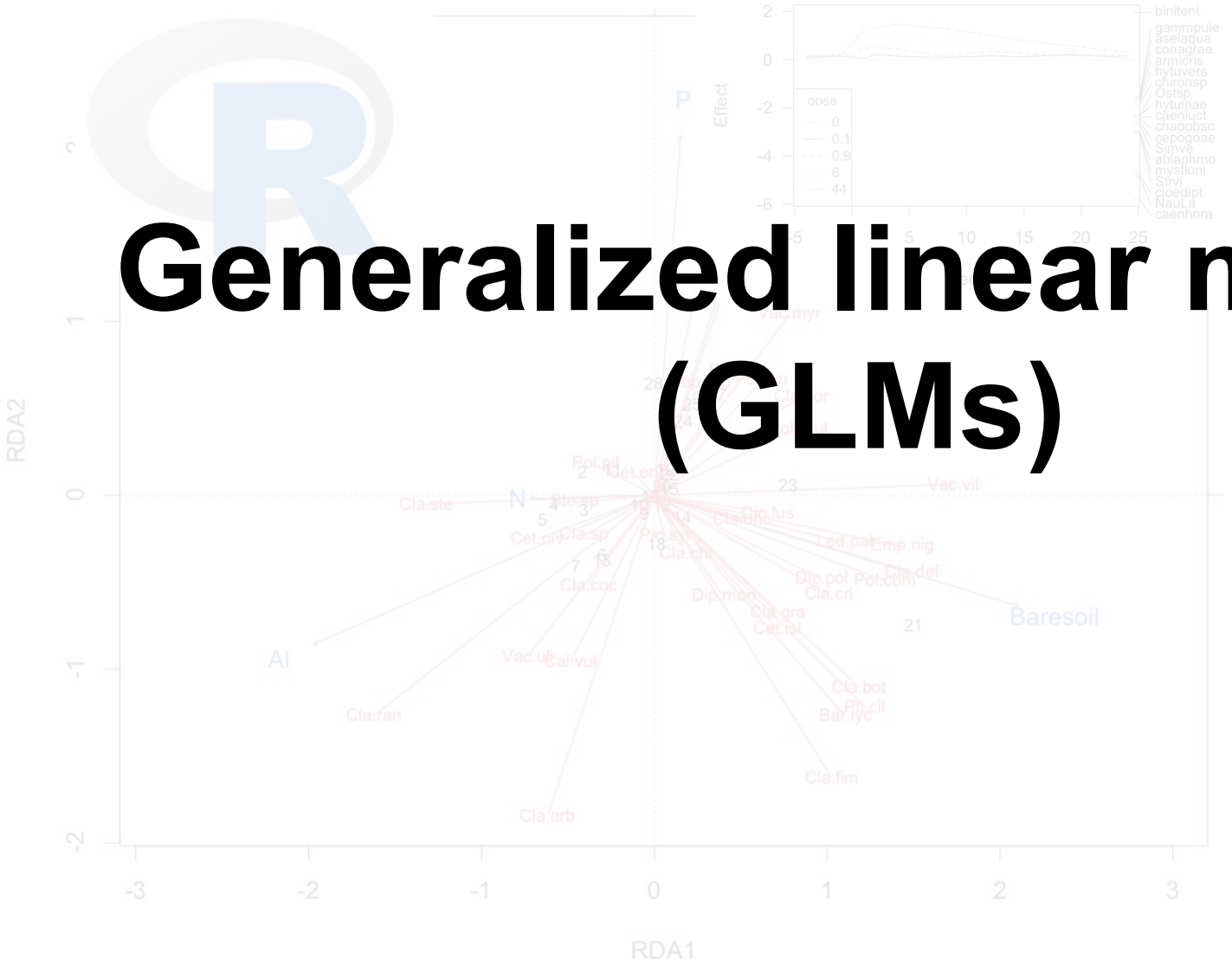
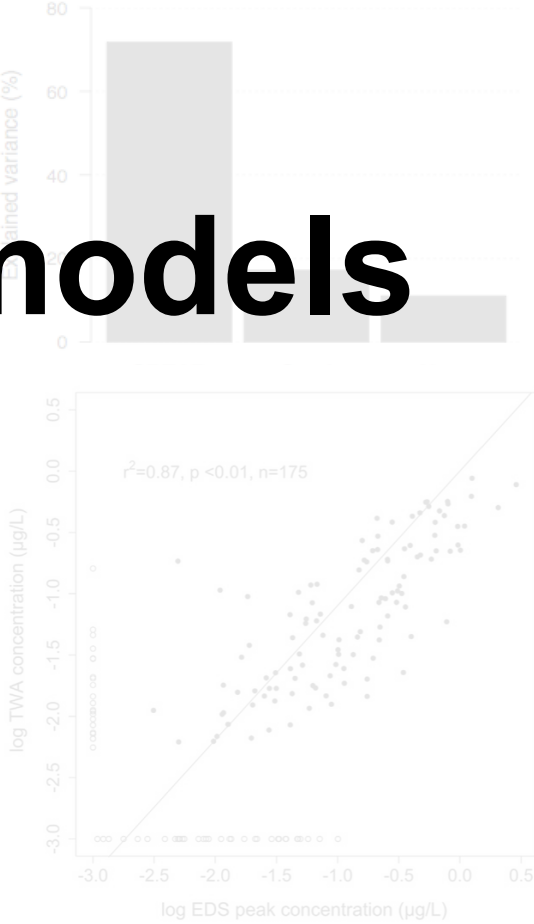
Tools for complex data analysis

University of Koblenz-Landau 2018/19





Generalized linear models (GLMs)

Ralf B. Schäfer

Learning targets

- Explaining and applying generalized linear models
- Explaining the concepts of maximum likelihood estimation and deviance
- Describe the specifics of GLMs regarding model selection and model assumptions

Learning targets and study questions

- Explaining and applying generalized linear models
  - Why should you use logistic regression for binomial data? Which assumptions of the linear model are violated and why?
  - How does the GLM deal with non-constant variance?
  - Outline differences in the model structure between a simple linear model and a GLM.
  - Describe typical error distribution and link functions for modelling a) species abundances and b) fraction of surviving organisms.
- Explaining the concepts of maximum likelihood estimation and deviance
  - Explain the core idea of maximum likelihood estimation and how it is done.
  - What is the difference between likelihood and probability?
  - Explain the concept of deviance for the GLM and why sum of squares can not be used.

Learning targets and study questions

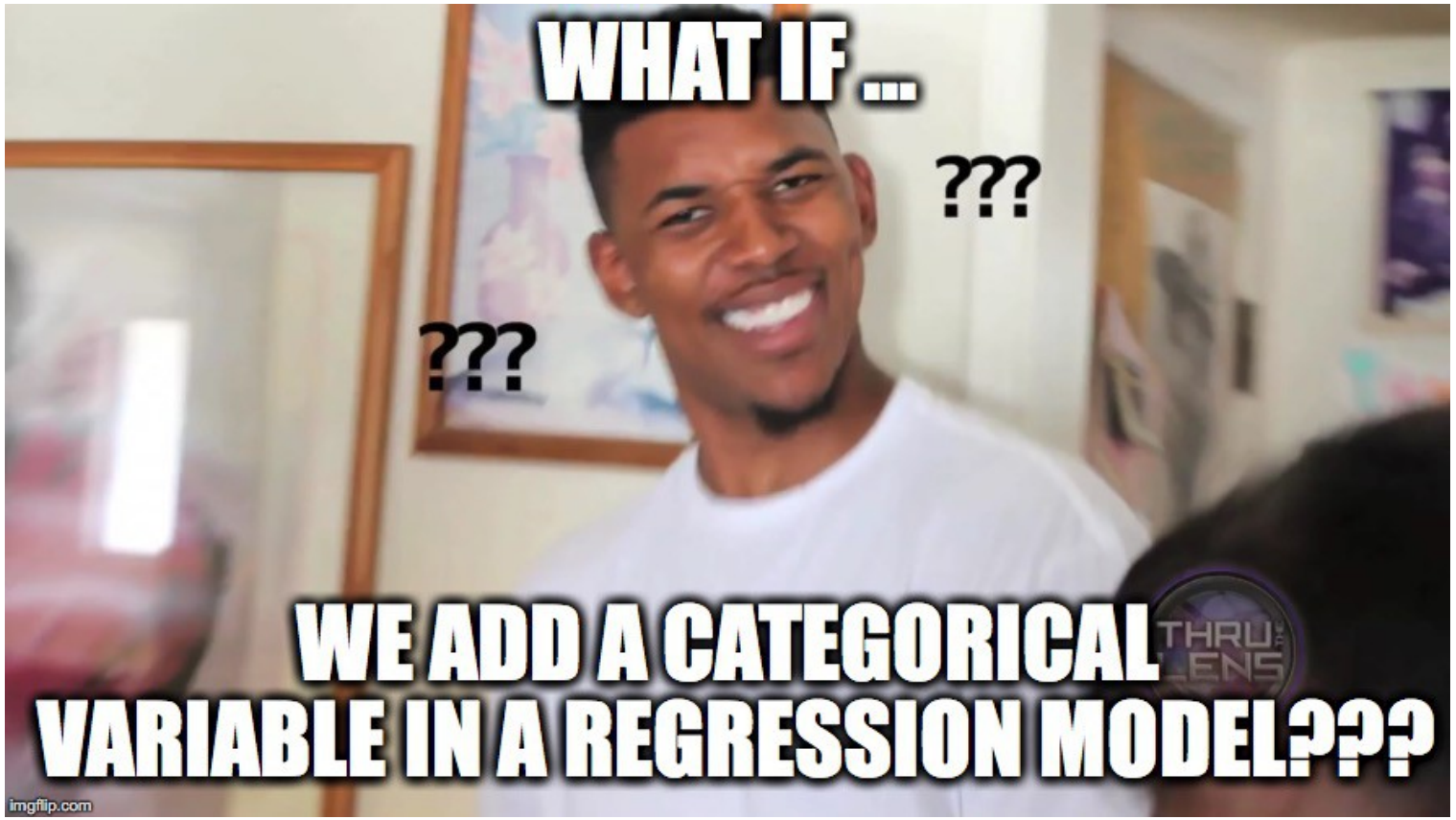
- Describe the specifics of GLMs regarding model selection and model assumptions
  - Describe the methods that can be used for model selection and specifics for GLMs.
  - Which types of model diagnostics are required for a GLM, and which of these are particular for this class of models?
  - Explain the issue of overdispersion and options to deal with it.

Generalized linear model

Contents

Extending the linear model: Motivation

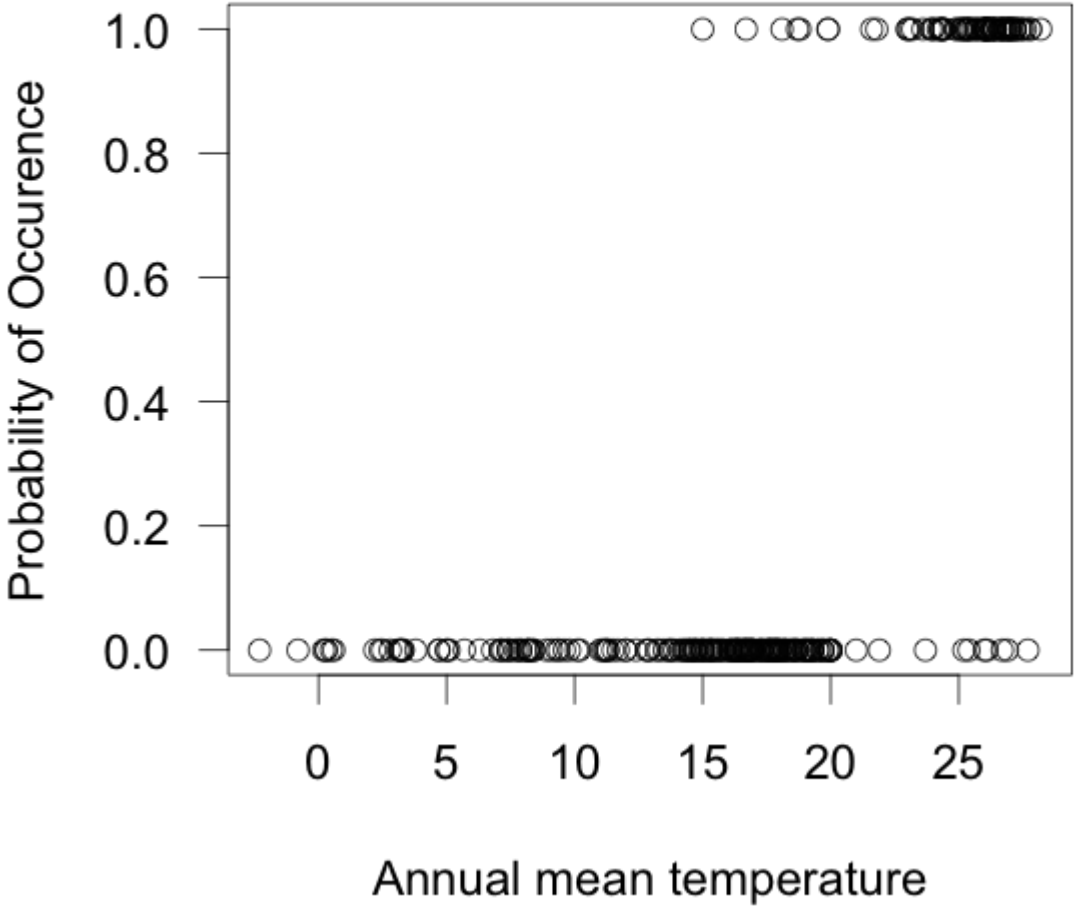
Linear model requires a continuous response, but responses can be discrete (e.g. number of events, objects or organisms) or categorical (e.g. occurrence (presence/absence) of events, objects or organisms)



Case study: Eastern brown snake

Research question: How does temperature influence the probability of occurrence of the eastern brown snake?

Study: Samples of potential habitats along temperature gradient

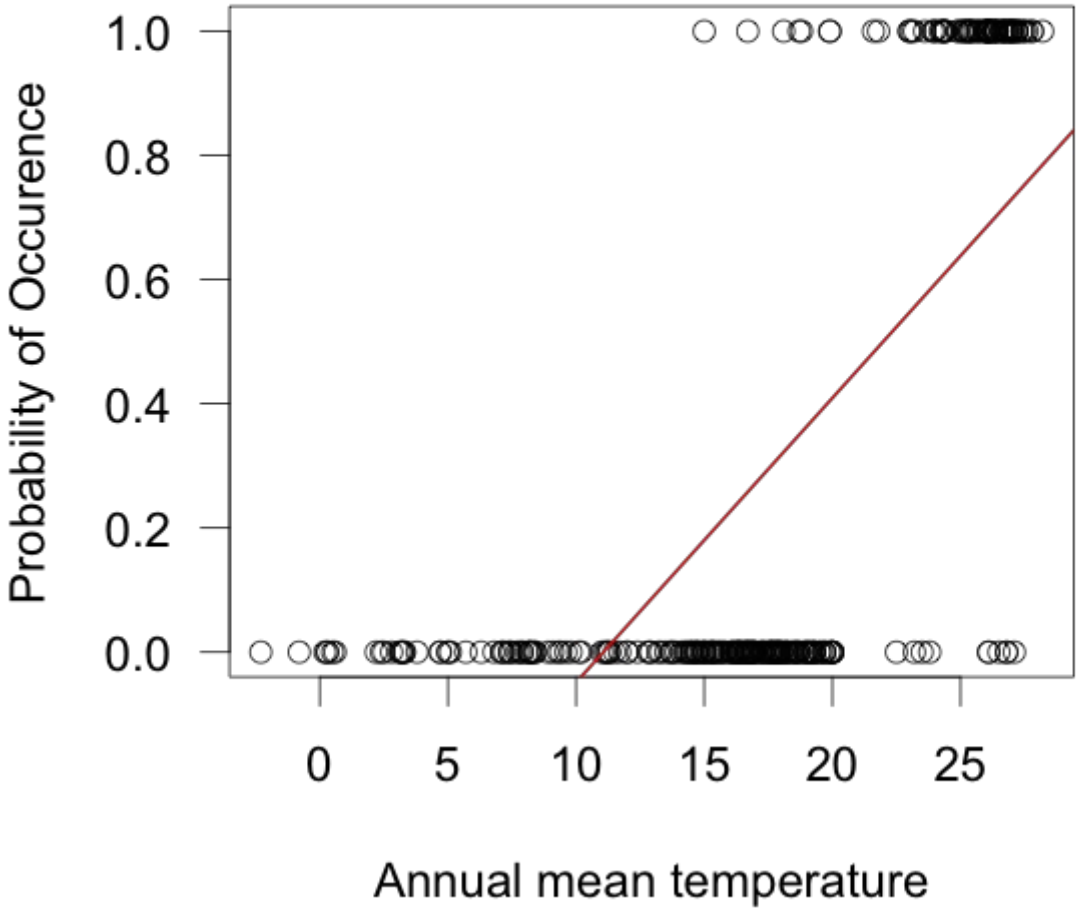


Case study: Eastern brown snake

Research question: How does temperature influence the probability of occurrence of the eastern brown snake?

