interaction Model

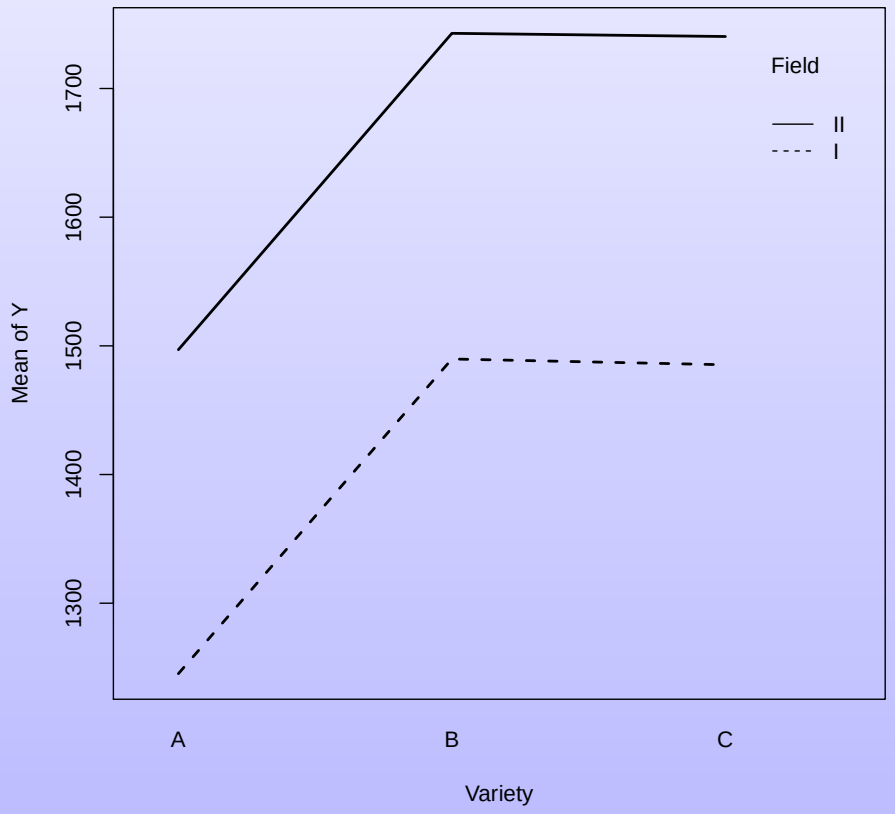
- Y_{vfb} is the random variable representing the averaged weight of the bth batch ($b = 1, ..., 50$) of the vth variety ($v = A, B, C$) from the fth field ($f = I, II$)
- The model assumes that the random variables $Y_{AI1}, \ldots, Y_{CII50}$ are:
  - independent,
  - normally distributed
  - have the same variance (say $Var(Y_{vfb}) = \sigma^2$)
  - have expectation depending on the combination of variety and field (say $E(Y_{vfb}) = \tau_{vf}$)
- In short,

$$Y_{vfb} \sim N\left(\tau_{vf}, \sigma^2\right), \text{ for } v = A, B, C, \ f = I, II \text{ and } b = 1, ..., 50,$$

where $Y_{AI1}, \ldots, Y_{CII50}$ are independent.

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Two-ways ANOVA - comparing three varieties in two fields - investigating additivity



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Two-ways ANOVA - comparing three varieties in two fields - the additive model

- Y_{vfb} is the random variable representing the averaged weight of the bth batch ($b = 1, ..., 50$) of the vth variety ($v = A, B, C$) from the fth field ($f = I, II$)

- The model assumes that the random variables $Y_{AI1}, \ldots, Y_{CII50}$ are:
  - independent,
  - normally distributed
  - have the same variance (say $Var(Y_{vfb}) = \sigma^2$)
  - The expectation can be written as a sum of a quantity depending on the variety and a quantity depending on the field (say $E(Y_{vfb}) = \tau_v + \beta_f$)

- In short,

$$Y_{vfb} \sim N\left(\tau_v + \beta_f, \sigma^2\right), \text{ for } v = A, B, C, \ f = I, II \text{ and } b = 1, ..., 50,$$

where $Y_{AI1}, \ldots, Y_{CII50}$ are independent.

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Overview of models

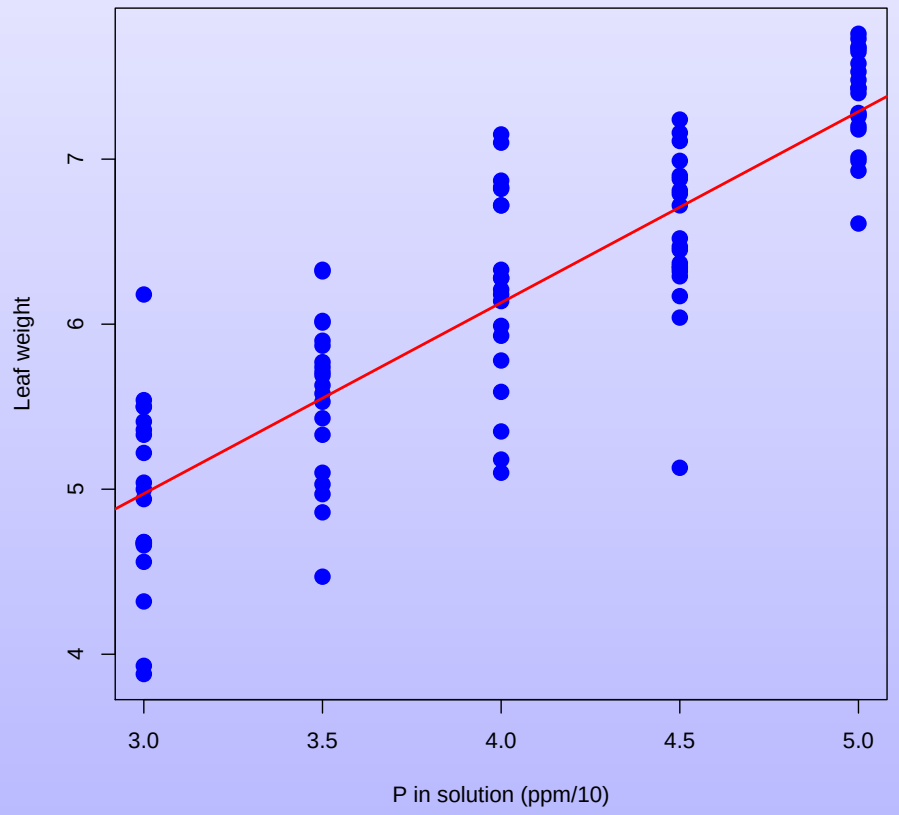
- Binomial - Proportions
  - One- way and two-ways classification
    Lecture 3: Seed germination under different watering levels and cancer prevalence
  - Logistic regression
    Lecture 4: Leave abscission of Radamachera
- Poisson - Counts
  - One- way and two-ways classification
    Deaths by horse kicks
  - Linear and non-linear regression
    CFU of *Penicillium verrucosum*
- Normal (Gaussian) - Continuous varying responses
  - One- way and two-ways classification
    Lecture 7: Seed weights of *Dolichos biflorus*
  - Linear and non-linear regression

    Lecture 7: Maize response to P

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Review, Normal linear regression models

Maize response to P



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Review, Normal linear regression models

Maize response to P

- Y_{pr} weight of the r-th repetition
  subject to the amount p of Phosphorous
- We assume that the expected weight depends linearly on the amount of Phosphorous
- In symbols

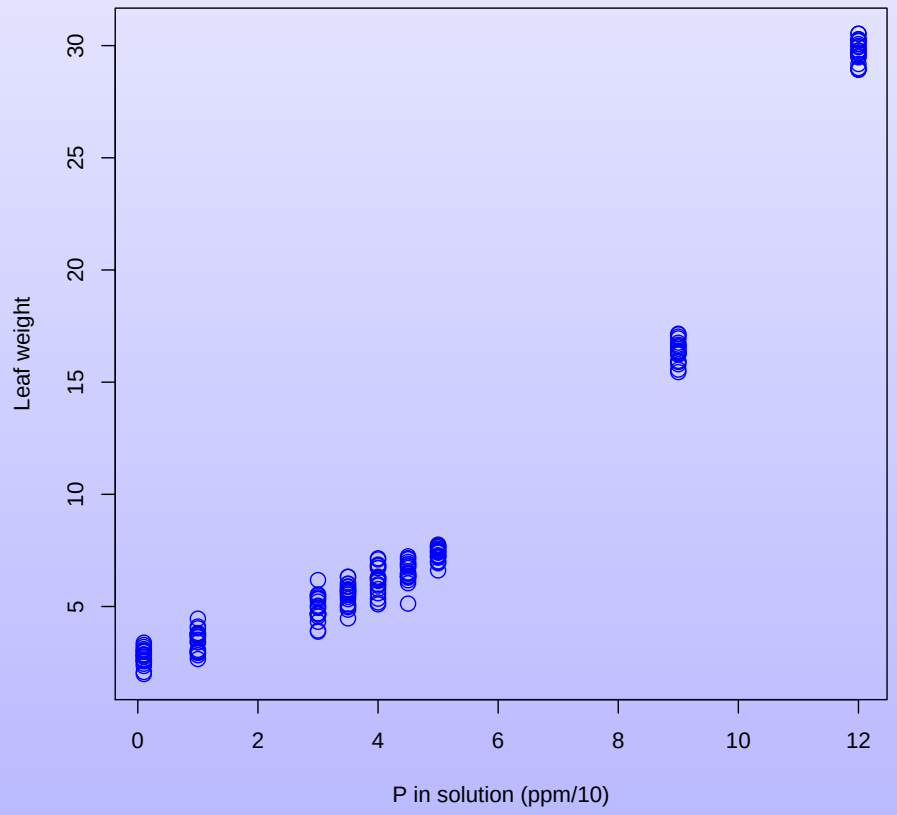
$$E\left(Y_{pr}\right) = \alpha + \beta p$$

- We assume, moreover, that Y_{pr} is normally distributed
  with constant variance and
  that the observations are independent

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Review, Normal non-linear regression models

Maize response to P



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Review, Normal non-linear regression models