Method: Linear regression model?

Provides meaningless probabilities <0 and >1!

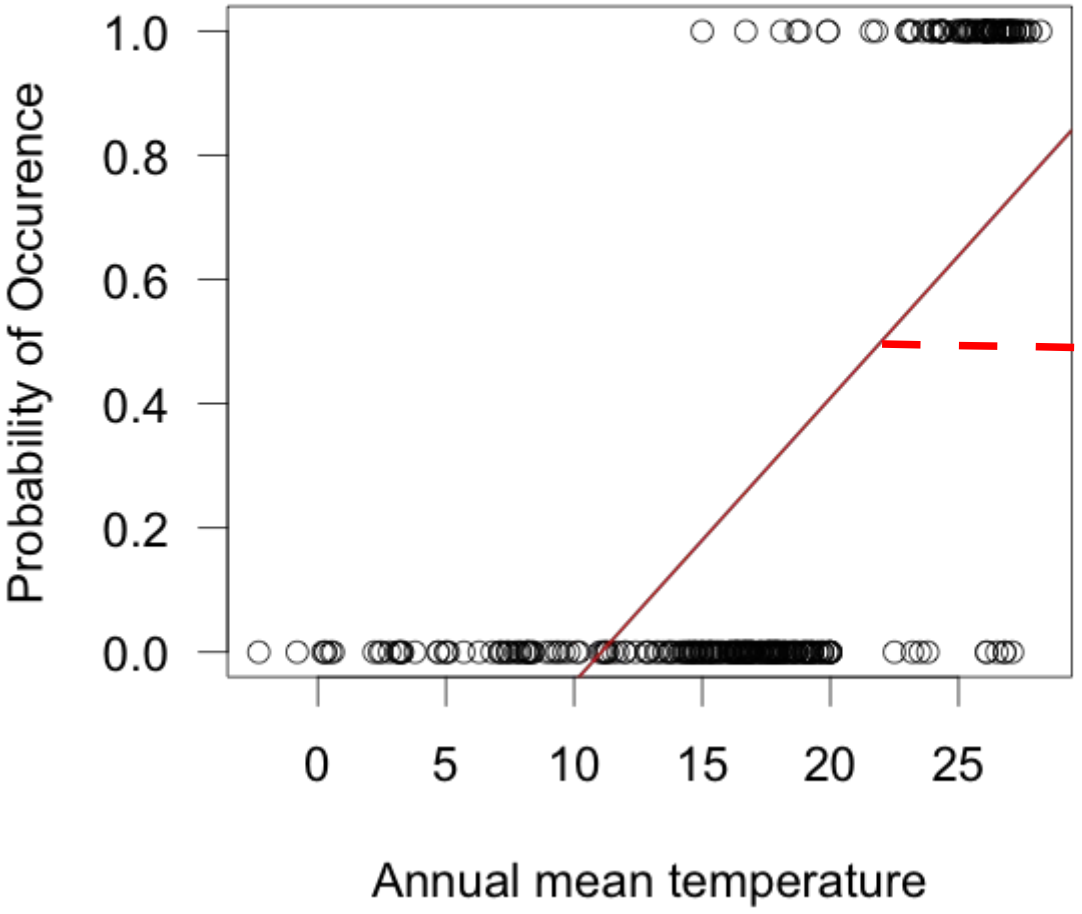


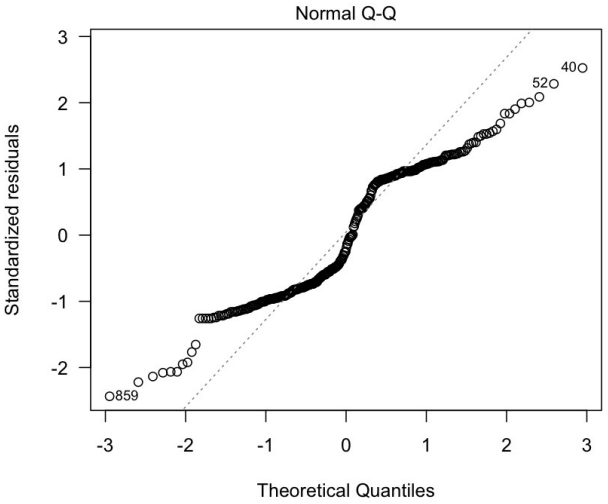
Case study: Eastern brown snake

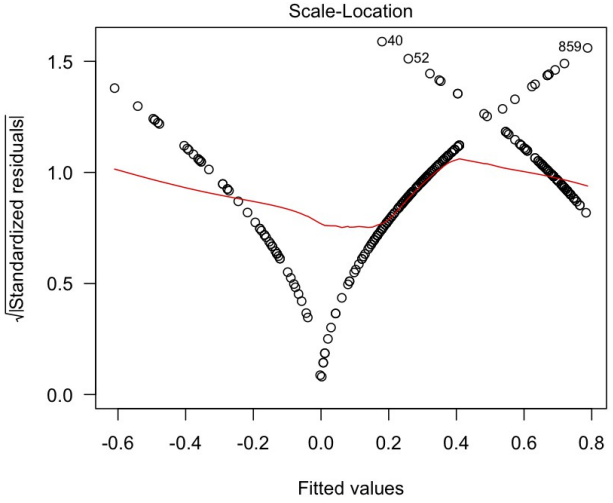
Research question: How does temperature influence the probability of occurrence of the eastern brown snake?

Method: Linear regression model?

Assumptions violated!







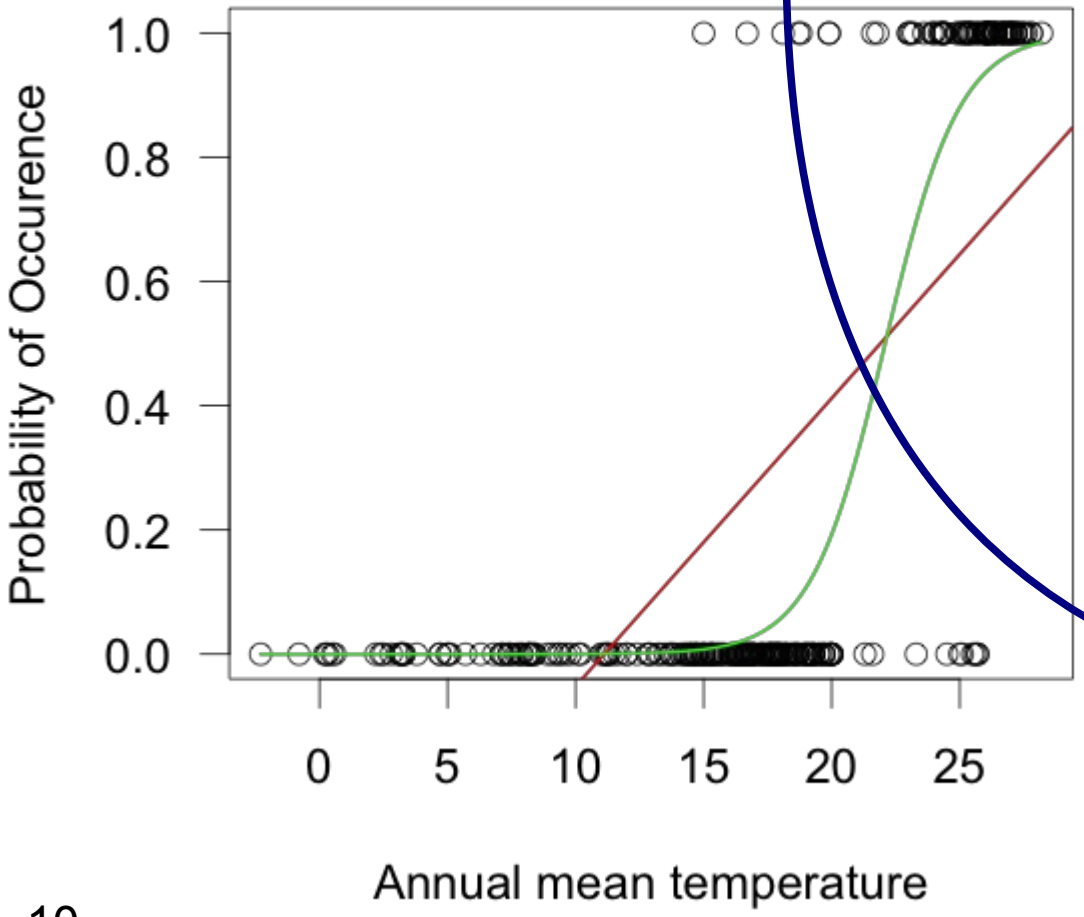
From simple to logistic regression

Research question: How does temperature influence the probability of occurrence of the eastern brown snake?

Method: Simple linear regression *vs.* logistic regression

$$E(Y|X=x) = \mu(X) = \beta_0 + \beta_1 X$$

$$E(Y=1|X=x) = \mu(X) = \pi$$



$$\pi = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \Leftrightarrow \pi = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

Reformulation

Logit

$$\log_e\left(\frac{\pi}{1-\pi}\right) = \beta_0+\beta_1 X$$

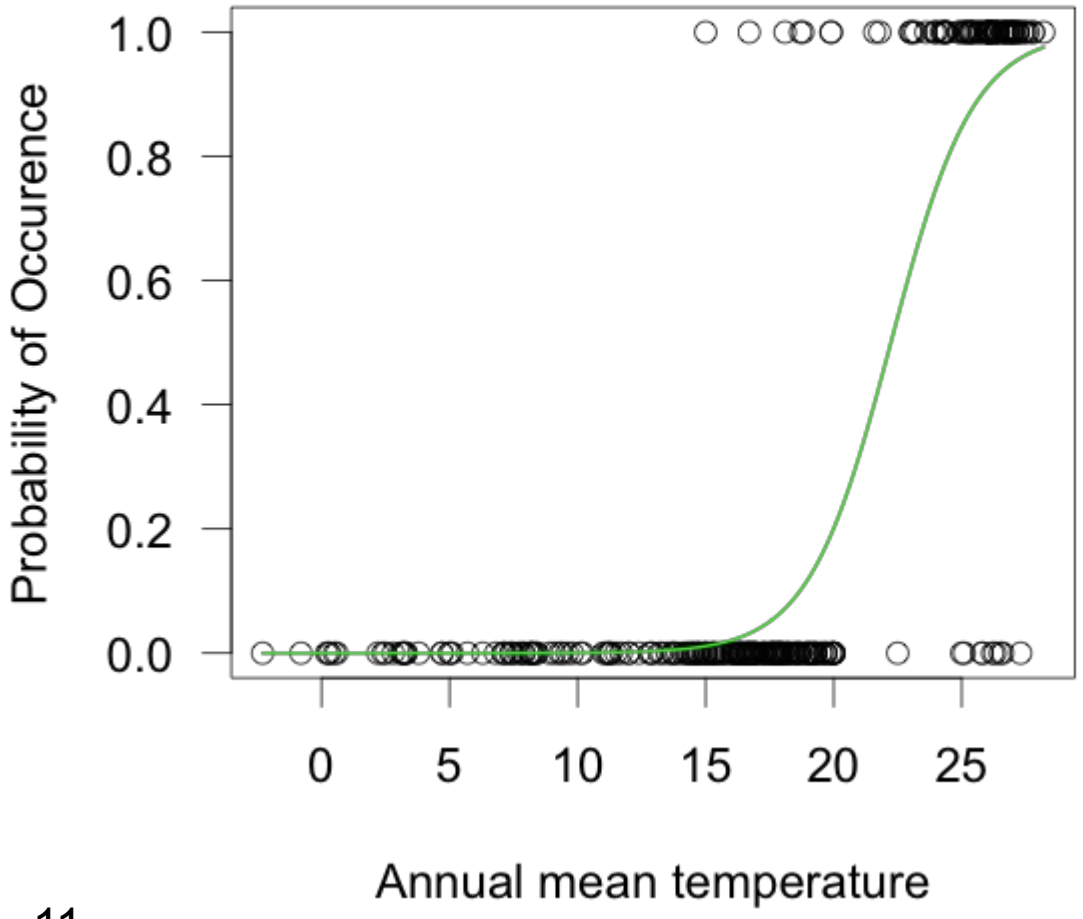
Linear component

Parameters in logistic regression

Formula for fitted model

$$\log_e\left(\frac{\hat{\mu}(x_i)}{1-\hat{\mu}(x_i)}\right) = \log_e\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = b_0+b_1 x_i$$

Given are parameters of the logistic model: b_0 = -15.9, b_1 = 0.72



Which x relates to a $\hat{\pi}$ = 0.5?

$$\log_e\left(\frac{0.5}{1-0.5}\right) = -15.9+0.72x$$
$$\Leftrightarrow \log_e(1) = -15.9+0.72x$$
$$\Leftrightarrow 0+15.9 = 0.72x$$
$$\Leftrightarrow \frac{15.9}{0.72} = x \Rightarrow x = 22.1$$