Maize response to P

- Y_{pr} weight of the r-th repetition subject to the amount p of Phosphorous
- We assume that

$$\log\left(E\left(Y_{pr}\right)\right) = \alpha + \beta p$$

or equivalently,

$$E\left(Y_{pr}\right) = \exp\left(\alpha + \beta p\right)$$

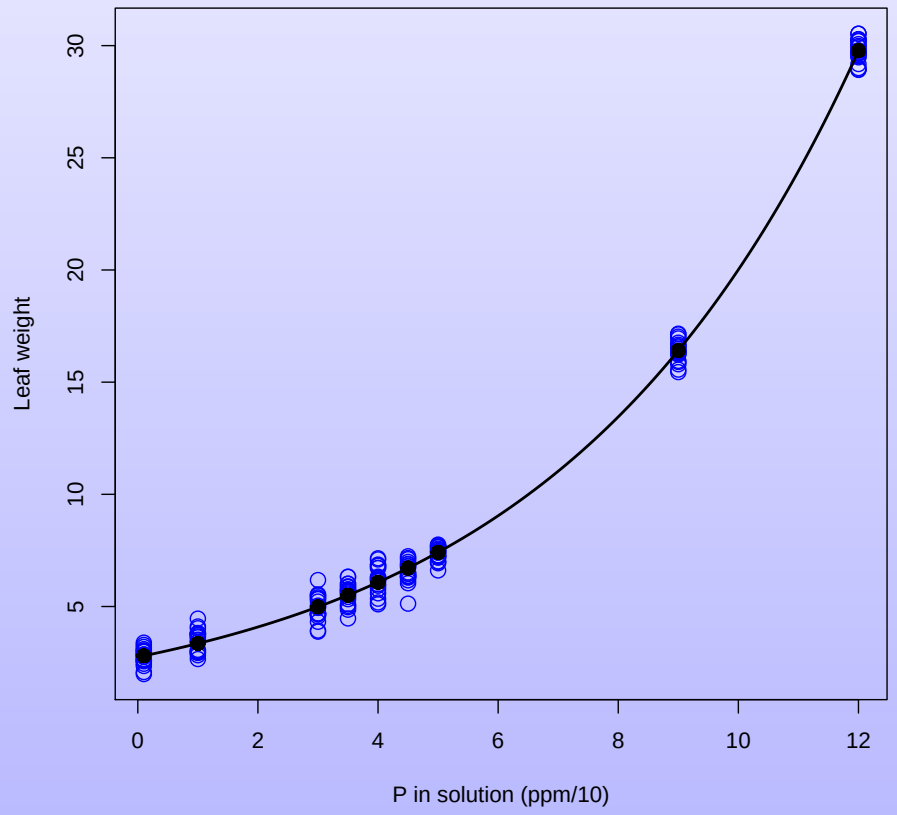
- We assume, moreover, that Y_{pr} is normally distributed with constant variance and

  that the observations are independent

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Review, Normal non-linear regression models

Maize response to P



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

GLM: basic setup

- Response variable (continuous or discrete)
- A range of explanatory variables:
  some continuous and some discrete
- Independent observations

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Classic normal linear model

- The typical formulation of a (simple normal) linear model is of the form

$$Y = X\beta + \epsilon$$

where Y is the response variable
X represents some explanatory variables
β is a vector of parameters
ϵ represents the residuals assumed to be normally distributed with mean zero and constant variance

- It is easy to see that

$$E(Y) = X\beta$$

i.e. the expected value of Y is a linear combination of the explanatory variables

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

- We might alternatively define the linear model with three assertions:
  - Y is normally distributed (with constant variance)
  - The mean of Y follows the relation

$$E(Y) = X\beta$$

  - The observations are independent
- Generalized linear models are defined also with three assertions:
  - Y is distributed according to a distribution contained in the class of the exponential dispersion models
    Examples: **Normal**, Gamma, Poisson, binomial, etc
  - The mean of Y follows the relation

$$g\left\{E(Y)\right\} = X\beta$$

    where g is a smooth monotone function
    (monotone = increasing or decreasing; smooth = has continuous derivatives of all orders)
  - The observations are independent
- g is called the *link function*
  $X\beta$ is called the *linear predictor*

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

GLM: three basic elements

- The distribution
  Exponential dispersion models
- The linear predictor
  The expectation can be described as a function of a linear combination of the explanatory variables
- The link function

  The function that connects the expectation with the linear predictor

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Exponential Dispersion Models

- We introduce now a range of families of probability distributions that allow a generalization of the model described:
  The "Exponential Dispersion Models" (EDM)
- They include many classic parametric families:
  normal, gamma, inverse gaussian, Poisson, negative binomial, binomial, etc
- EDM have many properties in common with the family of normal distributions
- EDM are the basis of Generalized Linear Models:
  linear, logistic, Poisson regressions, etc

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Exponential Dispersion Models formal definition (not important at this level)

- A family of distributions with density or probability function of the form:

$$p(y;\theta,\sigma^2) = \exp\left[\frac{y\theta - b(\theta)}{\sigma^2} + c(y,\sigma^2)\right],$$

  is called an *exponential dispersion model* (EDM) (Jørgensen, 1987)
  Here θ and σ^2 are parameters indexing the family.
- If Y is distributed according to an EDM, then

$$\mathrm{E}(\,Y) = \mu = b'(\theta) \text{ and } \mathrm{Var}(\,Y) = \sigma^2 b''(\theta) = \sigma^2 V(\mu)$$

  The function $V(\,\cdot\,)$ **characterizes uniquely** the EDM!
- EDM can be parametrized by the mean μ and the scale σ^2.

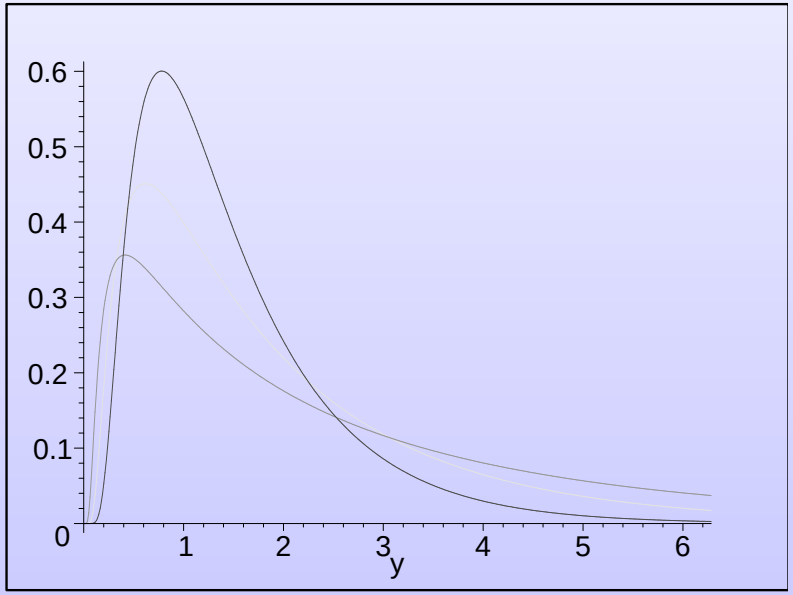
$$Y \sim \mathrm{ED}(\mu,\sigma^2)$$

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Examples of continuous EDM:

- The normal distribution is an exponential dispersion model (EDM)
  Using the normal distribution and the identity link function yields the classic normal model
- A gamma distribution is an EDM
  Gives a model suitable for positive skewed data with constant coefficient of variation
  Classic example: growth
- Inverse Gaussian: time for a Brownian motion hits a barrier
  Typical example: liquid percolation through a membrane
  Meat drip loss

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Some Exponential Dispersion Models

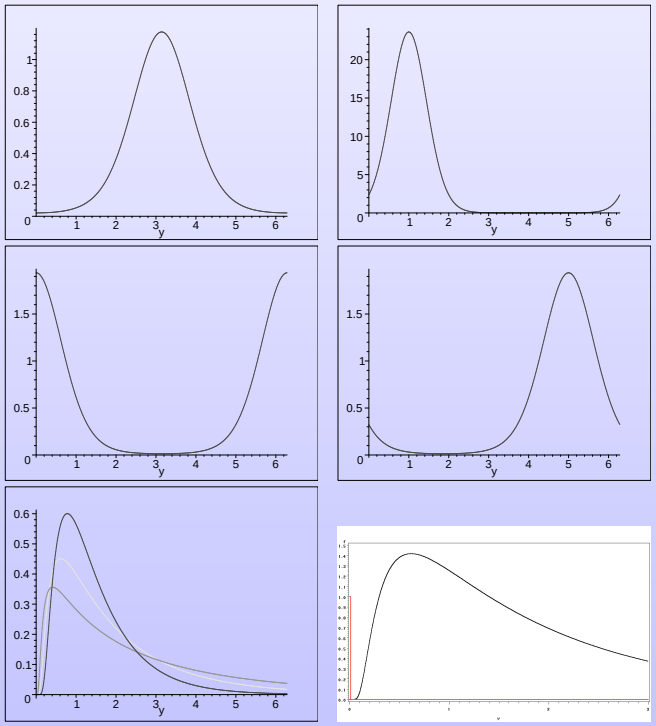


Figure: Density of some Exponential Dispersion Models

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Examples of discrete EDM:

- Binomial $\rightarrow$ proportions (logistic, probit regressions, ets)
- Poisson $\rightarrow$ counting data
- Negative binomial $\rightarrow$ waiting time for n successes in a Bernoully essay
  Alternative for counting data
- Compound Poisson
  Positive responses with zeroes

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Example: activity of toposisomerase for detecting malaria

- Experiment on activity of *Plasmodium* topoisomerase in blood
- Used to detect malaria
  (with hight sensitivity and specificity)
- Steps of the laboratory analysis:
  - Step 1: Extraction of topoisomerase from the blood
  - Step 2: Series of reactions to form a detectable product producing a colour in the suspension
  - Step 3: Measures the absorbance (at a specific colour)
- Preliminary study for the article "Droplet Microfluidics Platform for Highly Sensitive and Quantitative Detection of Malaria-Causing Plasmodium Parasites Based on Enzyme Activity Measuremet" by Juul, Nielsen, Labouriau et al. *ACS Nano* 2012.