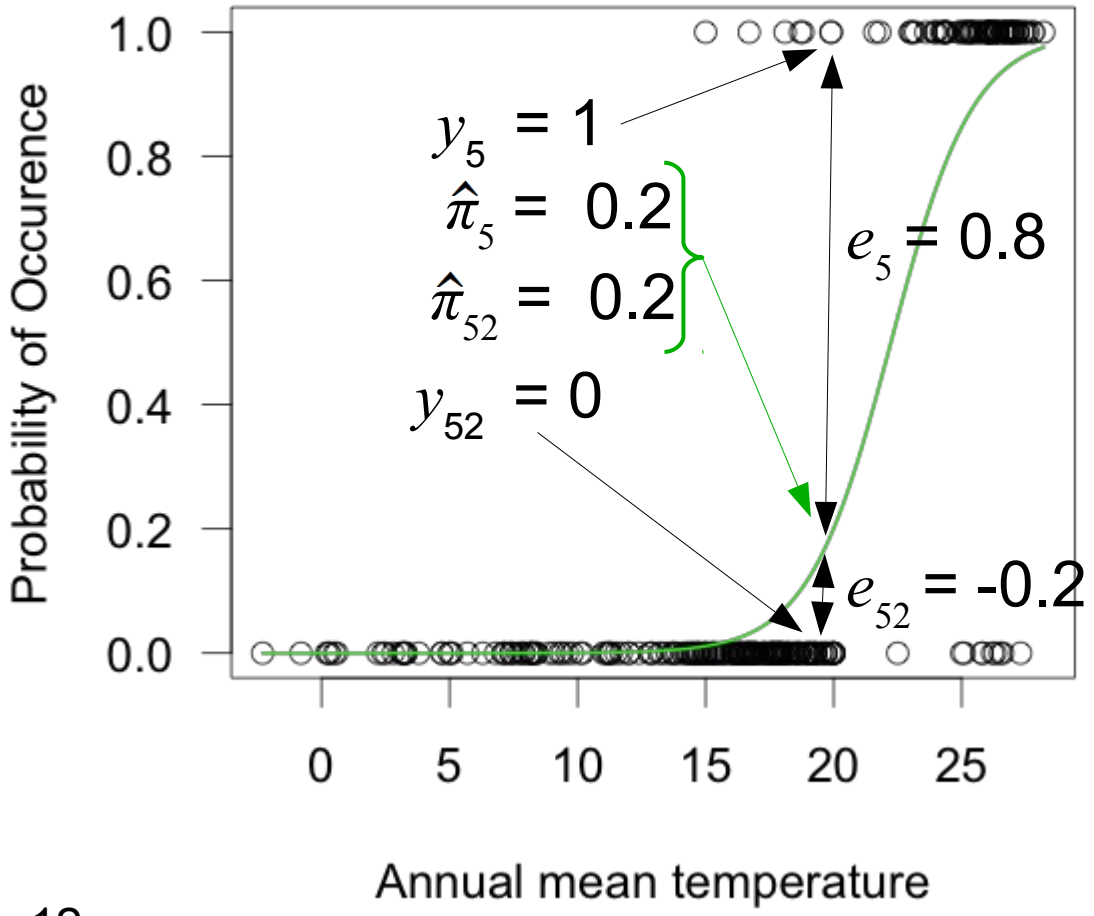
Repeat the calculation for $\hat{\pi}$ = 0.1 and $\hat{\pi}$ = 0.9

Variance and logistic regression

$$\log_e\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \text{logit}(\hat{\pi}_i) = b_0 + b_1 x_i \Leftrightarrow \hat{\pi}_i = \text{logit}^{-1}(b_0 + b_1 x_i)$$

We define residuals as: $e_i = y_i - \text{logit}^{-1}(b_0 + b_1 x_i) = y_i - \hat{\pi}_i$

➡ Residuals are dichotomous: For $\hat{\pi}_i$ = 0.2, e_i is either 0.8 or -0.2.



Generally, the (true) variance of Y is given by:

$$\text{Var}(Y) = \pi(1-\pi) = \mu(1-\mu)$$

This means:

- Variance not constant, depends on π (i.e. μ) → Quadratic function
- Variance not normally distributed

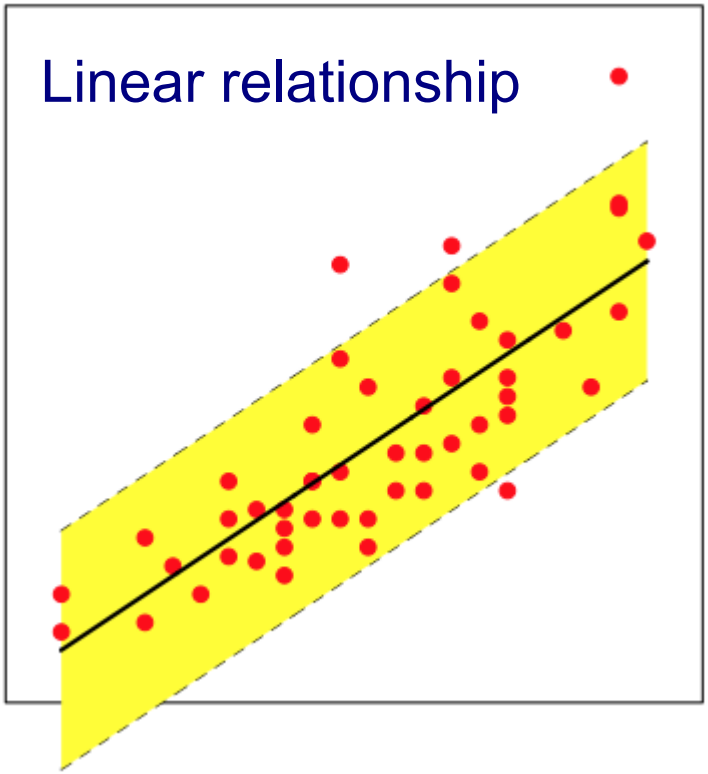
Generalized linear model

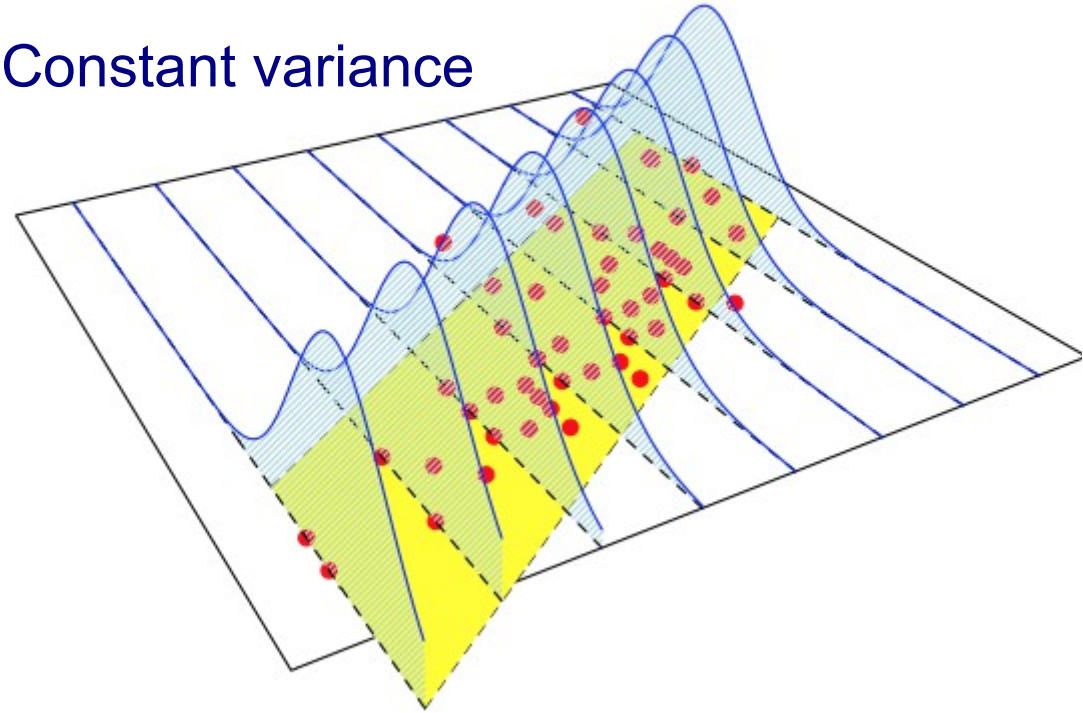
Contents

1. Case study: Logistic regression
2. **Extending the linear model**
3. Definition of the GLM
4. Deviance and Likelihood
5. Model selection and diagnostics

Remember: Simple linear regression

Linear model assumes linear relationship between explanatory variable(s) and response variable as well as a constant variance



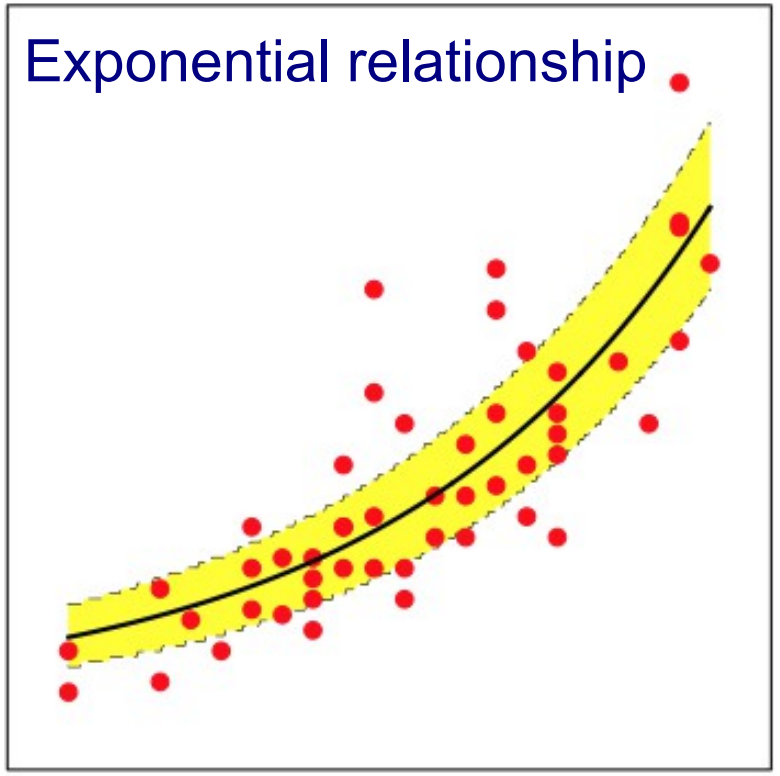


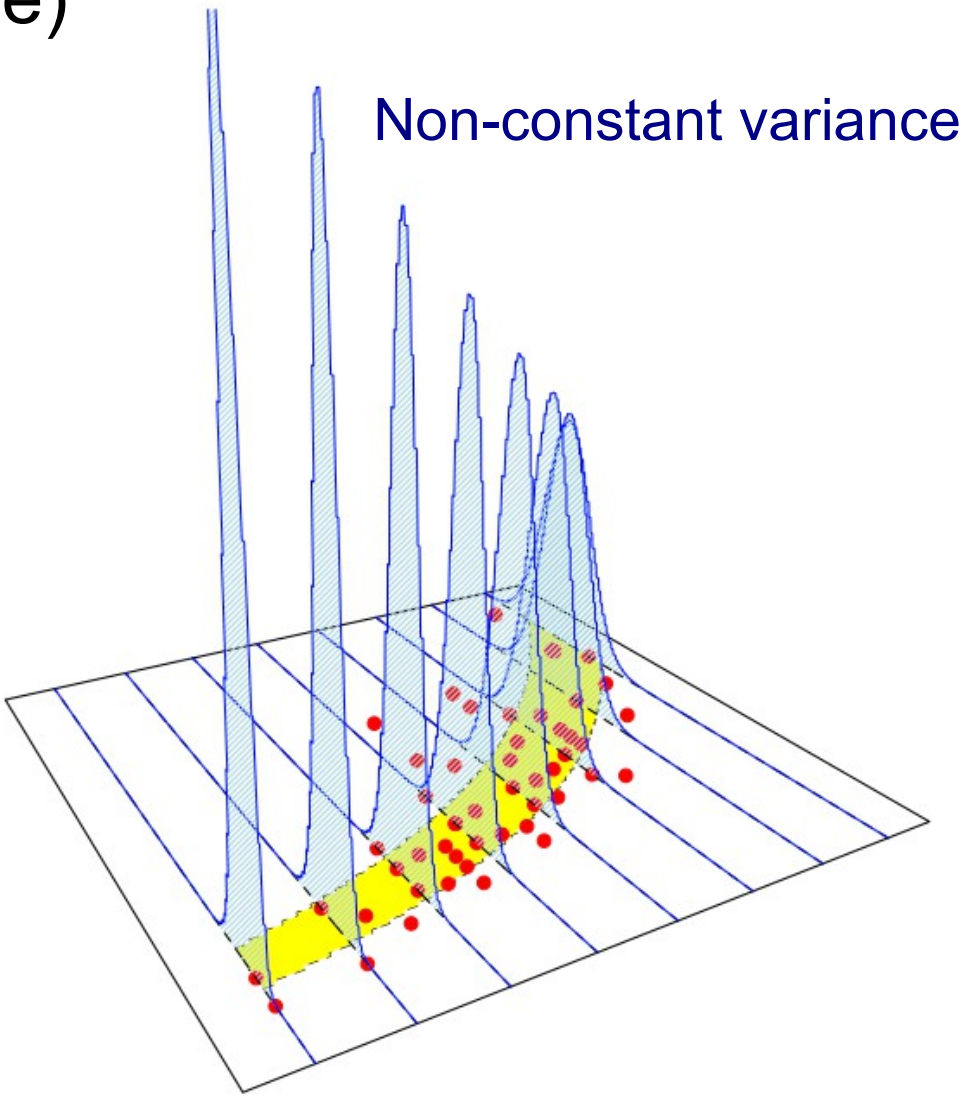
For ecological data, relationship with response variable can be non-linear and variance is often not constant

http://freakonometrics.hypotheses.org/tag/poisson

Extending the linear model: Motivation

Example for non-linear relationship and non-constant variance with continuous response (in contrast to logistic regression with binary response)