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria:

another preliminary experiment, now for characterising sources of variability

- Steps of the laboratory analysis:
  - Step 1: Sampling and extraction of topoisomerase from the blood
  - Step 2: series of reactions to form a detectable product producing a colour in the suspension
  - Step 3: Measures the absorbance (at a specific colour)
- Now we would like to characterise how much of the variability of the final results (step 3) arises from step 1 and step 2
- New (preliminary) experiment:
  10 extractions (step 1) of the same blood pool (patient) for two patients
  For each extraction 5 separate reactions (step 2) were performed
  For each of the 50 combinations of extraction and reactions 5 colorimetric determinations were performed
- In total 2x10x5x5=500 observations were generated

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: a model with random components

- We do not expect large systematic differences in the means of results obtained with different extractions or reactions from the same patient
- The observations are probably not independent, since some samples are taken from the same blood sample of the same patient ...
- A way to circumvent this problem (dependency) is to insert in the model two variables representing a common latent effect of the extraction and the reaction
- Two consequences of inserting these variables in the model:
  - The model then accounts for possible dependencies of the observations
  - The total variability can be decomposed in different sources of variability
- Such a model is called a mixed model

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: examining the data

```
> str(PrelMalaria)
'data.frame':        500 obs. of  4 variables:
 $ Patient   : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ Extraction: Factor w/ 10 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1
 $ Reaction  : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 2 2 2 2
 $ Absorbance: num  49.6 50.6 51.7 51.2 51.5 ...
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: examining the data

```
> table(Reaction, Extraction, Patient)

 Patient = 1                                  Patient = 2

        Extraction                                   Extraction

Reaction 1 2 3 4 5 6 7 8 9 10   Reaction 1 2 3 4 5 6 7 8 9 10

       1 5 5 5 5 5 5 5 5 5  5          1 5 5 5 5 5 5 5 5 5  5

       2 5 5 5 5 5 5 5 5 5  5          2 5 5 5 5 5 5 5 5 5  5

       3 5 5 5 5 5 5 5 5 5  5          3 5 5 5 5 5 5 5 5 5  5

       4 5 5 5 5 5 5 5 5 5  5          4 5 5 5 5 5 5 5 5 5  5

       5 5 5 5 5 5 5 5 5 5  5          5 5 5 5 5 5 5 5 5 5  5
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: examining the data

```
> summary(PrelMalaria)
 Patient   Extraction  Reaction   Absorbance
 1:250   1      : 50   1:100    Min.   :34.95
 2:250   2      : 50   2:100    1st Qu.:49.46
         3      : 50   3:100    Median :52.82
         4      : 50   4:100    Mean   :53.92
         5      : 50   5:100    3rd Qu.:59.35
         6      : 50            Max.   :76.01
         (Other):200
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: fitting a suitable model

```
> library(lme4)
> Fit <- lmer(Absorbance ~ Patient + 0 + (1|Extraction) +
+                                          (1|Reaction:Extraction))
```

. Assumptions:

- The observations $\mathcal{Y}_{peri}$s are normally distributed
  (for patient $p = 1, 2$, extraction $e = 1, \ldots, 10$, reaction $r = 1, \ldots, 5$ and repetition $i = 1, \ldots, 5$)
- $\mathcal{Y}_{peri} = \mu_p + \mathcal{X}_e + \mathcal{Z}_{r,e} + \mathcal{E}_{peri}$
  with $\mathcal{X}_e \sim N(0, \sigma_e^2)$, $\mathcal{Z}_{r,e} \sim N(0, \sigma_r^2)$ and $\mathcal{E}_{peri} \sim N(0, \sigma^2)$ independent
- Consequences:
  $Var(\mathcal{Y}_{peri}) = \sigma_e^2 + \sigma_r^2 + \sigma^2$
  Observations from the same extraction and same reaction are correlated!
  (therefore not assumed to be independent!)

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Calculation of the variances and covariances

$$y_{peri} = \mu_p + X_e + Z_{re} + \varepsilon_{peri}$$

$$\left.\begin{array}{l} X_e \sim N(0, \sigma_e^2) \\ Z_{re} \sim N(0, \sigma_r^2) \\ \varepsilon_{peri} \sim N(0, \sigma^2) \end{array}\right\} \text{Independent}$$

$$Var(y_{peri}) = Var(\mu_p + X_e + Z_{re} + \varepsilon_{peri})$$
$$= \sigma_e^2 + \sigma_r^2 + \sigma^2$$

$$Cov(y_{peri}, y_{p'eri'})$$
$$= Cov(\mu_p + X_e + Z_{re} + \varepsilon_{peri}, \mu_{p'} + X_e + Z_{re} + \varepsilon_{peri'})$$
$$= 0 + 0 + 0 + 0 + 0 + \sigma_e^2 + 0 + 0$$
$$0 + 0 + \sigma_r^2 + 0 + 0 + 0 + 0 + 0$$

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: fitting a model

```
> library(lme4)
> Fit <- lmer(Absorbance ~ Patient + 0 + (1|Extraction) +
+                                        (1|Reaction:Extraction))
> summary(Fit)
Linear mixed model fit by REML ['lmerMod']
Formula: Absorbance ~ Patient + 0 + (1 | Extraction) + (1 | Reaction:Ex
...
Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.33789 -0.65969 -0.00224  0.60396  2.56435
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: fitting a model

```
> library(lme4)
> Fit <- lmer(Absorbance ~ Patient + 0 + (1|Extraction) + (1|Reaction:E
> summary(Fit)
...
Random effects:
 Groups               Name        Variance Std.Dev.
 Reaction:Extraction (Intercept)  5.222   2.285
 Extraction          (Intercept) 12.259   3.501
 Residual                        16.533   4.066
Number of obs: 500, groups:  Reaction:Extraction, 50; Extraction, 10
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: quantifying the sources of variability

```
> library(lme4)
> Fit <- lmer(Absorbance ~ Patient + 0 + (1|Extraction) +
+                                          (1|Reaction:Extraction))
```

. Assumptions:

- $\mathcal{Y}_{peri} = \mu_p + \mathcal{X}_e + \mathcal{Z}_{r,e} + \mathcal{E}_{peri}$
  with $\mathcal{X}_e \sim N(0, \sigma_e^2)$, $\mathcal{Z}_{r,e} \sim N(0, \sigma_r^2)$ and $\mathcal{E}_{peri} \sim N(0, \sigma^2)$ independent
- Random effects:
  $Var(\mathcal{Y}_{peri}) = \sigma_e^2 + \sigma_r^2 + \sigma^2 = 12.259 + 5.222 + 16.533 = 34.012$
  Observations from the same extraction and same reaction are correlated!
  (therefore not assumed to be independent!)
- The variation due to the extraction represented
  15.3% of the total variance
- The variation due to the reaction represented
  36.0% of the total variance

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: inferring the fixed effects

```
> library(lme4)
> Fit <- lmer(Absorbance ~ Patient + 0 + (1|Extraction) +
    (1|Reaction:Extraction), REML=FALSE)
> summary(Fit)
...
Fixed effects:
         Estimate Std. Error t value
Patient1   49.291      1.182   41.71
Patient2   58.554      1.182   49.55
...
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: inferring the fixed effects

```
> confint(Fit)
              2.5 %    97.5 %
.sig01     1.704909  3.063553
.sig02     2.069330  5.731202
.sigma     3.809988  4.342075
Patient1 46.871760 51.710146
Patient2 56.135188 60.973574
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: inferring the fixed effects

```
> library(lme4);library(lmerTest)
> Fit <- lmer(Absorbance ~ Patient + 0 + (1|Extraction) +
+ (1|Reaction:Extraction))
> summary(Fit)
Linear mixed model fit by REML t-tests use Satterthwaite approximations
to degrees of freedom [merModLmerTest]
...
Fixed effects:
         Estimate Std. Error     df t value Pr(>|t|)
Patient1   49.291      1.182  9.442   41.71 4.99e-12 ***
Patient2   58.554      1.182  9.442   49.55 9.88e-13 ***
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: inferring the fixed effects

```
> library(lme4);library(lmerTest)
> Fit0 <- lmer(Absorbance ~ Patient + (1|Extraction) +
+  (1|Reaction:Extraction))
> summary(Fit0)
...
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  49.2910     1.1817   9.4000   41.71 4.99e-12 ***
Patient2      9.2634     0.3637 449.0000   25.47  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Activity of toposisomerase for detecting malaria: testing fixed effects

```
> anova(Fit)
Analysis of Variance Table of type 3  with  Satterthwaite
approximation for degrees of freedom
        Df    Sum Sq   Mean Sq F value  Denom Pr(>F)
Patient  1 0.0091619 0.0091619  0.0018 484.98 0.9666
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

A Complex Design

The biomass yield was studied using an incomplete block design with six blocks; each block was split into two sub-blocks containing two experimental units (plots). In each block the control GWT management was applied in the two experimental units of one of the sub-blocks and whether the full irrigation or the intermediate irrigation regimen was applied to the two experimental units of the other sub-block. Repeated measurements at two different times of cut at two different years were performed in all the experimental units (keeping the allocation of the GWT management constant in each experimental unit).

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

A Complex Design

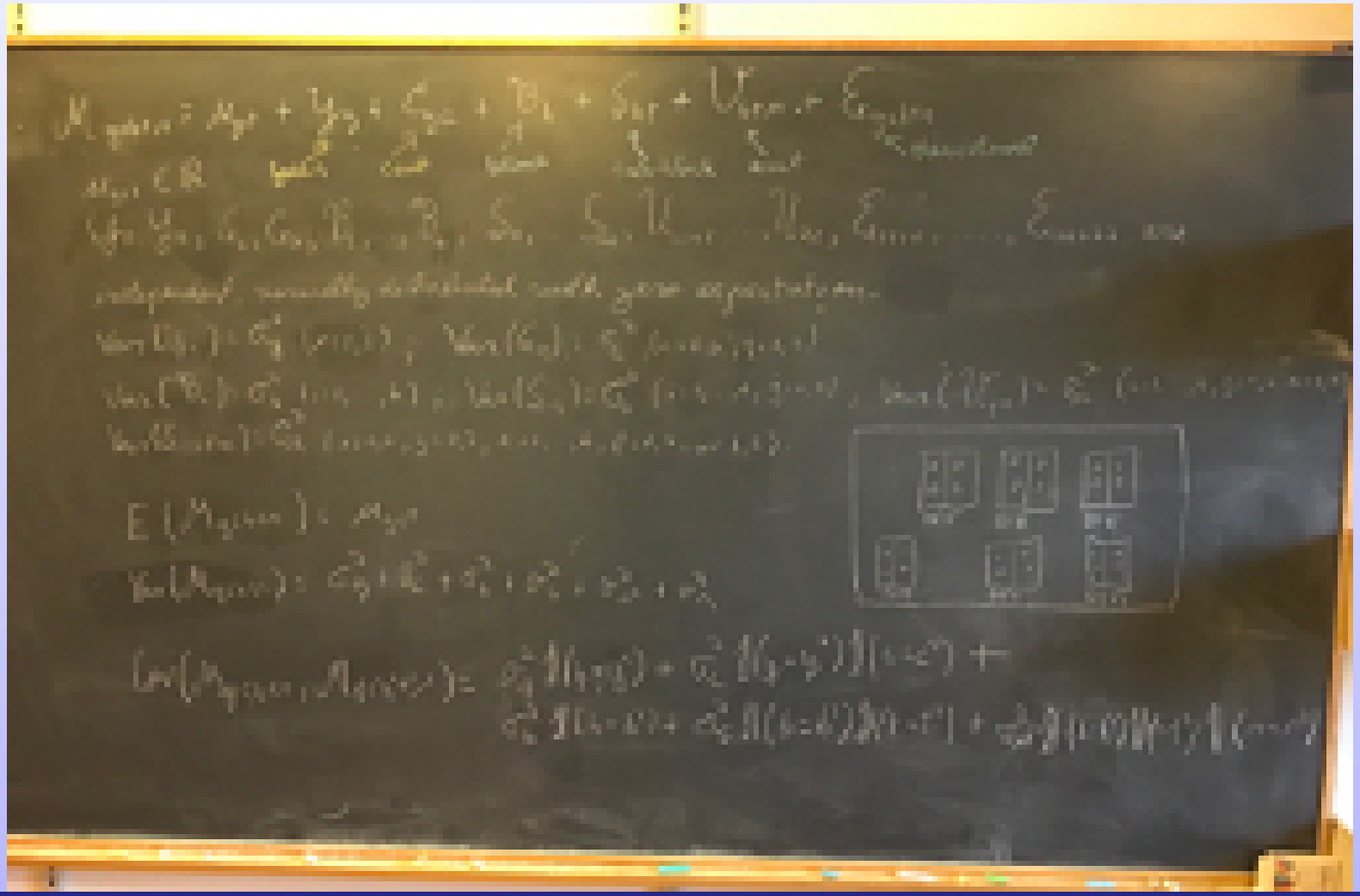
The effects of the GWT management, year and time of cut on the biomass yield was modelled using the following Gaussian mixed model. Denote by M_{ycbtr} the biomass yield the rth repetition ($\underline{r}$ = 1,2) of the cth time of cut (c = 1, 2) at the yth year (y = 20015-2016, 2016-2017) at the tth sub-block (t = "none", "intermediate" , "full" irrigation) at the bth block (b = 1, …,6) . According to the Gaussian mixed model used, for y = 20015-2016, 2016-2017, c = 1, 2, t = "none", "intermediate" , "full" , b = 1, …,6 and $\underline{r}$ = 1,2,

$$M_{ycbtr} = \mu_{yct} + Y_y + C_{\mathrm{yc}} + B_b + S_{\mathrm{bt}} + U_{\mathrm{cbtr}} + E_{\mathrm{ycbtr}} ,$$

where μ_{yct} is a fixed effect representing the mean biomass yield of the cth time of cut at the yth year of an experimental unit subject to the tth GWT management. Here Y_y, C_{yc}, B_b, S_{bt}, U_{cbtr} and E_{ycbtr} are Gaussian independent random components representing the year, cut time for each year, block, sub-block, experimental unit and the residual variation, respectively. The statistical inference of the Gaussian mixed model described above was performed using the 'lme4' package in R (R Core CRAN, 2016).

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

A Complex Design



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

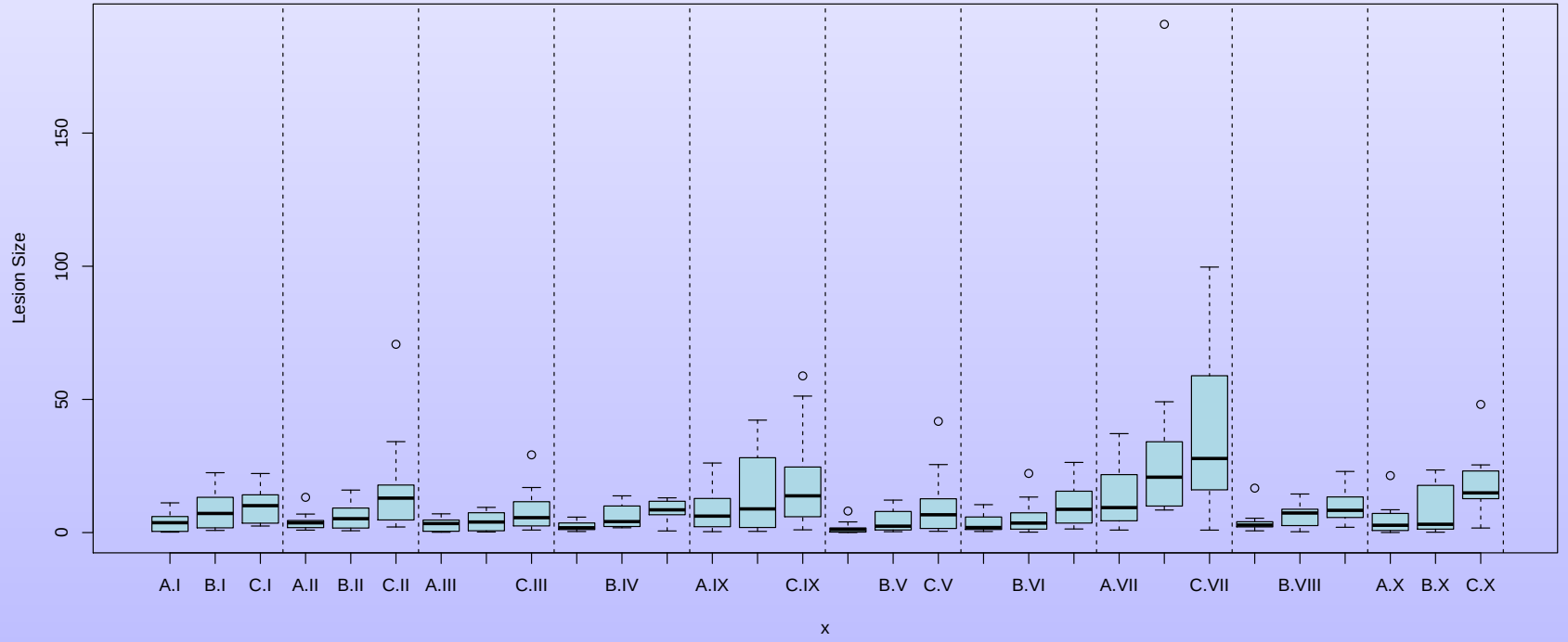
Initial Example of Non-Gaussian Models - Fungal resistance essay

- Several measurements of fungal resistance in a cultivated plant
- Three fungal strains: A, B and C.
- 10 plants, 10 repetitions (leaves) inoculated
- Responses:

  Lesion size
- Different leaves used for the three determinations
- We analyse the lesion sizes in detail

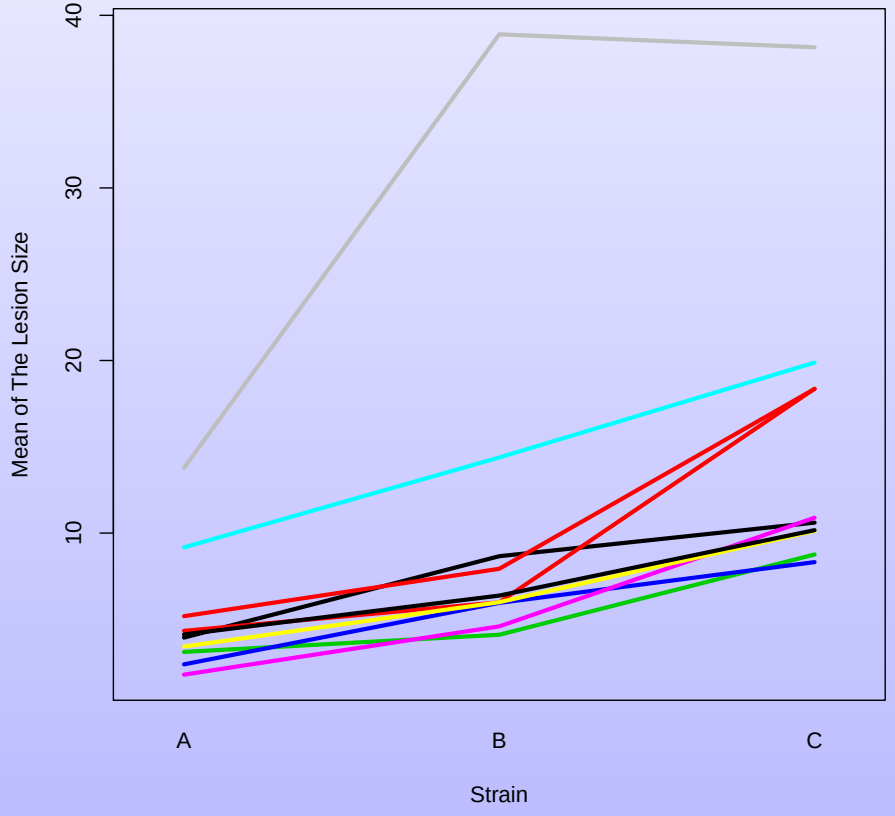
 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Initial Example of Non-Gaussian Models - Fungal resistance essay



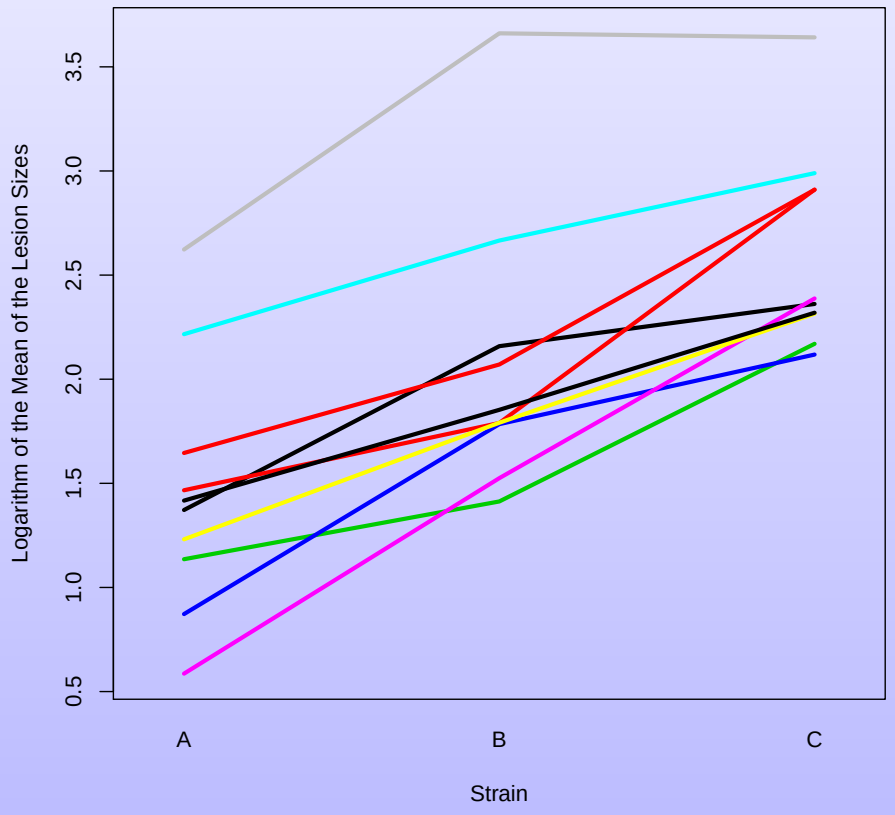
 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Initial Example of Non-Gaussian Models - Fungal resistance essay



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Initial Example of Non-Gaussian Models - Fungal resistance essay



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Lesion size, modelling

- We will use a GLMM defined with the gamma distributions and a logarithmic link function
- The model will contain a fixed effect representing the effect of the strains and

  a random component representing the plant.