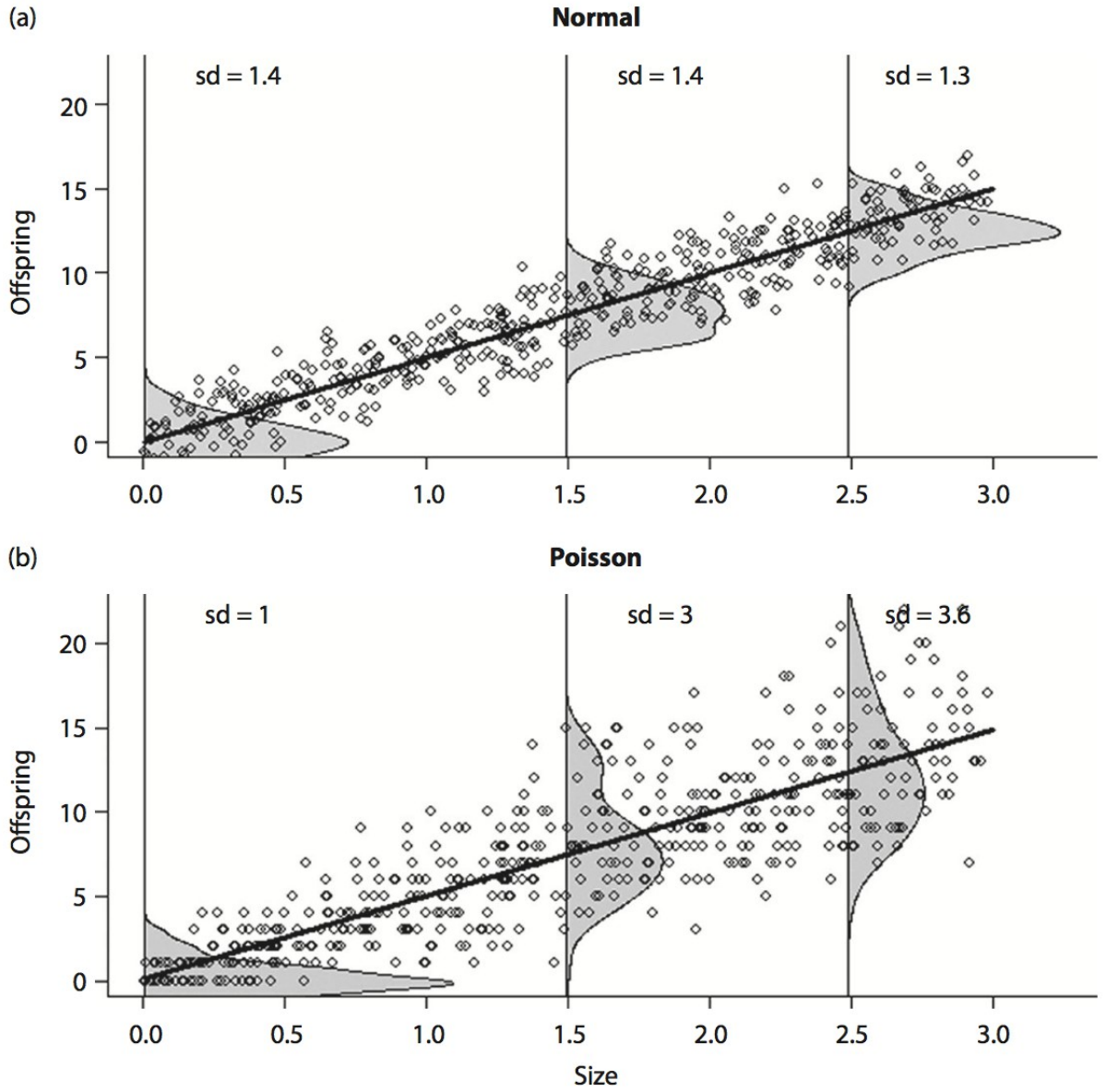




http://freakonometrics.hypotheses.org/tag/poisson

Extending the linear model: Motivation

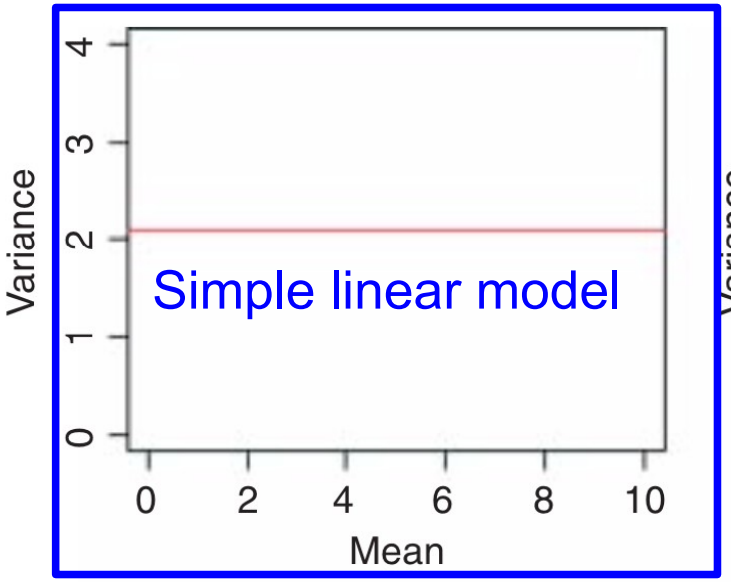
Example: Increasing variability in number of offsprings with increasing body size of individuals

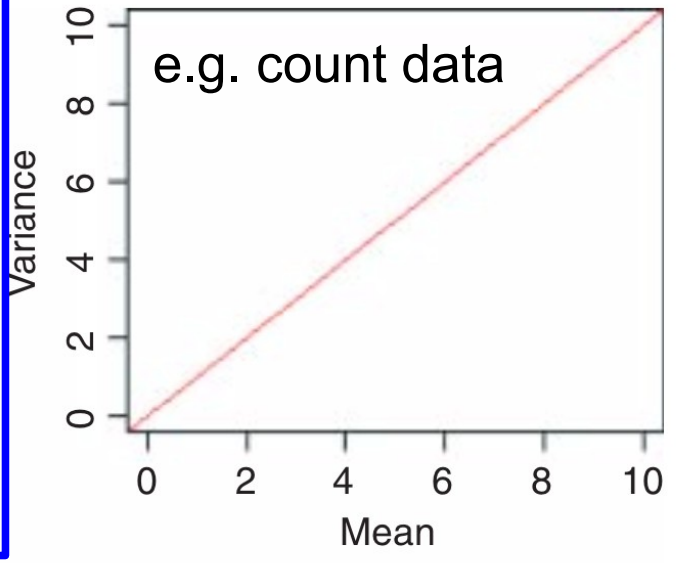


Buckley 2015: 134

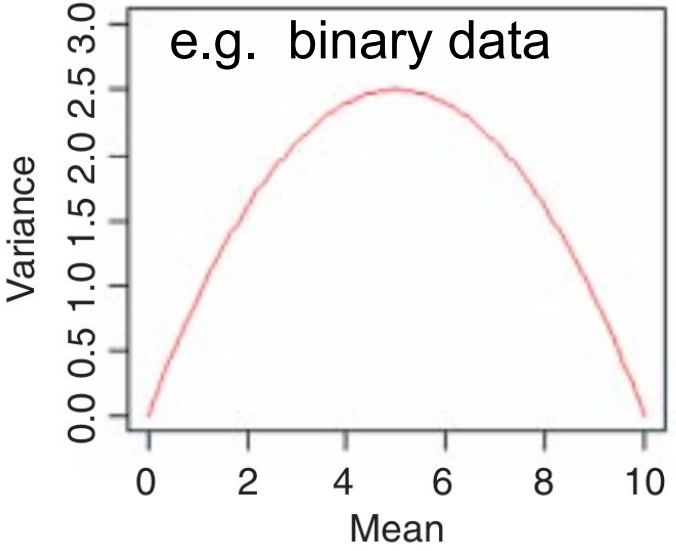
Modelling the mean-variance relationship

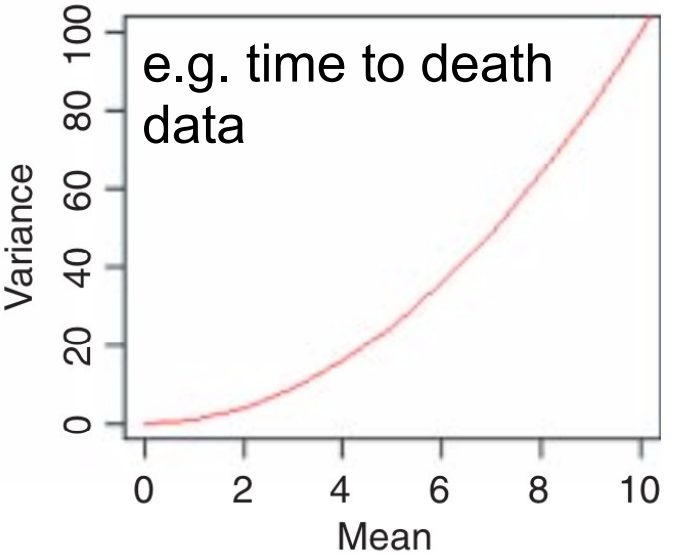
Idea: Express variance as a function of the mean!





taken from
Crawley 2012: 557





Generalized linear model

Contents

1. Case study: Logistic regression
2. Extending the linear model
3. **Definition of the GLM**
4. Deviance and Likelihood
5. Model selection and diagnostics

Defining the GLM

Linear model: $Y = \beta_0 + \beta_1 X + \varepsilon$

Generalised linear model:

1. Linear predictor: $\eta = \beta_0 + \beta_1 X$
2. Link function: $g(\mu) = \eta$ with $E(Y|X = x) = \mu$
3. Distribution of Y with related $\text{Var}(Y) = \phi\, V(\mu)$

Error structure with related variance function and typical link function

Family (error structure)	Default Link	Link name	Variance function
`gaussian`	$\eta = \mu$	identity	1
`poisson`	$\eta = \log_e \mu$	log	μ
`binomial`	$\eta = \log_e\left(\frac{\mu}{(n-\mu)}\right)$	logit	$\frac{\mu(n-\mu)}{n}$
`Gamma`	$\eta = \mu^{-1}$	inverse	μ^2
`inverse.gaussian`	$\eta = \mu^{-2}$	inverse square	μ^3

modified from Crawley 2012: 562

General and specific GLMs

Response Y follows distribution from exponential family:

$$f_\theta(y) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$$

Specific (exponential) distributions:

Gaussian

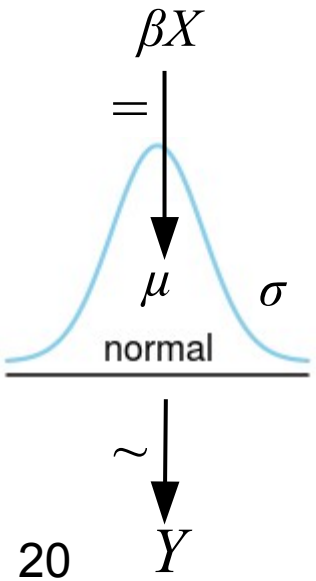
$Y \sim \text{Normal}(\mu, \sigma)$

$E(Y) = \mu$

$\text{Var}(Y) = \sigma^2$

$\mu = \beta X$

$\varepsilon = y - \mu$



Binomial

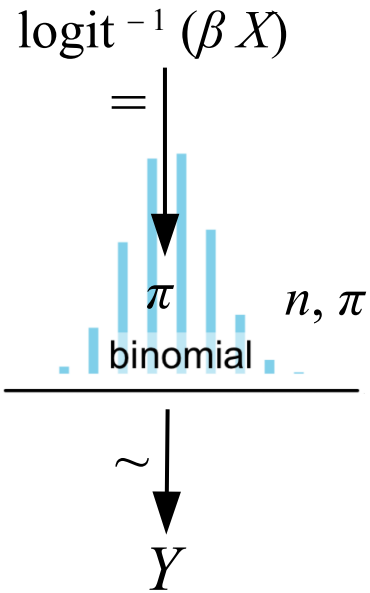
$Y \sim \text{Bin}(n, \pi)$

$E(Y) = \pi$

$\text{Var}(Y) = \frac{\pi(n-\pi)}{n}$

$\text{logit}(\pi) = \beta X$

$\varepsilon = y - \pi$



Poisson

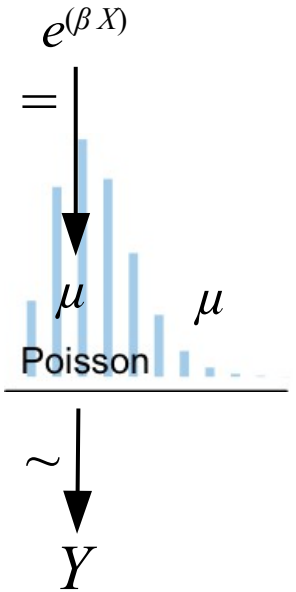
$Y \sim \text{Pois}(\mu)$

$E(Y) = \mu$

$\text{Var}(Y) = \mu$

$\log(\mu) = \beta X$

$\varepsilon = y - \mu$



Negative binomial

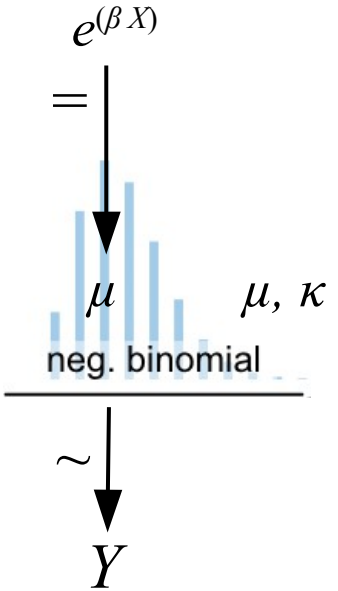
$Y \sim \text{Neg.Bin}(\mu, \kappa)$

$E(Y) = \mu$

$\text{Var}(Y) = \mu + \frac{\mu^2}{\kappa}$

$\log(\mu) = \beta X$

$\varepsilon = y - \mu$



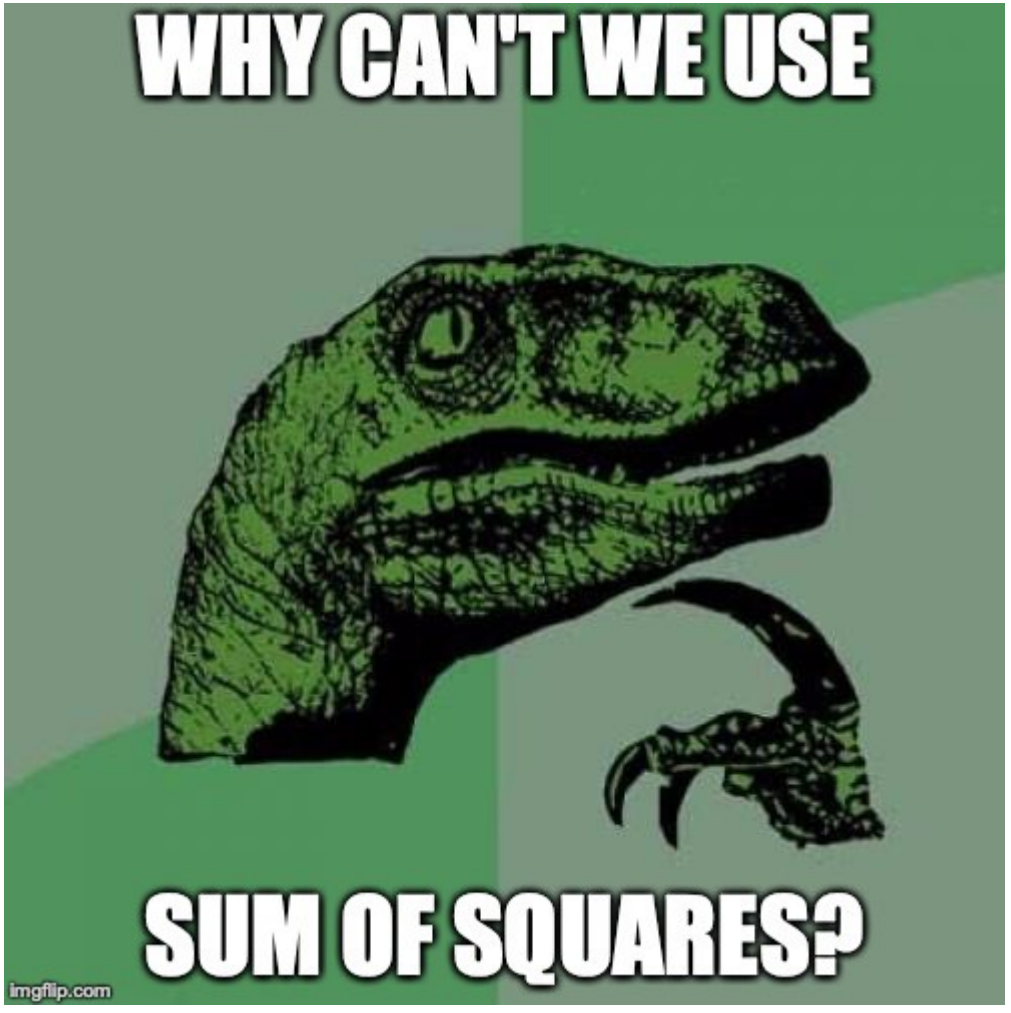
Data type and GLM specification

Response variable	Error distribution	Canonical link function	Alternative link functions
Continuous positive and negative values	Gaussian/Normal	Identity	Log, Inverse
Counts	Poisson	Log	Identity, Sqrt
Counts with over-dispersion	Negative Binomial, Quasi-Poisson	Log Log	As per Poisson
Proportions (no. successes/total trials)	Binomial	Logit	Probit, Cauchit, Log, Complementary Log-Log
Binary (male/female, alive/dead)	Binomial (Bernoulli)	Logit	As per Binomial
Proportions or binary with overdispersion	Quasi-Binomial	logit	As per Binomial
Time to event (germination, death)	Gamma	Inverse	Inverse, Identity, Log

Buckley 2015: 137

Deviance: Goodness of fit for GLM

- GLMs minimize Deviance (D) instead of Sum of Squares in simple linear regression model
- Deviance derived by *Maximum Likelihood Estimation* (MLE)



Why can't we use sum of squares?