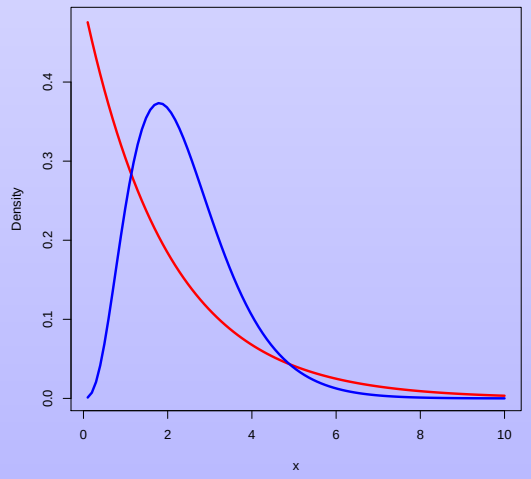
 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Lesion size, the Gamma distribution

- A probability distribution on the positive real numbers with probability density of the form, for $\alpha > 0$ and $\beta > 0$,

$$p(y; \alpha, \beta) = y^{\alpha-1} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \exp\left(-y/\beta\right), \text{ for } y > 0,$$

  is said to be a *Gamma distribution*. Notation $X \sim G(\alpha, \beta)$
- The parameters $\alpha > 0$ and $\beta > 0$ are called the *shape* and the *scale* parameters, respectively.



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Lesion size, defining a GLMM

- Denote by $\mathcal{Y}_{b,t,r}$ the random variable representing the lesion size of the rth replicate ($r = 1, \ldots, 10$) of the experimental units of the bth plant (or cluster, $b = I, \ldots, X$) that received the tth strain ($t = A, B, C$).
- Suppose that there exist $\mathcal{U}_I, \ldots, \mathcal{U}_X$ i.i.d. random variables , with $\mathcal{U}_1 \sim N(0, \sigma_U^2)$ such that, $\mathcal{Y}_{I,A,1}, \ldots, \mathcal{Y}_{X,C,10}$ are conditionally independent and Gamma distributed given $\mathcal{U}_1, \ldots, \mathcal{U}_B$, and for $b = 1, \ldots, X$, $t = A, B, C$ and $r = 1, \ldots, 10$,

$$\log\left\{E\left(\mathcal{Y}_{btr} | \mathcal{U}_b = u\right)\right\} = \tau_t + u, \quad \text{for all } u \in \mathbb{R},$$

or equivalently,

$$E\left(\mathcal{Y}_{btr} | \mathcal{U}_b = u\right) = \exp\left(\tau_t + u\right) = \exp(\tau_t)\exp(u), \quad \text{for all } u \in \mathbb{R}.$$

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Lesion size, Fitting the model in R

```
> library(GLMMstudy)
> data("FungusResistance")
> D <- FungusResistance
> M <- glmer(LesionSize ~ Strain + 0 + (1|Plant),
+            family = Gamma(link = "log") ,data = D)
> summary(M)
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Lesion size, Fitting the model in R

```
Random effects:

 Groups   Name        Variance Std.Dev.

 Plant    (Intercept) 0.2507   0.5007

 Residual             0.9753   0.9876

Number of obs: 300, groups:  Plant, 10

Fixed effects:

        Estimate Std. Error t value Pr(>|z|)

StrainA   1.4698     0.1869   7.864 3.73e-15 ***

StrainB   2.0929     0.1871  11.185  < 2e-16 ***

StrainC   2.6252     0.1869  14.044  < 2e-16 ***
```

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Three Examples of Non-Gaussian Models - Number of spots, some inference, marginal means

```
> ( FixedEffects <- M@beta ); exp(FixedEffects);exp(FixedEffects) * exp(0.2507/2)
```

```
[1] 1.4698  2.0929  2.6252
```

```
[1]  4.348365  8.108395 13.807335
```

```
[1]  4.929068  9.191232 15.651238
```

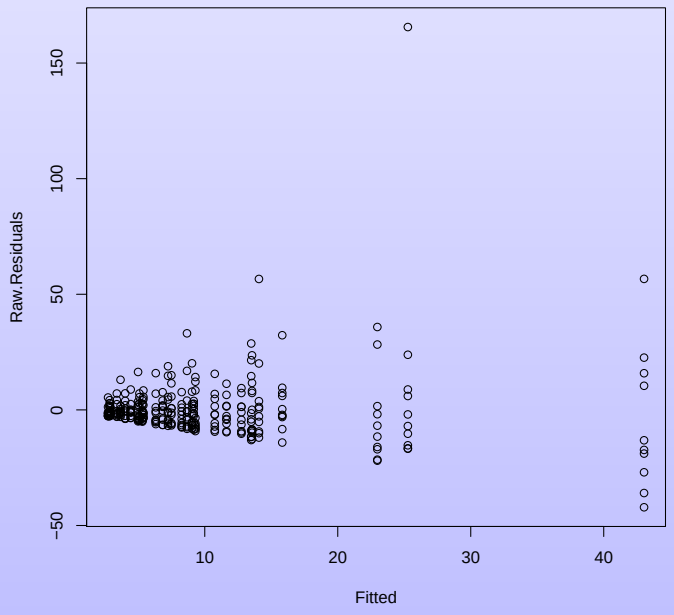
$$
\begin{aligned}
E\left(\mathcal{Y}_{btr}\right) &= E_{\mathcal{U}}\left\{E\left(\mathcal{Y}_{btr}|\mathcal{U}_b\right)\right\} = \int_{-\infty}^{\infty} \exp\left(\tau_t + u\right)\, \phi(u; 0, \sigma_U^2)\, du \\
&= \exp\left(\tau_t\right) \int_{-\infty}^{\infty} \exp\left(u\right)\, \phi(u; 0, \sigma_U^2)\, du = \exp\left(\tau_t\right) \exp\left(\sigma_U^2/2\right) .
\end{aligned}
$$

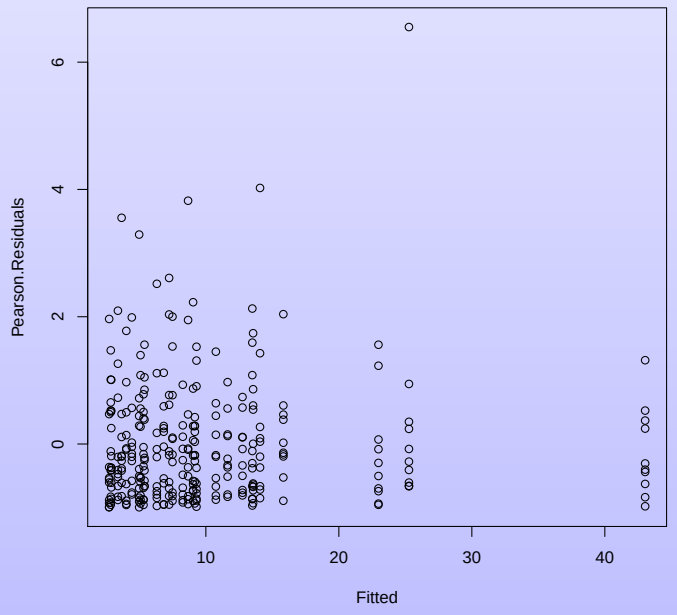
Here we can argue that the expectation of a log-normal distribution $logN(0, \sigma^2)$ is $\exp\left(\sigma^2/2\right)$ or that the MGF of $N(0, \sigma^2)$ is $\exp(\sigma^2/2)$. Note that $exp(0.2507/2) \approx 1.133545$.

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Lesion size, model control

The trombone form of the graph of the raw-residuals against the fitted values is what one expects for a model based on the Gamma distribution; this pattern is not present in the graph of the Pearson-residuals against the fitted values





 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Three Model Control Techniques

a test of goodness of fit specially designed for Gamma GLMs and GLMMs

- If a continuous real random variable Y has cumulative distribution function F_Y (that is, F_Y is the function defined by $F_Y(y) = P(Y \leq y)$, for all $y \in \mathbb{R}$),

  then $Z = F_Y(Y) \sim U(0,1)$.

- We will take advantage of this fact to construct a notion of generalised residuals
- The *uniform residuals*, obtained by applying the cumulative distribution function of the Gamma distribution to the observations

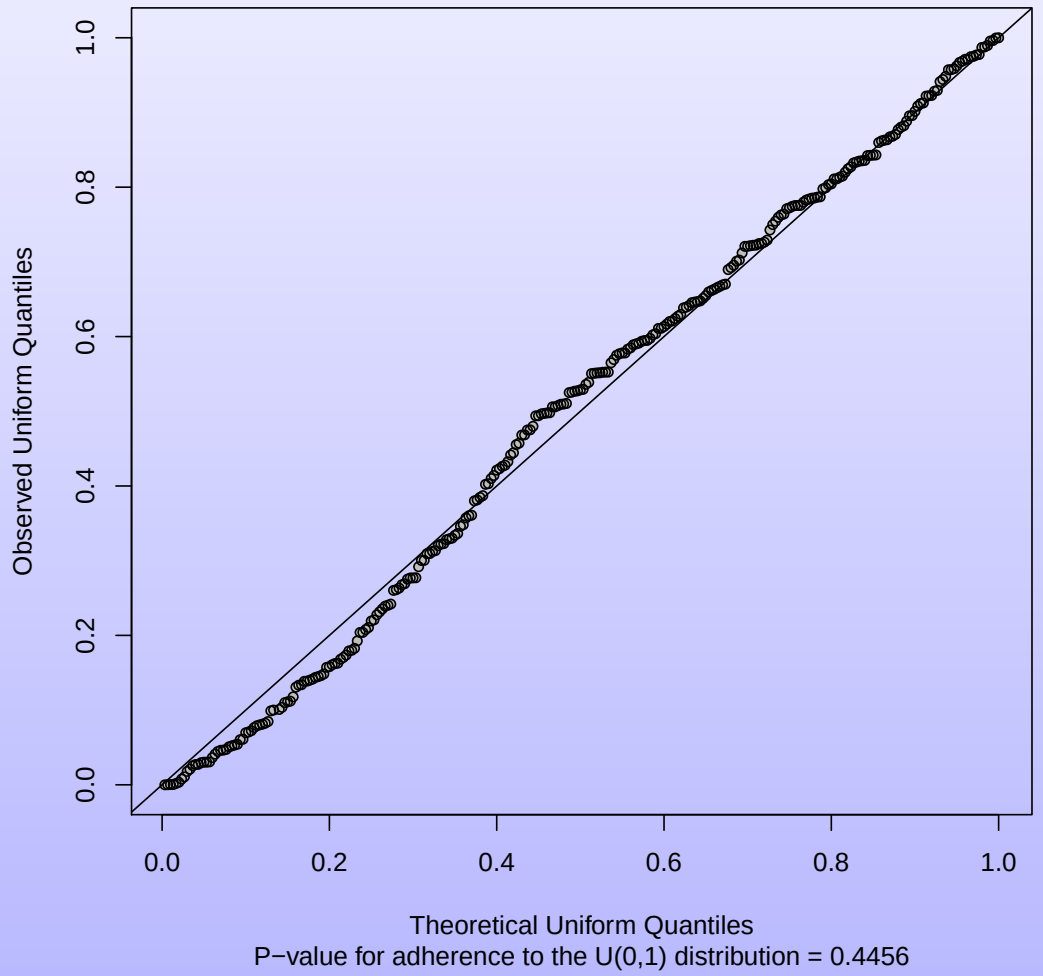
  (different functions will be used for different observations with different means).

- We test then whether the uniform residuals follow the standard uniform distributions.

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Three Model Control Techniques

a test of goodness of fit specially designed for Gamma GLMs and GLMMs



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Mixed Models

- Mixed models describe the relationship between a response variable and some explanatory variables
- In this course we considered only categorical (classification) response variables, called **factors** with several classification **levels**
- The parameters (coefficients) associated with the classification levels of a factor are called **effects**
- In mixed models, we define two types of effects:
  - fixed: affect the assumed mean of the observations
  - random effects: affect the assumed structure of the variance (and covariance)
- It is relatively easy to fit mixed models in R, but the definition of the model and the interpretation might be tricky!

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Population genetic example

- Does inbreeding affect human fertility?
  How much?
- The Danish study based on the cohort of all women born in Denmark in 1954
  who were alive and living in Denmark in 1969
  totaling 42,165 women.
- The cohort was followed up to the end of 1999.
  Fertility: The number of children born to each mother had between the ages of 15 and 45 years old was determined

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

- Fertility: The number of children born to each mother between the ages of 15 and 45 years old was determined
- Genetic distance (proxy):
  The mean marital radius (MR) associated with each mother in the cohort was estimated using the distance between the centroids of the parish where she was born, and the parishes where the partners with which she had children were born