Example for logistic regression

Remember: For (simple) linear model, we determine coefficients by minimizing residual sum of squares (RSS):

$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - \left(b_0 + b_1 x_i \right) \right)^2 \;\Leftrightarrow\; \arg\min_{b_0, b_1} \sum_{i=1}^{n} e_i \;\Leftrightarrow\; \arg\min_{b_0, b_1} \sum \mathrm{RSS}$$

Application of concept to (simple) logistic regression:

$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} e_i \;\Leftrightarrow\; \arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - \frac{1}{1 + e^{-\left(b_0 + b_1 x_i \right)}} \right)^2$$

→ No simple algebraic matrix solution as for linear model
→ Solution based on maximum likelihood estimation

Generalized linear model

Contents

1. Case study: Logistic regression
2. Extending the linear model
3. Definition of the GLM
4. **Deviance and Likelihood**
5. Model selection and diagnostics

Maximum Likelihood Estimation (MLE)

Method of estimating model parameters that maximise the likelihood function given the data. In other words, model parameters are estimated that have the highest likelihood to produce the sample data.

General likelihood function: $L(\theta|y) = f_\theta(y)$

Search for maximum: $\arg\max_\theta L(\theta|y) = \arg\max_\theta f_\theta(y)$

The ML estimate $\hat{\theta}$ is defined as: $\hat{\theta} \in \{\arg\max_\theta L(\theta|y)\}$

<u>How to identify the maximum likelihood estimate?</u>

→ Set first derivative equal to zero: $\frac{\partial L(\theta|y)}{\partial \theta} = 0$

Example: Likelihood calculation

Research question: What is the probability p of an insect to be killed at a specific chemical concentration?

Study: Laboratory test with 15 insects. 10 died.

$$L(p|y,n) = \binom{15}{10} p^{10}(1-p)^5$$

p = ?; probab. of death in single trial (i.e. insect)
n = 15; no. of trials (i.e. insects)
y = 10; no. of deaths

→ Estimate parameter p with MLE

$$\hat{p} \in \{\arg\max_p L(p|x)\}$$

Log likelihood simplifies derivation

$$\log L(p|y,n) = \log\left(\binom{15}{10} p^{10}(1-p)^5\right) = \log 3003 + 10 \log p + 5 \log(1-p)$$

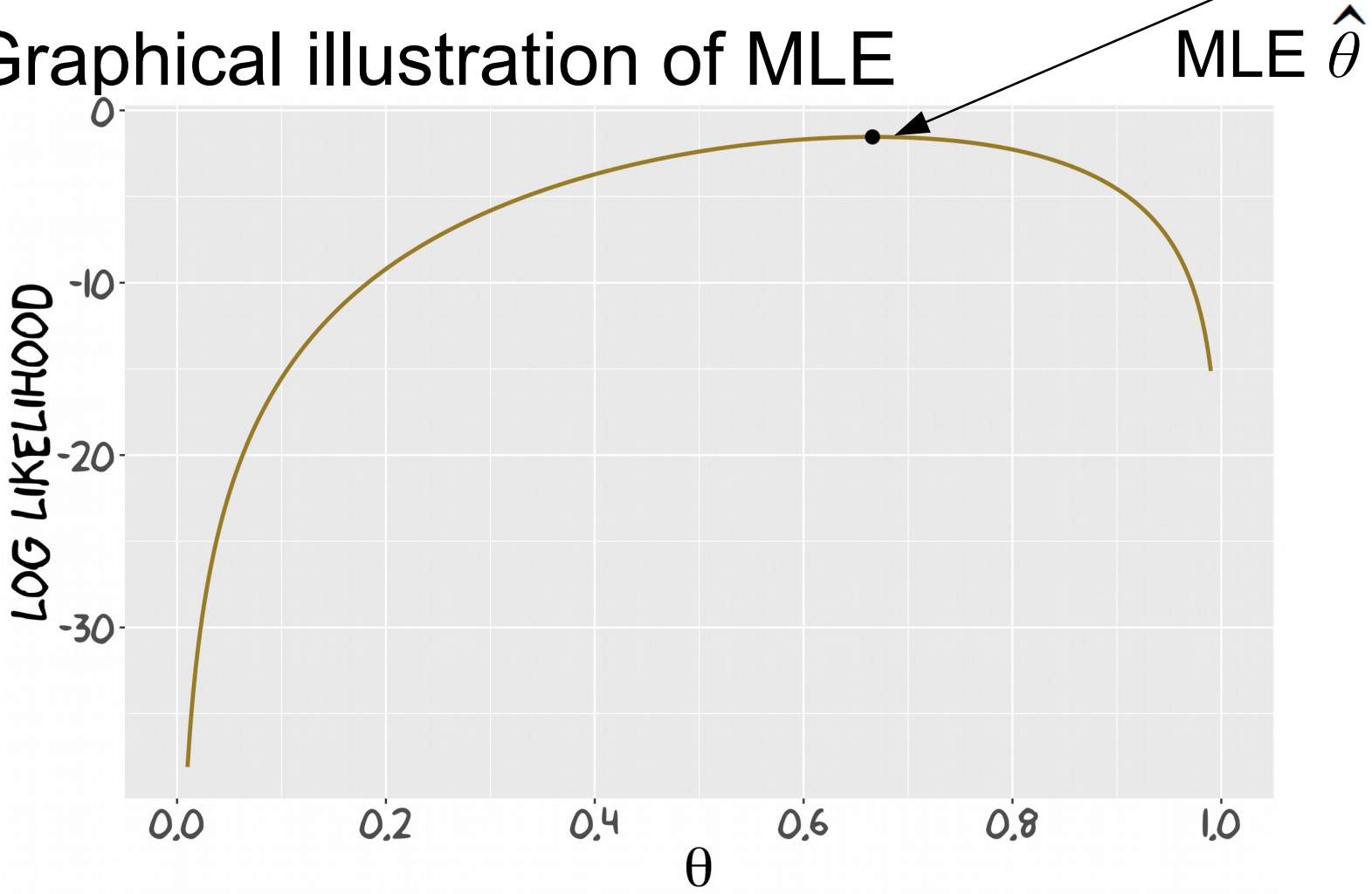
$$\frac{\partial L(\theta|x)}{\partial \theta} = 0 \Rightarrow \frac{\partial \log L(\hat{p}|y,n)}{\partial p} = 0$$

Example: Likelihood calculation

$$\frac{\partial \log L(\hat{p}|y,n)}{\partial p} = 0 \Leftrightarrow \frac{\partial(8+10\log\hat{p}+5\log(1-\hat{p}))}{\partial p} = 0 \Leftrightarrow$$

$$\frac{10}{\hat{p}}+\frac{5}{(\hat{p}-1)} = 0 \Leftrightarrow \frac{5}{(\hat{p}-1)} = -\frac{10}{\hat{p}} \Leftrightarrow 5\hat{p} = -10\hat{p}+10 \Leftrightarrow \hat{p} = \frac{2}{3}$$

Graphical illustration of MLE



MLE and logistic regression

Probability P for simple logistic regression

$$P(y_i=0|x_i, b_0, b_1) = 1-\frac{1}{1+e^{-(b_0+b_1 x_i)}}$$

$$P(y_i=1|x_i, b_0, b_1) = \frac{1}{1+e^{-(b_0+b_1 x_i)}}$$

Likelihood L of any observation (x_i, y_i) in simple logistic regression:

$$L(b_0, b_1|y_i, x_i) = \left(\frac{1}{1+e^{-(b_0+b_1 x_i)}}\right)^{y_i}\left(1-\frac{1}{1+e^{-(b_0+b_1 x_i)}}\right)^{1-y_i}$$

Likelihood L of all n observations (x, y) is:

$$L(b_0, b_1|y, x) = \prod_{i=1}^{n}\left(\frac{1}{1+e^{-(b_0+b_1 x_i)}}\right)^{y_i}\left(1-\frac{1}{1+e^{-(b_0+b_1 x_i)}}\right)^{1-y_i}$$

MLE and GLMs

log Likelihood L of all n observations (x, y) is:

$$\log L(b_0, b_1|y, x) = \sum_{i=1}^{n} \left(-\log\left(1+e^{-(b_0+b_1 x_i)}\right)+ y_i\left(b_0+b_1 x_i\right)\right)$$

General: log Likelihood for GLMs

$$\log L(\theta, \phi|y) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$$

Solution to MLE based on complex matrix algebra and different algorithms, for details see: Dobson & Barnett (2018: 71-76), Wood (2017: 105-107) and Hilbe (2017: 51-62)

Deviance: Goodness of fit for GLM

Definition of residual deviance D:

$$D(y|\hat{\mu}) = -2(\log L_m - \log L_s) = -2[\log L(\hat{\mu}, \phi|y) - \log L(y, \phi|y)]$$

L_m is the maximised likelihood of our current model that is nested in a saturated model with maximised likelihood L_s (i.e. a model fitting the data as closely as possible)

Deviance can generally be used for model comparison

log likelihood and Information criteria

The information-theoretic criteria AIC and BIC also rely on the log likelihood $\log L$:

$$\text{AIC} = -2\log L + 2p$$

$$\text{BIC} = -2\log L + \ln(n)p$$

n = sample size

p = parameters in model

Generalized linear model

Contents

1. Case study: Logistic regression
2. Extending the linear model
3. Definition of the GLM
4. Deviance and Likelihood
5. **Model selection and diagnostics**

Model selection for GLM

- Same methods as for multiple linear regression model
- Best subset and multi-model averaging
- Hypothesis-based stepwise model selection:
  - Wald test for individual regression coefficients
  - Log-likelihood ratio test for complete model comparison
- Information-theoretic stepwise model selection (e.g. AIC, corrected AIC, BIC)
- Post-selection shrinkage and LASSO

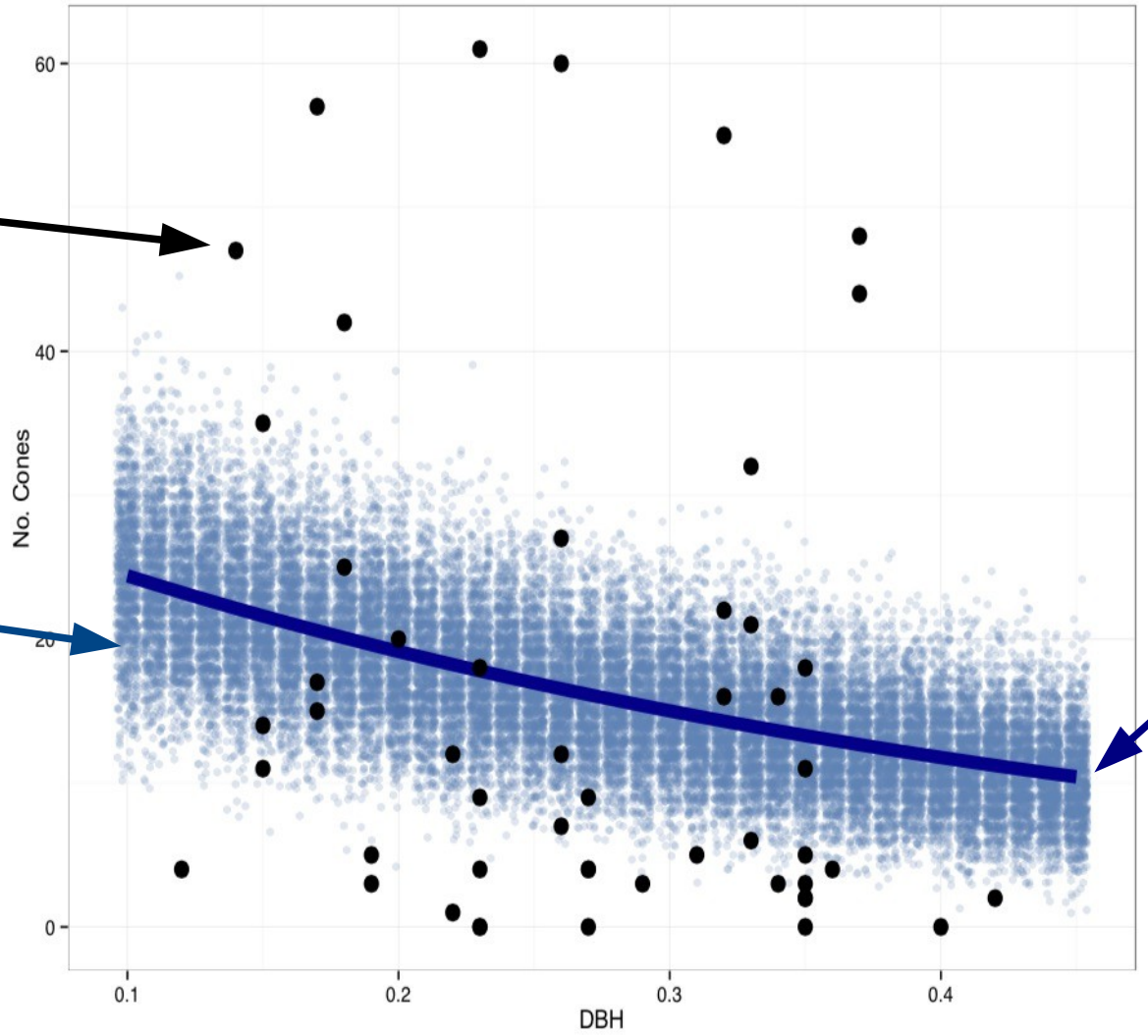
GLM assumptions and diagnostics

- Assumptions & tools from linear model largely apply to GLM
- Independence of observations
  - For spatiotemporal autocorrelation: GLMMs (see Bolker 2009)
- Assumed mean-variance relationship matches data (no over- or underdispersion in poisson and binomial GLM) (→ check with dispersion parameter and graphical diagnostics)

Overdispersion

Observations

Simulated data from Poisson distribution

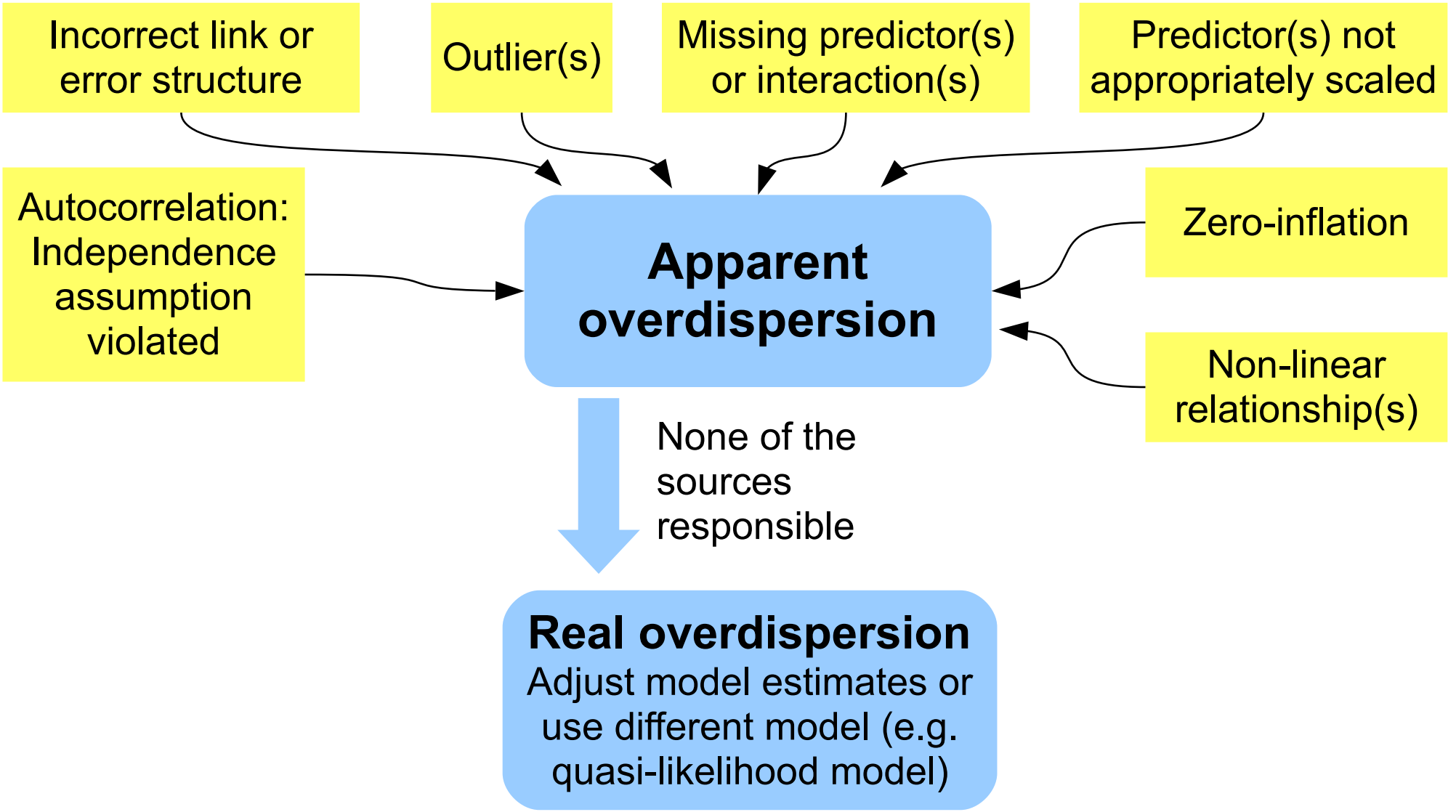


Poisson GLM

→ Too narrow standard errors (and in turn *p*-values or CIs) and too high estimates for regression coefficients

How to deal with overdispersion?