- Classic measure of human genetic proximity (Malecot, etc)
- The Spearman correlation between the MR and the fertility in the cohort was 0.38 ($P < 0.0001$), indicating a positive association
- Form of the response of fertility to MR: Unknown
  Relationship between MR and Inbreeding: exponential
  Relationship between inbreeding and Fertility: unknown
- Labouriau and Amorim (2008) *Genetics* **178**: 601:606

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

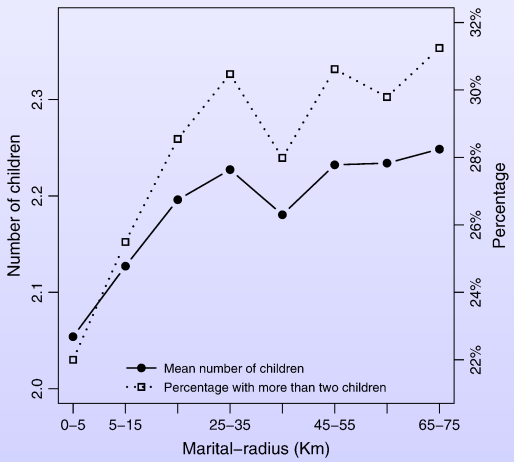


FIGURE 1.—Mean number of children (continuous line and left scale) and percentage of mothers that had more than two children (interrupted line and right scale) determined for mothers with marital radiuses contained in different intervals arranged (in scale) along the horizontal axis.

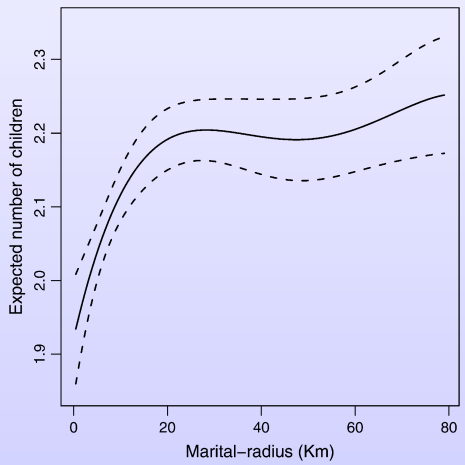


FIGURE 2.—Expected number of children as a function of marital radius (and the limits of a pointwise 95% C.I., interrupted lines), inferred by a truncated polynomial Poisson regression.

The association was further characterized by fitting a truncated Poisson regression for predicting the number of children as a fourth degree polynomial function of the marital radius. A likelihood ratio test for checking constancy of the expected number of children based on the regression model above formally confirmed the reported association (P-value < 0.0001). Moreover, visual inspection of the graph of the expected number of children as a function of the marital radius (Figures 1 and 2) confirms that fertility and marital radius are positively associated.

Fertility, measured in a complex population such as this, can be affected by socioeconomic factors. Therefore we also performed a conditional analysis involving three key socioeconomic indicators: education, income, classified in an ordered scale with five levels: 1, countryside with low population density; 2, town with <20,000 inhabitants; 3, town with 20,000–39,999 inhabitants; 4, city with 40,000–99,999 inhabitants; and 5, city with >100,000 inhabitants (including the capital and its surroundings).

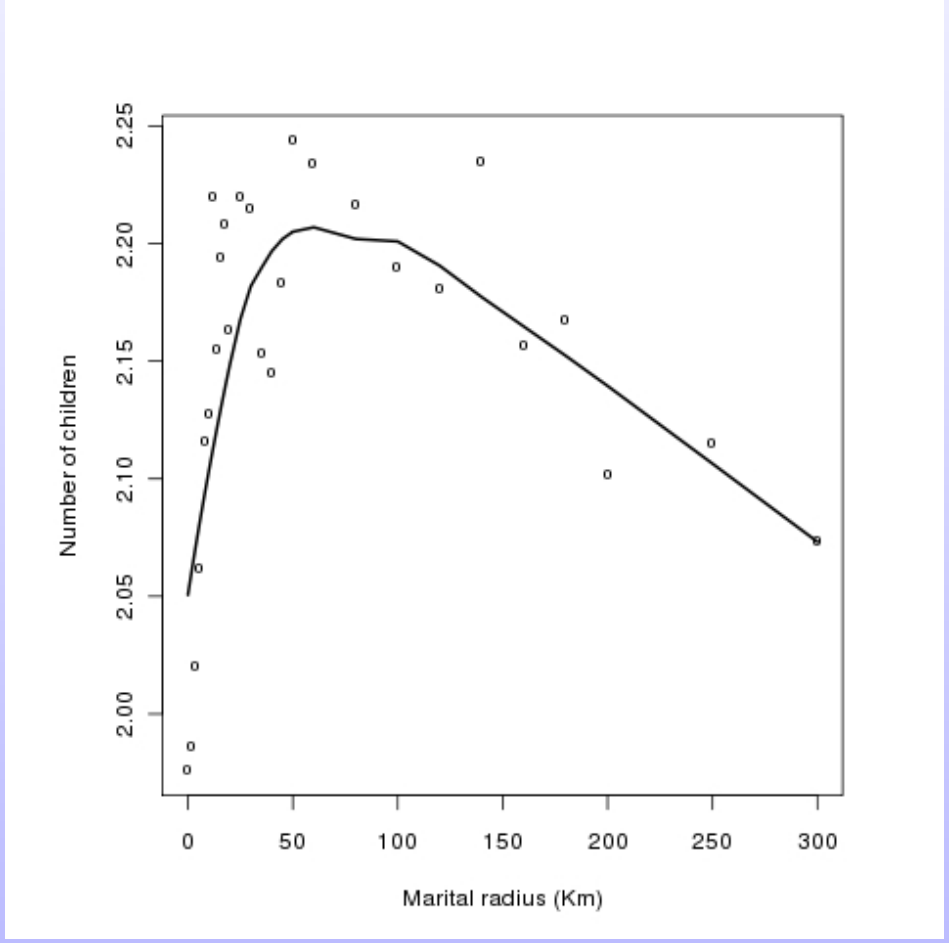
The Spearman partial correlation between the number of children and marital radius conditional on urbanicity, education, and income is 0.041 (P-value < 0.0001), indicating that the raw positive association between fertility and marital radius reported above is not a mere artifact due to spurious association with these socioeconomic factors.

We take advantage of the theory of graphical models to extract further relevant aspects of the correlation structure and the distribution of the information between

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

- Discussion on effect of outbreeding X inbreeding
- Interest on estimating effect of MR on fertility for less related matings
- Nonparametric regression
- Labouriau and Amorim (2008). *Science*

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate



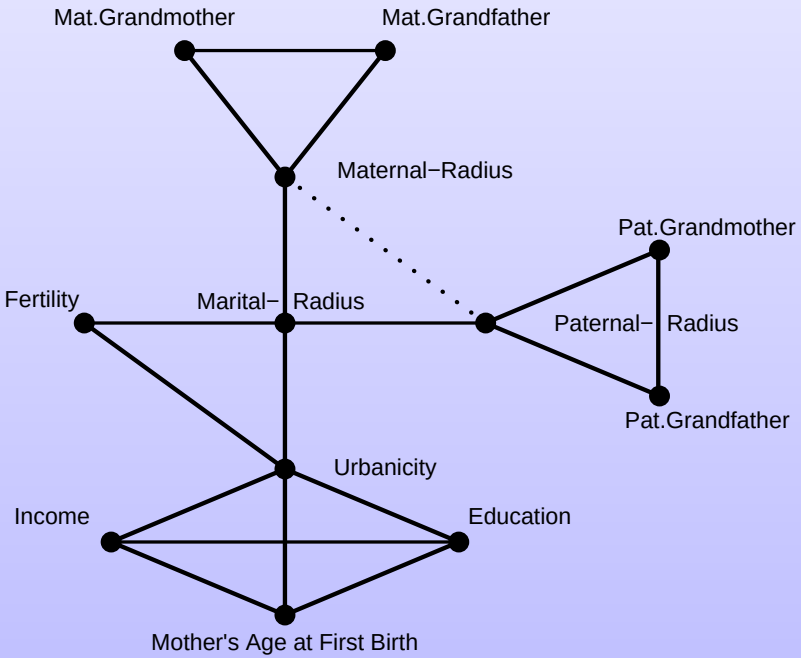
 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Multivariate

- Discussion on effect of outbreeding X inbreeding
- Nonparametric regression
- Are the effects of MR on fertility socially determined?
- The Danish registers contain information on socioeconomic variables at **individual** level!
- Multivariate analysis: graphical models
  Discussion on information distribution

  Labouriau and Amorim (2008). *Science*

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

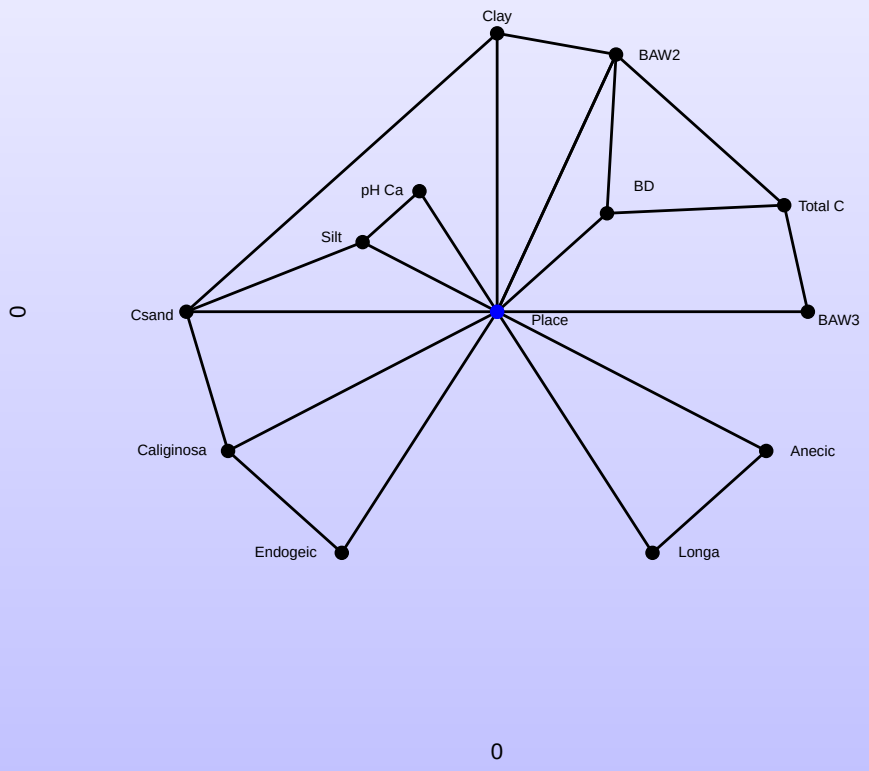