First step: Check for potential source of overdispersion



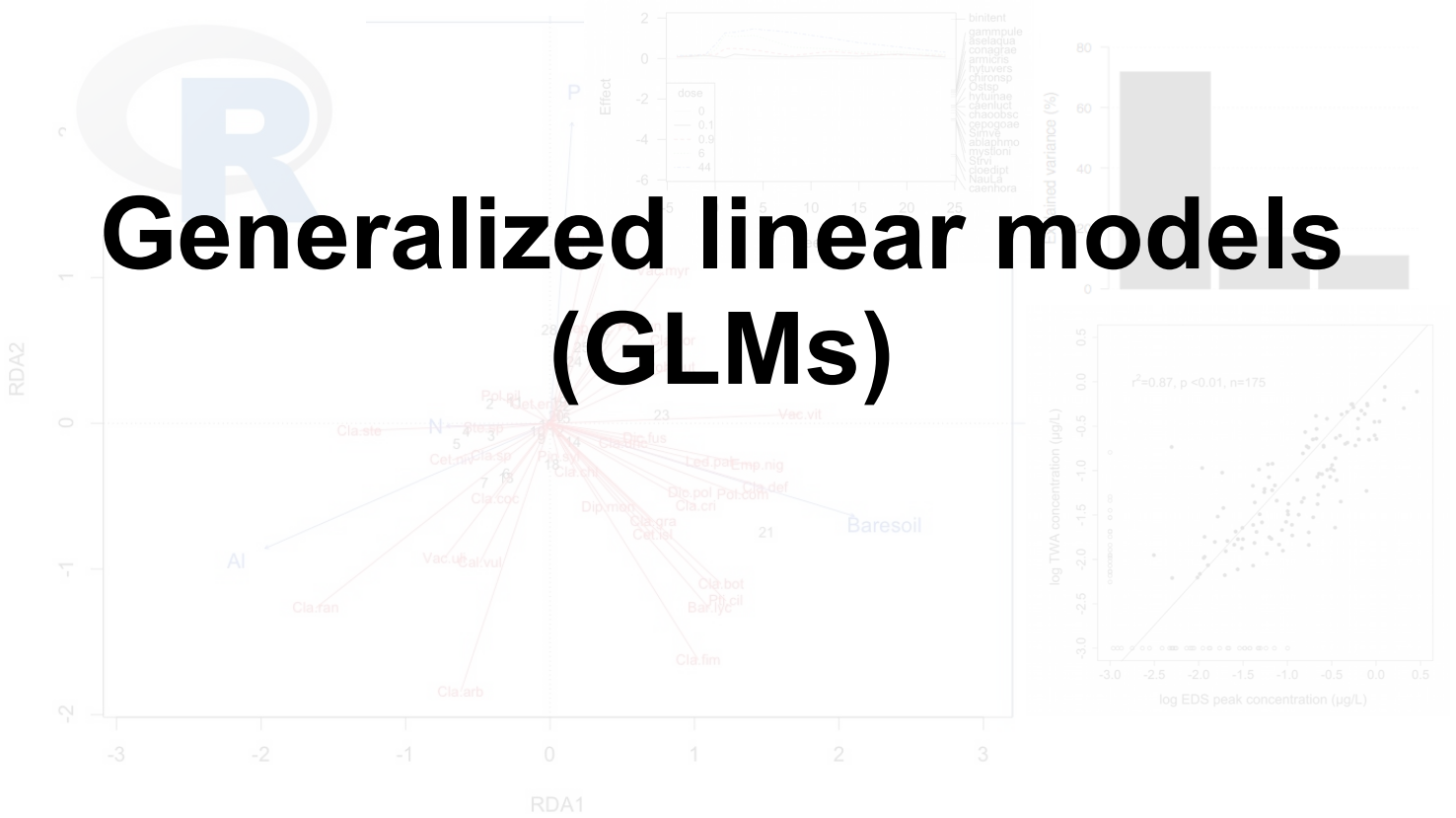
modified from Zuur et al. 2013: 20

GLM assumptions and diagnostics

- Assumptions & tools from linear model largely apply to GLM
- Independence of observations
  - For spatiotemporal autocorrelation: GLMMs (see Bolker 2009)
- Assumed mean-variance relationship matches data (no over- or underdispersion in poisson and binomial GLM)(→ check with dispersion parameter and graphical diagnostics)
- Linear relationship between η and predictor (→ check with Component-residual plot)
  - Non-linearity: Use nonlinear or nonparametric (e.g. GAMs) regression (see Zuur 2007)
- No observation overly influential (→ check with measures e.g. Cooks distance and graphical diagnostics)

Tools for complex data analysis

University of Koblenz-Landau 2018/19



Generalized linear models (GLMs)

Ralf B. Schäfer

These slides and notes complement the lecture with exercises “Tools for complex data analysis” for ecotoxicologists and environmental scientists. Do not hesitate to contact me if you have any comments or you find any errors (slides, slide notes, or code): schaefer-ralf@uni-landau.de

While I made notes below the slides, some aspects are only mentioned in the R demonstration associated with the lecture.

Learning targets

- Explaining and applying generalized linear models
- Explaining the concepts of maximum likelihood estimation and deviance
- Describe the specifics of GLMs regarding model selection and model assumptions

2

Learning targets and study questions

- Explaining and applying generalized linear models
  - Why should you use logistic regression for binomial data? Which assumptions of the linear model are violated and why?
  - How does the GLM deal with non-constant variance?
  - Outline differences in the model structure between a simple linear model and a GLM.
  - Describe typical error distribution and link functions for modelling a) species abundances and b) fraction of surviving organisms.
- Explaining the concepts of maximum likelihood estimation and deviance
  - Explain the core idea of maximum likelihood estimation and how it is done.
  - What is the difference between likelihood and probability?
  - Explain the concept of deviance for the GLM and why sum of squares can not be used.

3

Learning targets and study questions

- Describe the specifics of GLMs regarding model selection and model assumptions
  - Describe the methods that can be used for model selection and specifics for GLMs.
  - Which types of model diagnostics are required for a GLM, and which of these are particular for this class of models?
  - Explain the issue of overdispersion and options to deal with it.

4

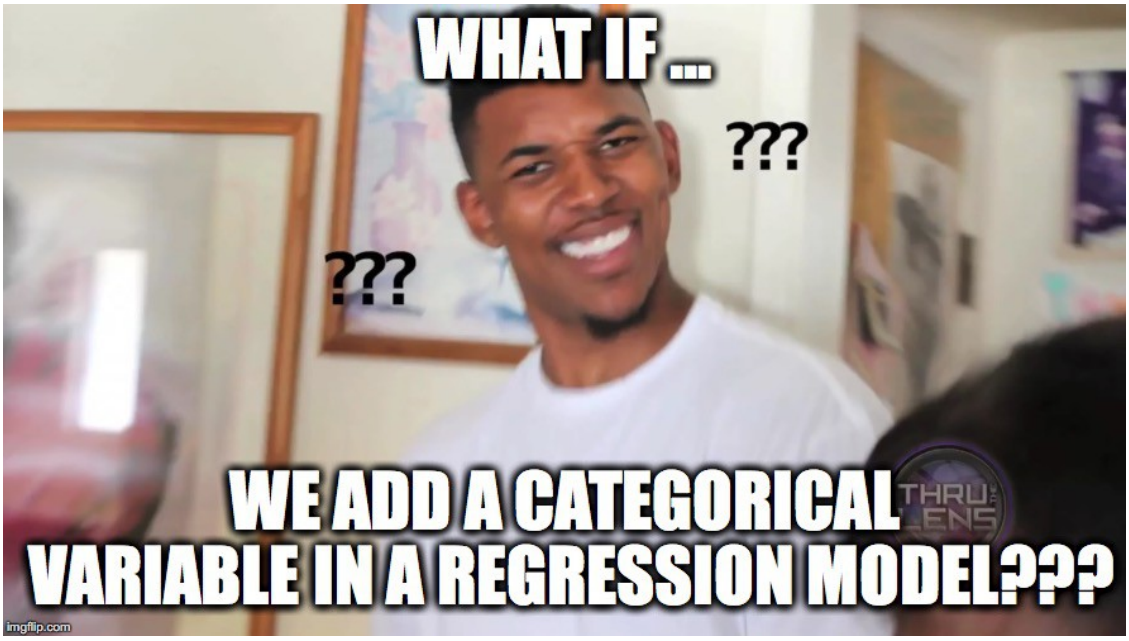
Generalized linear model

Contents

5

Extending the linear model: Motivation

Linear model requires a continuous response, but responses can be discrete (e.g. number of events, objects or organisms) or categorical (e.g. occurrence (presence/absence) of events, objects or organisms)



6

Imagine a study where the number of lizards was counted in different habitat patches and the research goal is to predict the number from environmental and spatial predictors.

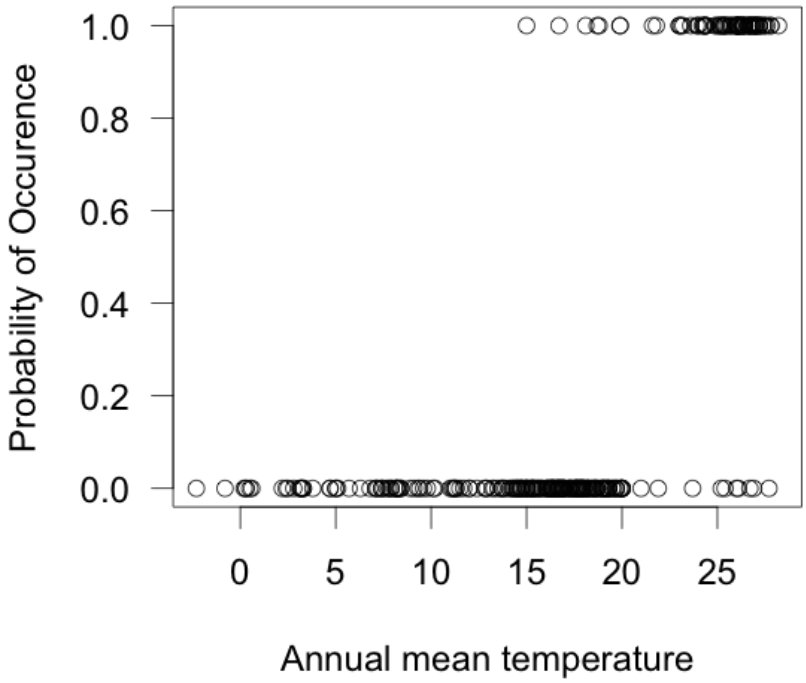
Similarly, imagine a study where the mortality (e.g. response is dead or alive) of ecotoxicological test organisms was assessed at different chemical concentrations and the research goal is to predict the mortality from the concentration.

In one of the previous sessions we used that meme and examined what happens if we include a categorical predictor in a linear model. Here, we include a discrete or categorical response, which requires an extension of the linear model.

Case study: Eastern brown snake

Research question: How does temperature influence the probability of occurrence of the eastern brown snake?

Study: Samples of potential habitats along temperature gradient



7

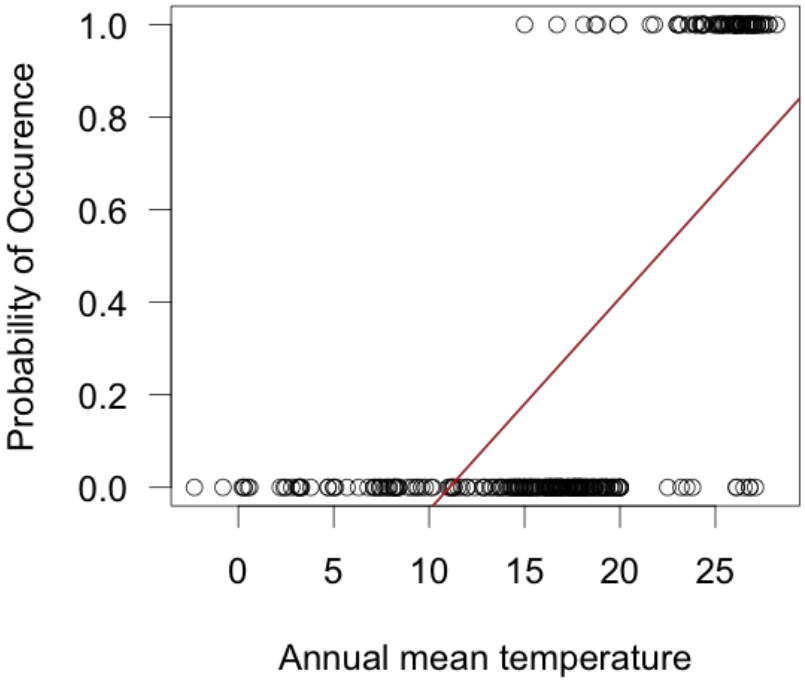
Picture taken from: https://en.wikipedia.org/wiki/Eastern_brown_snake#/media/File:Eastern_Brown_Snake_-_Kempsey_NSW.jpg

Case study: Eastern brown snake

Research question: How does temperature influence the probability of occurrence of the eastern brown snake?

Method: Linear regression model?

Provides meaningless probabilities <0 and >1!



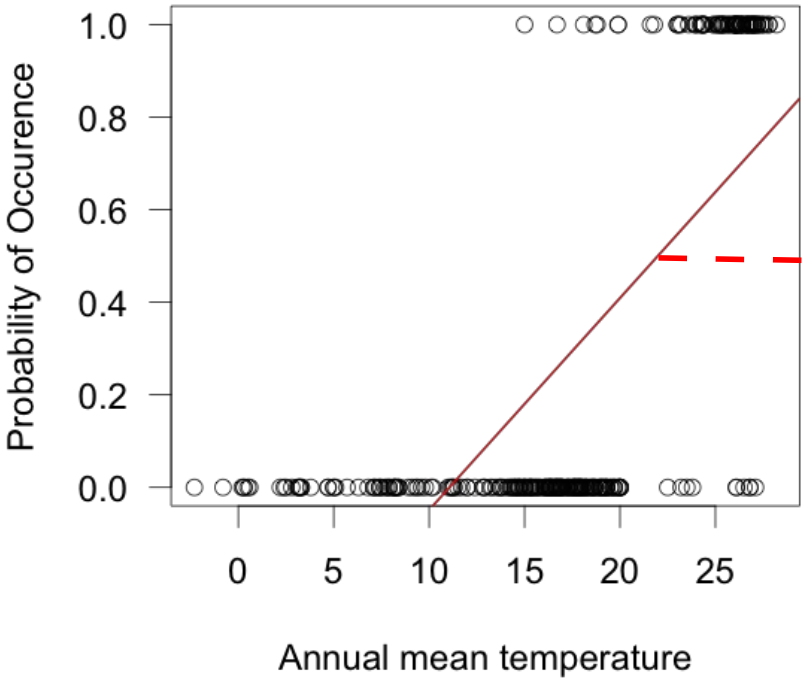
8

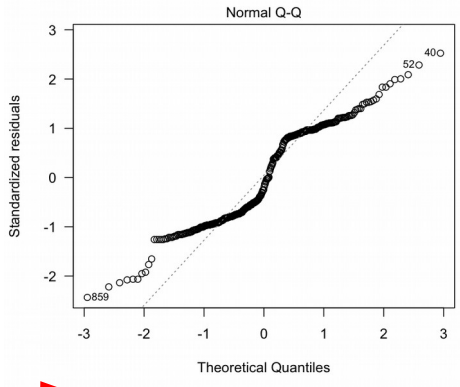
Case study: Eastern brown snake

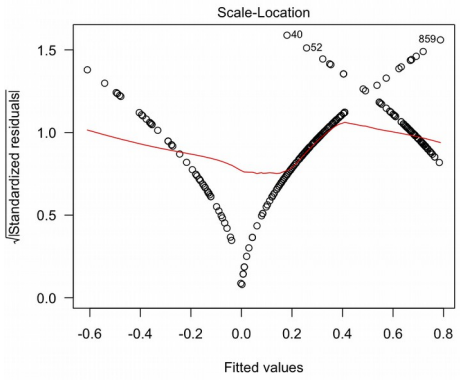
Research question: How does temperature influence the probability of occurrence of the eastern brown snake?

Method: Linear regression model?

Assumptions violated!







9

We note that the assumptions of variance homogeneity and of normal distribution of the error are clearly violated in this case. Such non-normal error distribution indicates that the relationship between the explanatory variable(s) and the response variable may be non-linear.

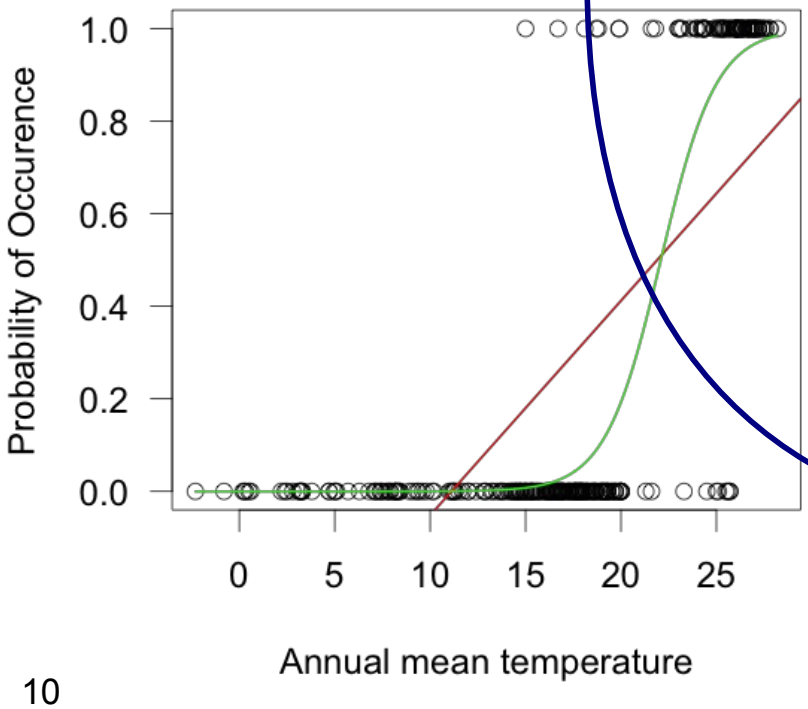
Further details on the problems of linear models to fit to dichotomous data are provided in Fox (2015) pp. 372-373.

From simple to logistic regression

Research question: How does temperature influence the probability of occurrence of the eastern brown snake?

Method: Simple linear regression *vs.* logistic regression

$$E(Y|X=x)=\mu(X)=\beta_0+\beta_1 X \qquad E(Y=1|X=x)=\mu(X)=\pi$$



$$\pi=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \Leftrightarrow \pi=\frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

Reformulation

Logit

$$\log_e\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1 X$$

Linear component

10

Remember that for the binomial distribution, in the beginning of the course (see Definitions document), we defined p as the probability of success. In the framework of logistic regression, this p corresponds to π following the convention that parameters from the statistical population are denoted with greek letters.

A thorough mathematical derivation of the logistic regression model is beyond the scope of this course, but can be found in almost any text book covering GLMs. See for example Matloff (2017: 154-158), Fox (2015: 375-378) or the seminal text book on GLMs: McCullagh & Nelder (1989: 98-114).

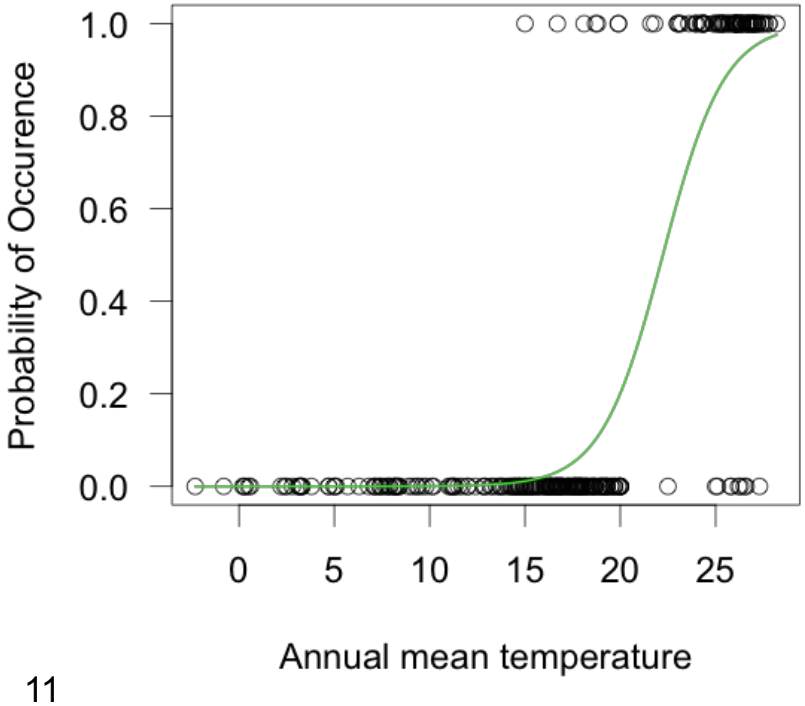
McCullagh P. & Nelder J.A. (1989) Generalized linear models, 2nd ed. Chapman & Hall/CRC, Boca Raton.

Parameters in logistic regression

Formula for fitted model

$$\log_e\left(\frac{\hat{\mu}(x_i)}{1-\hat{\mu}(x_i)}\right) = \log_e\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = b_0+b_1 x_i$$

Given are parameters of the logistic model: b_0 = -15.9, b_1 = 0.72



Which x relates to a $\hat{\pi}$ = 0.5?

$$\log_e\left(\frac{0.5}{1-0.5}\right) = -15.9+0.72x$$
$$\Leftrightarrow \log_e(1) = -15.9+0.72x$$
$$\Leftrightarrow 0+15.9 = 0.72x$$
$$\Leftrightarrow \frac{15.9}{0.72} = x \Rightarrow x = 22.1$$

Repeat the calculation for $\hat{\pi}$ = 0.1 and $\hat{\pi}$ = 0.9

Solution to exercise:

$$\log\left(\frac{0.1}{0.9}\right) = -\log\left(\frac{0.9}{0.1}\right)$$

The symmetry of the logistic model means that the left hand side of the equation for a given $\hat{\pi}$ (denoted $\hat{\pi}_{orig}$ hereafter) can be multiplied by -1 to solve the equation for a corresponding $\hat{\pi}_{corres} = 1-\hat{\pi}_{orig}$.

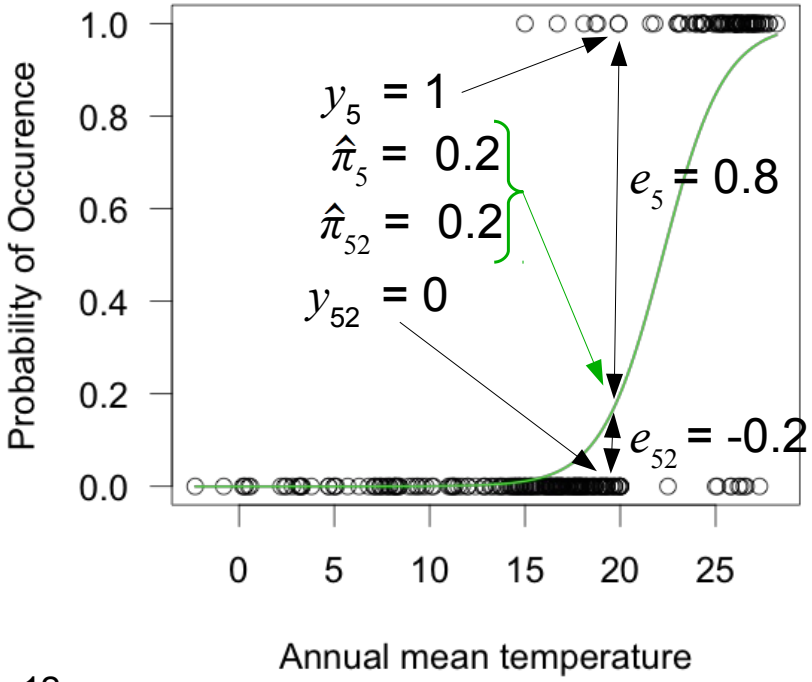
Note that $\hat{\pi}$ = 0.5 would represent the *LC50* (i.e. the median lethal effect concentration) in an ecotoxicological context.

Variance and logistic regression

$$\log_e\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \text{logit}(\hat{\pi}_i) = b_0+b_1 x_i \Leftrightarrow \hat{\pi}_i = \text{logit}^{-1}(b_0+b_1 x_i)$$

We define residuals as: $e_i = y_i - \text{logit}^{-1}(b_0 + b_1 x_i) = y_i - \hat{\pi}_i$

➡ Residuals are dichotomous: For $\hat{\pi}_i$ = 0.2, e_i is either 0.8 or -0.2.



Generally, the (true) variance of Y is given by:

$\text{Var}(Y) = \pi\,(1-\pi) = \mu\,(1-\mu)$

This means:

- Variance not constant, depends on π (i.e. μ) → Quadratic function
- Variance not normally distributed

12

For further details on logistic regression including an overview of different model types see for example Fox (2015) pp. 339, Gelman & Hill (2007) pp. 79 and Hilbe (2017).

In the context of GLMs, $\mu(x)$ is typically shortened to μ.

Gelman A. & Hill J. (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge; New York.

Hilbe J.M. (2017) Logistic regression models. CRC Press, Boca Raton.

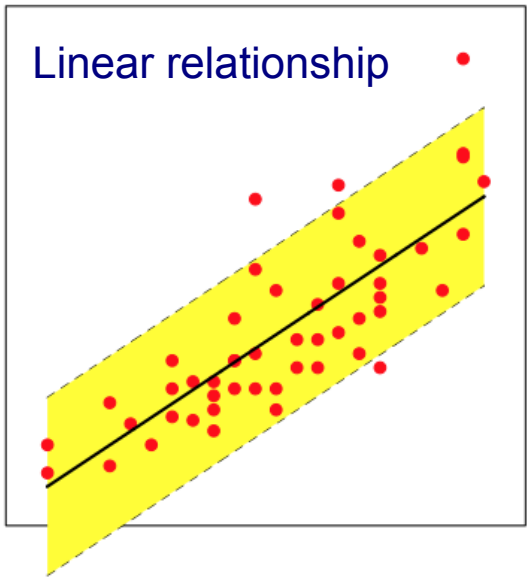
Generalized linear model

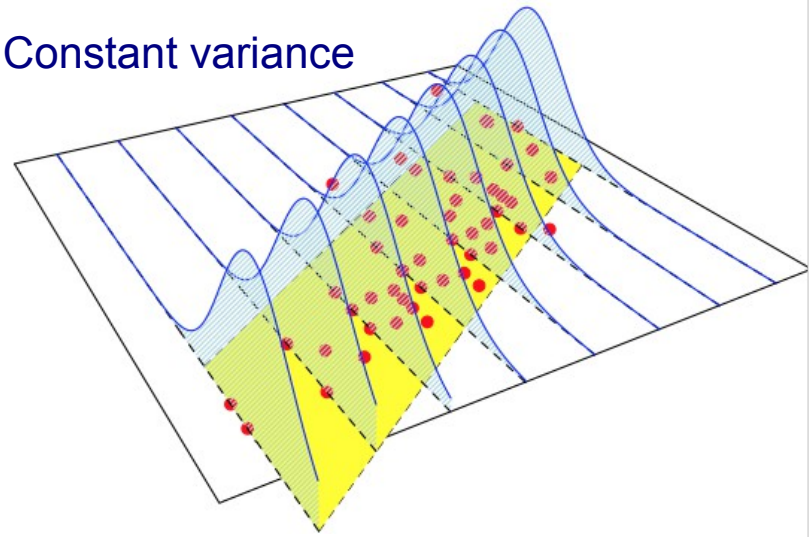
Contents

13

Remember: Simple linear regression

Linear model assumes linear relationship between explanatory variable(s) and response variable as well as a constant variance





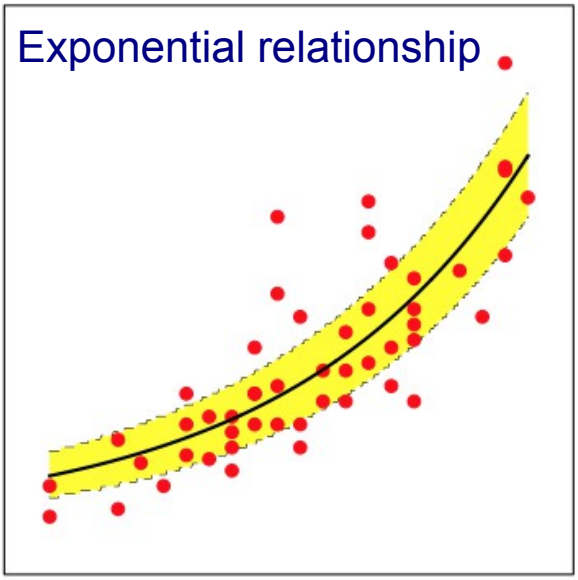
For ecological data, relationship with response variable can be non-linear and variance is often not constant

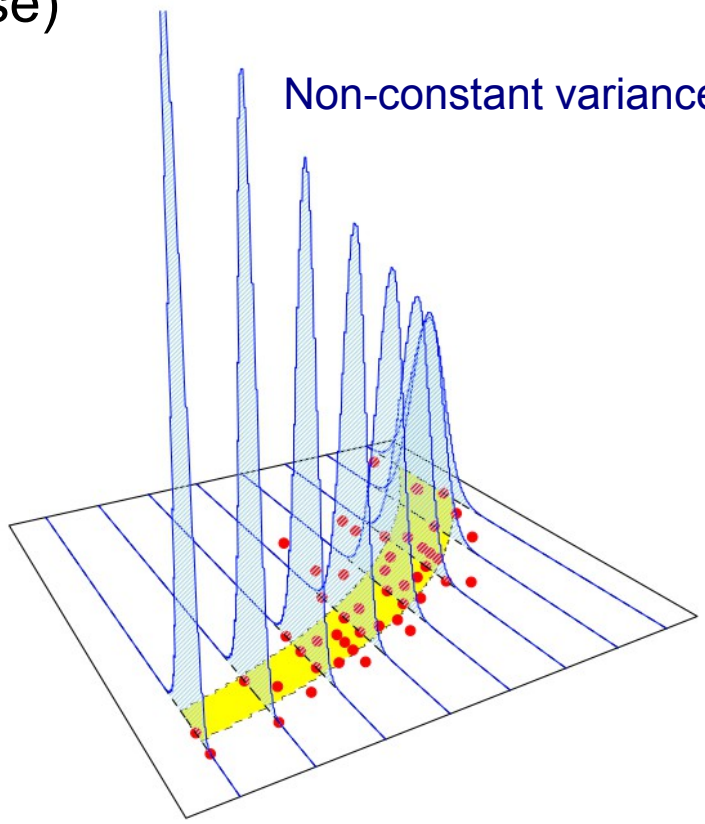
14 http://freakonometrics.hypotheses.org/tag/poisson

For the simple linear regression model, the variance of the response variable is constant (homoscedasticity).