 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Multivariate - other examples

- Worm distribution in soil and soil texture
- Very large network: Gene expression in infected pigs

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

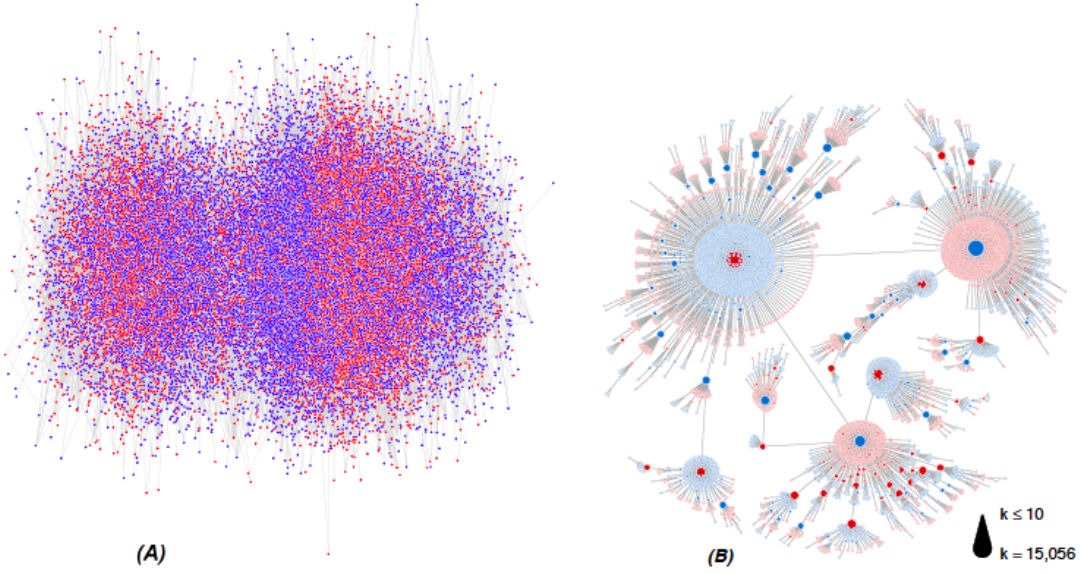


Figure 3: Co-expression network of Example 1. (A) Raw representation of the decomposable graphical model with minimum BIC; red and blue points represent differentially expressed and non-differentially expressed probes, respectively. (B) Network representation obtained by the clustering procedure; each point represents a cluster, which size is proportional to the number of probes in the cluster; clusters with predominance of differentially expressed probes are marked in red and the others in blue.

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Couts of Collembola

- The data we will analyse in this third example consists of counts of several species of Collembola or spring tail in soil.
  This data was produce by Alessandra D'Annibbale at the Department of Agroecology, Aarhus University
- The main interest is to compare and characterise the abundance of those animals when subject to five different treatments.

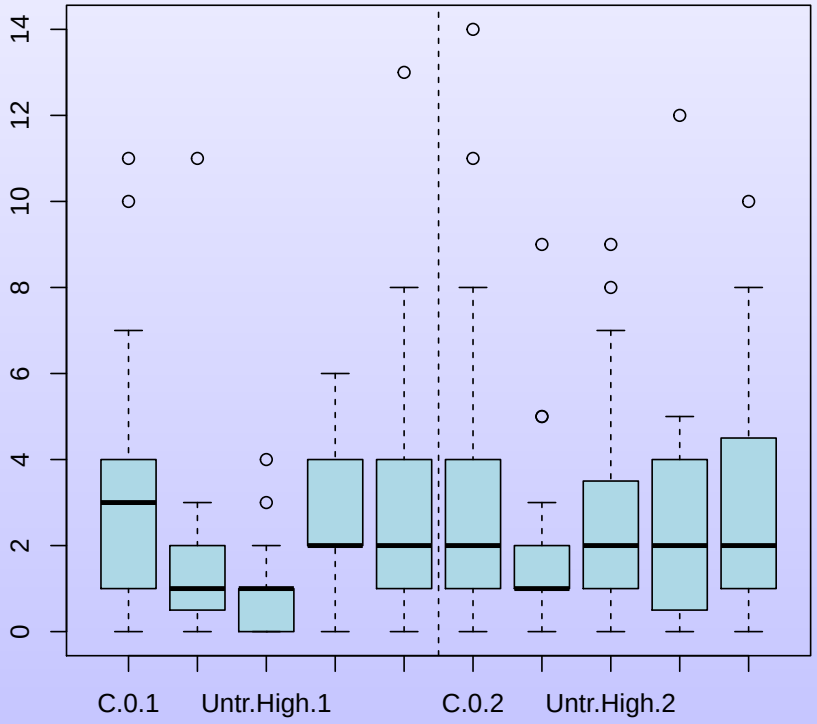
 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Couts of Collembola

- A naive modelling describing the counts of Collembola as normally distributed fails to describe the data well and to detect differences between the treatments, differences that are indeed clearly visible.
- Less naive alternative, assuming a Poisson distribution, also fail describe the data well and to detect differences between the treatments!

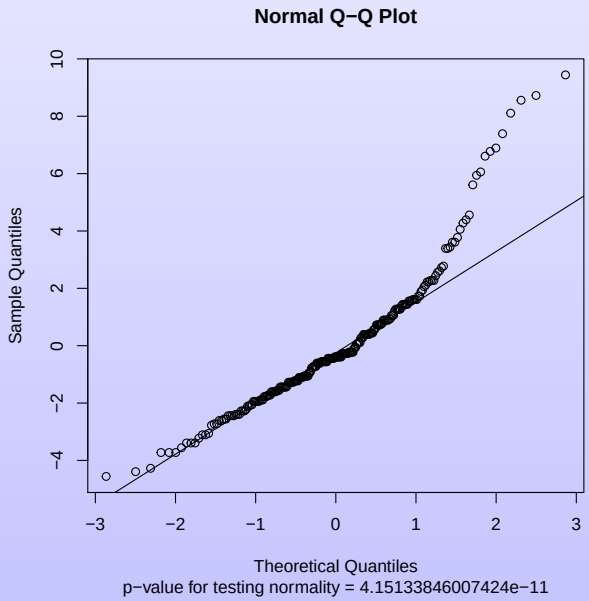
 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

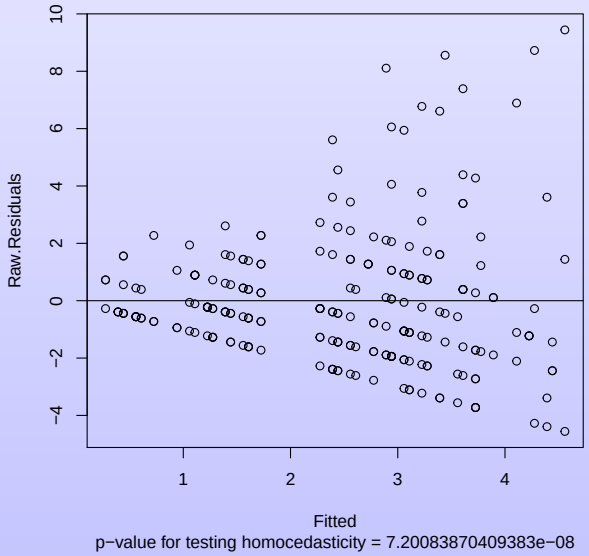
A Poisson model with random components ...



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

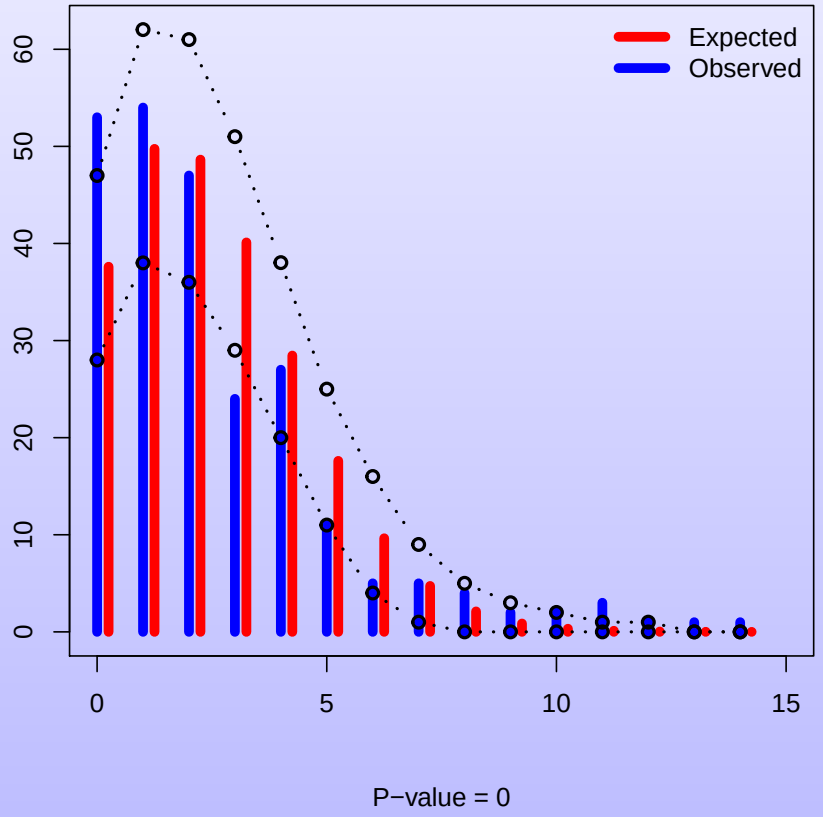
Model control for a Gaussian Model





 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

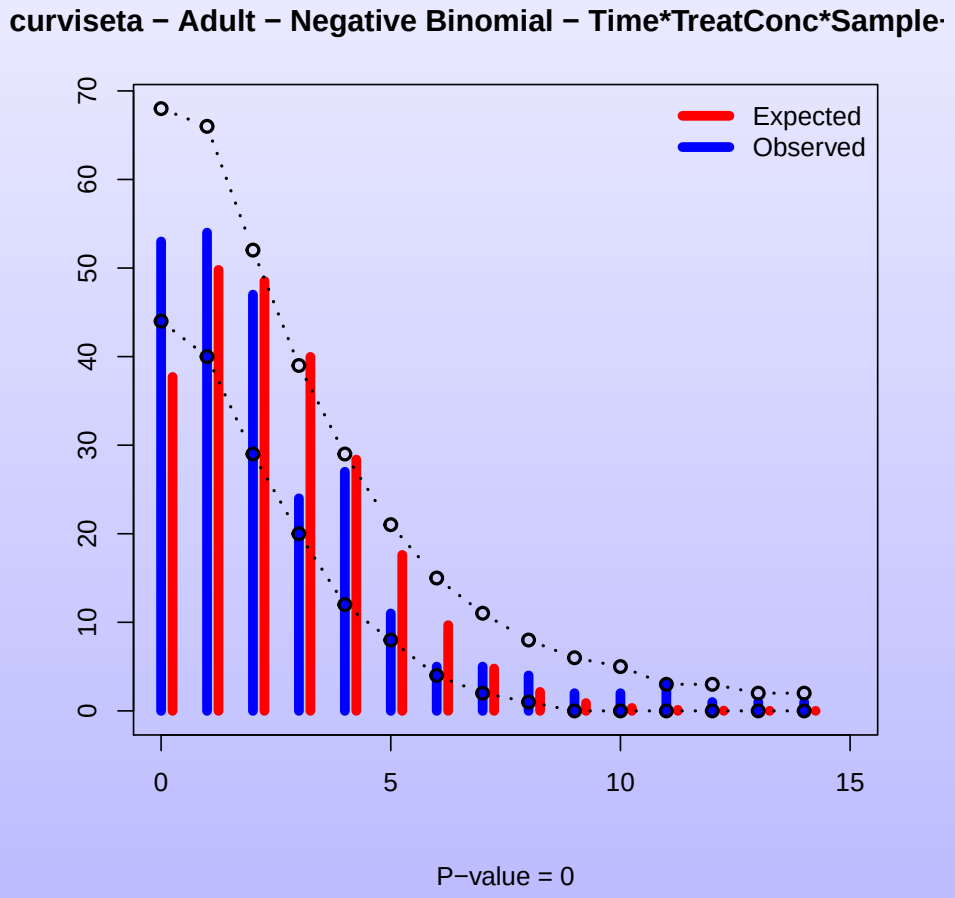
Model control for a Poisson Model



S. curviseta − Adult − Time*TreatConc*Sample+Block

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Model control for a Negative Binomial Model



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

A Poisson model with random components ...

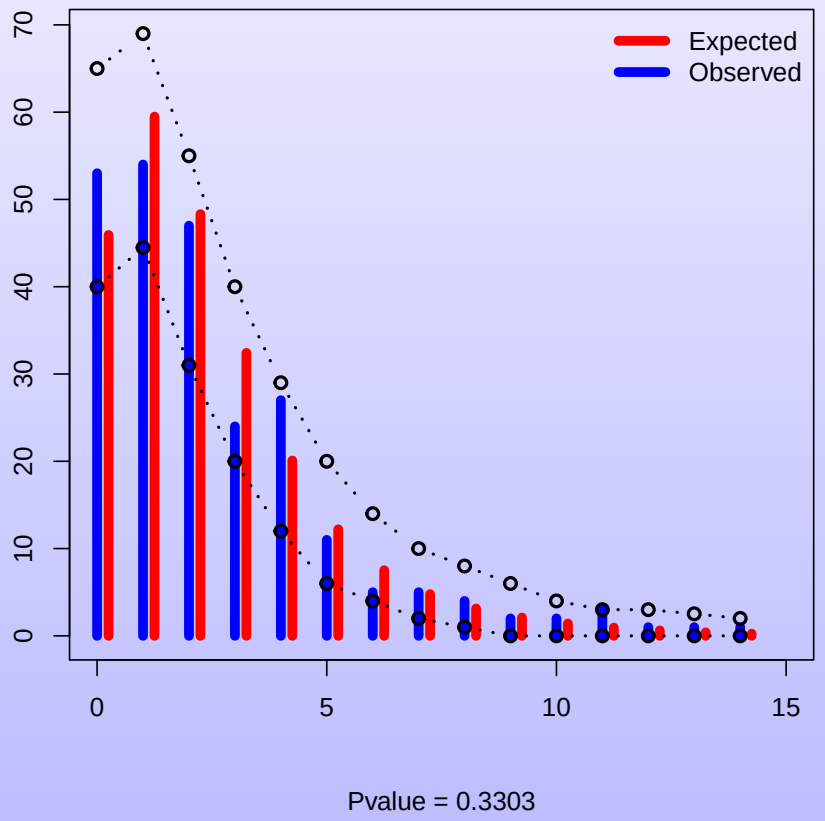
- Assume the observations (i.e. the counts) are conditionally Poisson distributed given two Gaussian random components ζ and ϵ. The conditional expectations of the observations are then given by ....

$$\log\left\{E\left(Y_{tcsbr}|\zeta = z_{tcbr}, \epsilon = e_{tcsbr}\right)\right\} = I_{tcs} + B_b + z_{tcbr} + e_{tcsbr}\,.$$

- The two random components are independent, normally distributed with $\zeta \sim N(0, \sigma_u^2)$ and $\epsilon \sim N(0, \sigma^2)$.
- The random component ζ represents the dependency of the observations arising from the same unit and the (residual) random component ϵ represents a possible overdispersion (due to clustering) of the counts.

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

Model control for Poisson Model with Random Components

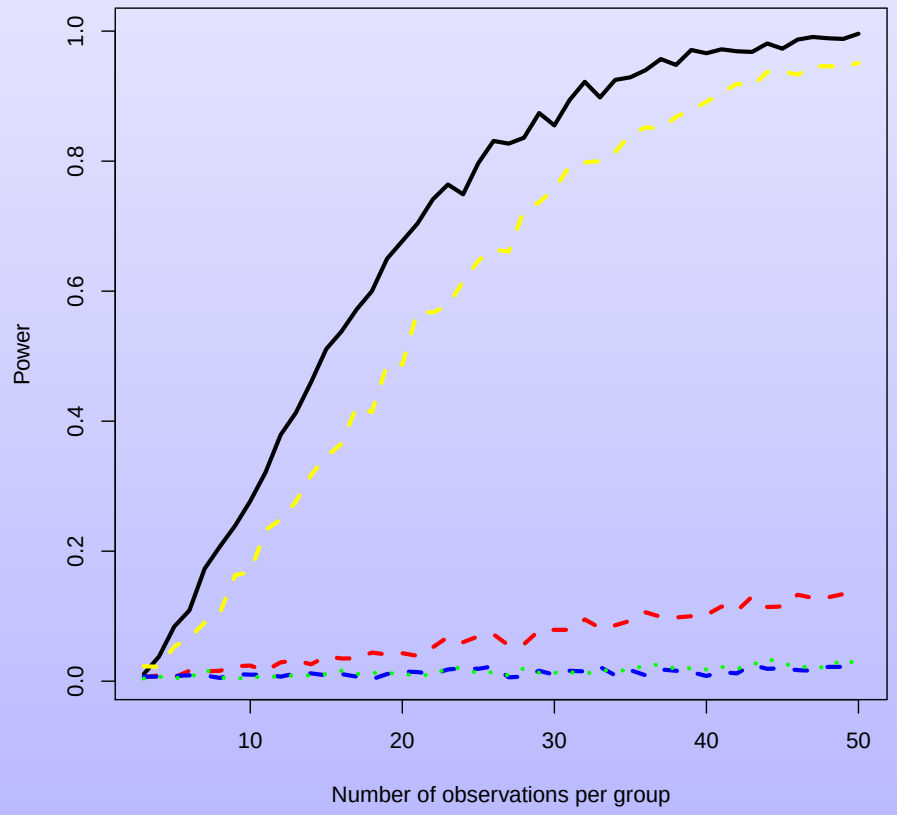


By the way, the model detects differences between the treatments!

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

All these models; does it matter?

Some power calculations: t-test with different distributions



 Copyright © 2023 by Rodrigo Labouriau, please do not circulate

It has been nice to be in contact with all of you!

 Copyright © 2023 by Rodrigo Labouriau, please do not circulate