Extending the linear model: Motivation

Example for non-linear relationship and non-constant variance with continuous response (in contrast to logistic regression with binary response)



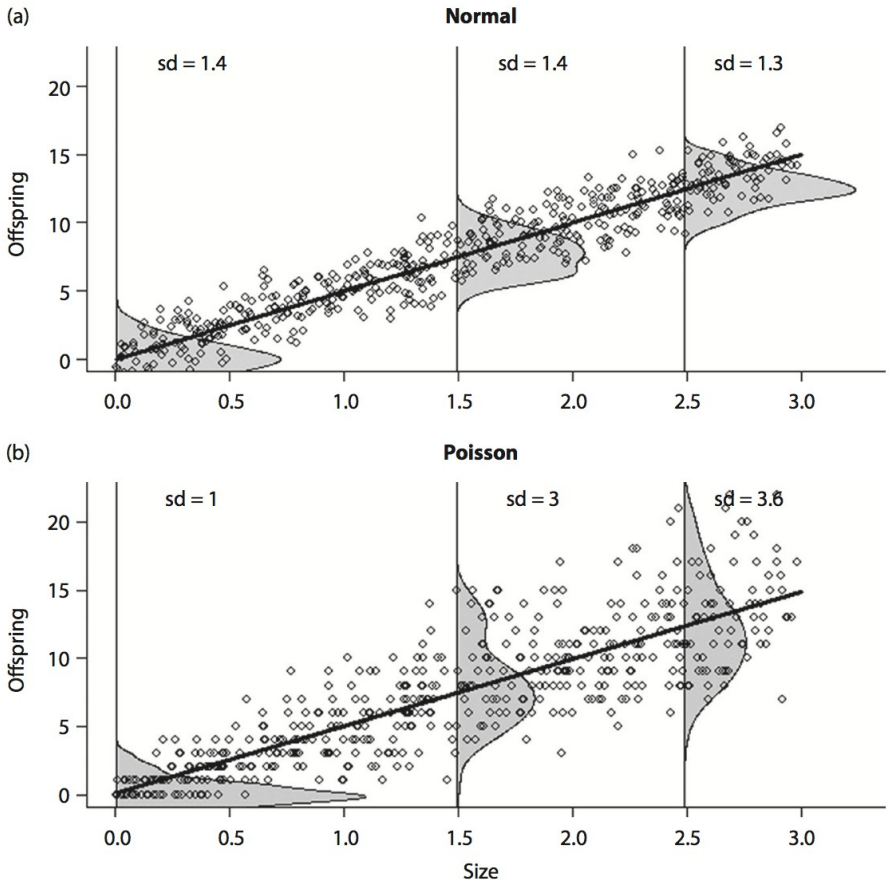


15 http://freakonometrics.hypotheses.org/tag/poisson

For example, the relationship between metabolism and body mass is represented by a power function (exponent = 0.75).

Extending the linear model: Motivation

Example: Increasing variability in number of offsprings with increasing body size of individuals



16

Buckley 2015: 134

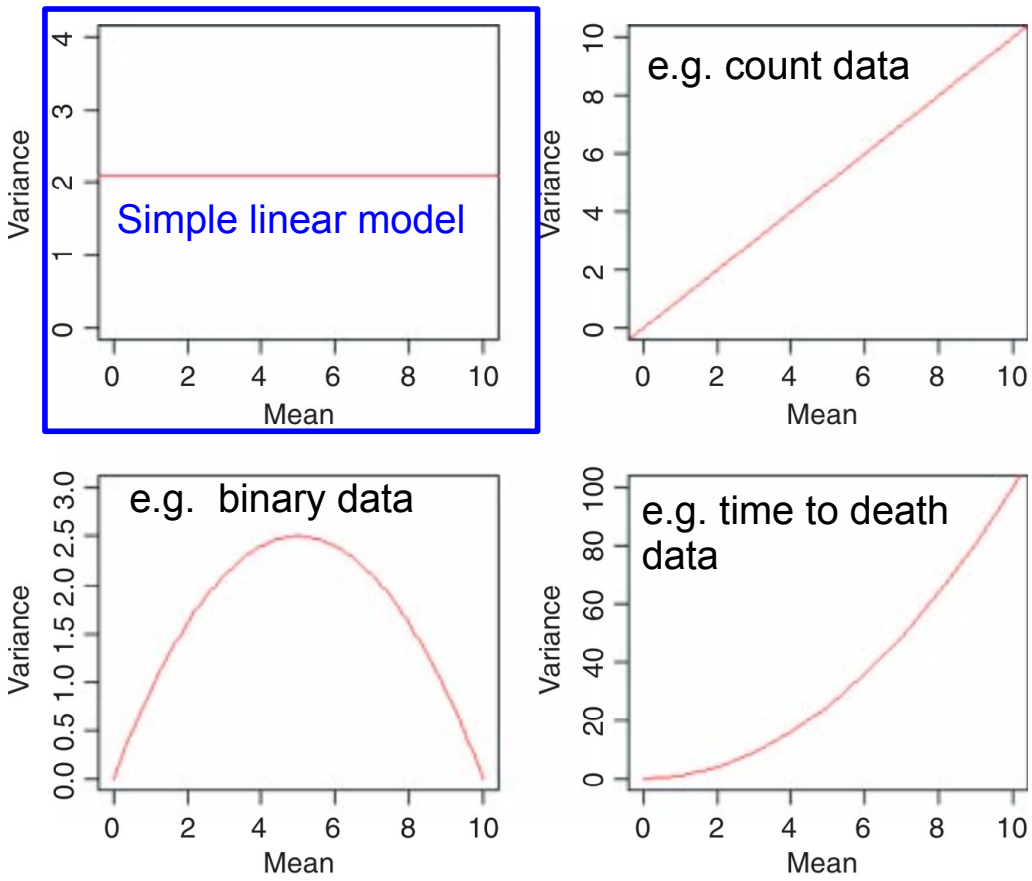
The top figure shows normally distributed residuals for the relationship between the number of offspring and body size. Although count data may approach a normal distribution in case of large data sets, the displayed residuals, which have been simulated, are not plausible, as they are associated with negative responses around 0 and generally should relate only to discrete numbers. Therefore, it is often no solution to transform data and then fit a linear model. This will be briefly discussed again later.

The figure below shows residuals that have been simulated using a Poisson distribution. Count data are typically Poisson distributed and we can see that the variance is not constant but increases with the mean, as for the logistic regression.

Buckley, Y.M. (2015): Generalized linear models in: Fox G.A., Negrete-Yankelevich S. & Sosa V.J. Eds: Ecological statistics: contemporary theory and application. Oxford University Press, Oxford. p. 132-148

Modelling the mean-variance relationship

Idea: Express variance as a function of the mean!



taken from Crawley 2012: 557

17

We have already discussed examples for binary data such as species presence-absence data or data from ecotoxicological experiments, where for each individual the response is dead or alive. The mean-variance relationship for binary data displayed in the bottom left figure follows that introduced in the context of logistic regression with $\text{Var}(Y) = \mu\ (1\text{-}\mu)$. In fact, the equation used here is adapted to μ being the number of successes rather than proportions: $\text{Var}(Y) = (\mu\ (n\text{-}\mu))\ n^{-1}$, where n is the total number of trials (in the figure n = 10).

An example for count data has been given on the previous slide.

Generalized linear model

Contents

18

Defining the GLM

Linear model: $Y = \beta_0 + \beta_1 X + \varepsilon$

Generalised linear model:

1. Linear predictor: $\eta = \beta_0 + \beta_1 X$
2. Link function: $g(\mu) = \eta$ with $E(Y|X = x) = \mu$
3. Distribution of Y with related $\text{Var}(Y) = \phi\, V(\mu)$

Error structure with related variance function and typical link function

Family (error structure)	Default Link	Link name	Variance function
`gaussian`	$\eta = \mu$	identity	1
`poisson`	$\eta = \log_e \mu$	log	μ
`binomial`	$\eta = \log_e\left(\frac{\mu}{(n-\mu)}\right)$	logit	$\frac{\mu(n-\mu)}{n}$
`Gamma`	$\eta = \mu^{-1}$	inverse	μ^2
`inverse.gaussian`	$\eta = \mu^{-2}$	inverse square	μ^3

modified from Crawley 2012: 562

Beside many other text books, Fox (2015) gives a thorough introduction to GLMs. Zuur et al. (2013), Faraway (2016) and Fox & Weisberg (2019) focus on the implementation in R.

In contrast to linear models, GLMs can model discrete and categorical responses and non-constant error variance by expressing the variance as a function of the mean and allowing for non-normal error distributions. In the past, data were often transformed to reach normal distribution. This is not necessary if the data can be directly modelled with a GLM. Transformed data can lead to biased estimates, higher variance and lower power. See Matloff (2017: pp. 137), O'Hara & Kotze (2010), Warton & Hui (2011) and Szöcs & Schäfer (2015) for more detailed discussions.

$E(Y)$ is the expected value, i.e. the mean, of the response variable Y, which is a random variable. $V(\mu)$ is the variance function. For the simple regression model, $g(\mu) = \mu$ and $\phi\, V(\mu) = \phi\, 1$. ϕ is the dispersion (or scale) parameter that is taken to be known with a value of 1 for a poisson and binomial model, and σ^2 for a linear model. Note that binomial data can be expressed in two ways: the number of trials n and successes k can be expressed as proportion, i.e. k/n or as absolute number of successes k given n trials. The table presents the link and variance function for the latter. If the GLM is specified using proportions, the n in the link and variance function is set to 1 and $\mu = k/n$ instead of $\mu = k$ (often μ is denoted with π).

The table gives error distributions with the related variance function and the typical (e.g. default in R) link function. However, alternative link functions could be used, which is elaborated later. For example, in ecological studies the gamma distribution is often combined with the log link. For an overview of the inverse link functions see Fox & Weisberg (2019: 274).

Faraway J.J. (2016) Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. CRC Press, Boca Raton FL, USA.

O'Hara R.B. & Kotze D.J. (2010) Do not log-transform count data. *Methods in Ecology and Evolution* 1, 118–122.

Szöcs E. & Schäfer R. (2015) Ecotoxicology is not normal. *Environmental Science and Pollution Research* 22, 13990–13999.

Warton D.I. & Hui F.K.C. (2011) The arcsine is asinine: the analysis of proportions in ecology. *Ecology* 92, 3–10.

Zuur A.F., Hilbe J.M. & Ieno E.N. (2013) A beginners guide to GLM and GLMM with R: a frequentist and bayesian perspective for ecologists. Highland Statistics, Newburgh.

19

General and specific GLMs

Response Y follows distribution from exponential family:

$$f_\theta(y) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)}$$

Specific (exponential) distributions:

Gaussian

$Y \sim \text{Normal}(\mu, \sigma)$
$E(Y) = \mu$
$\text{Var}(Y) = \sigma^2$
$\mu = \beta X$
$\varepsilon = y - \mu$

Binomial

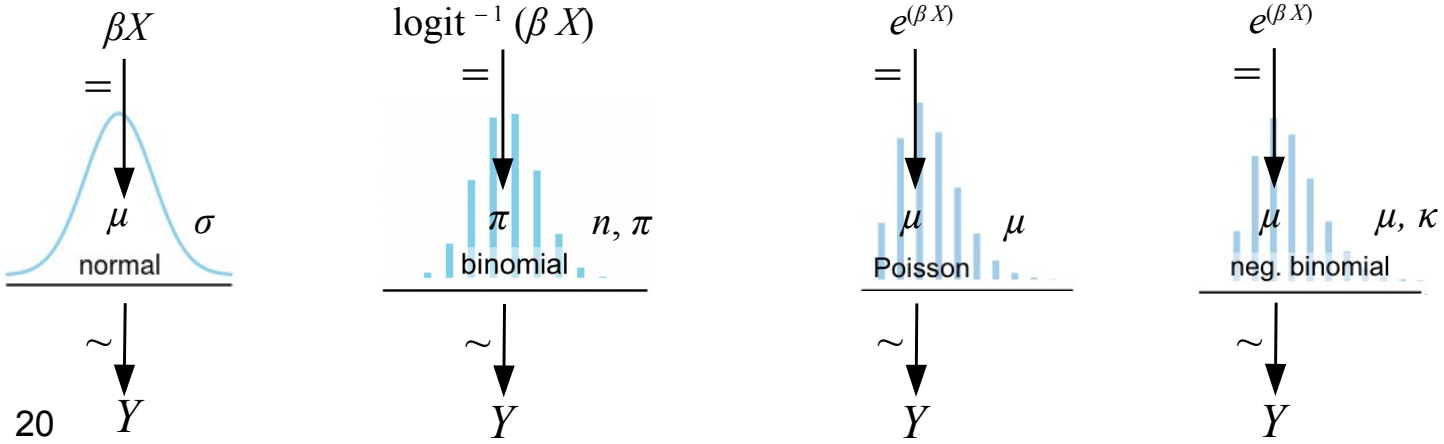
$Y \sim \text{Bin}(n, \pi)$
$E(Y) = \pi$
$\text{Var}(Y) = \frac{\pi(n-\pi)}{n}$
$\text{logit}(\pi) = \beta X$
$\varepsilon = y - \pi$

Poisson

$Y \sim \text{Pois}(\mu)$
$E(Y) = \mu$
$\text{Var}(Y) = \mu$
$\log(\mu) = \beta X$
$\varepsilon = y - \mu$

Negative binomial

$Y \sim \text{Neg.Bin}(\mu, \kappa)$
$E(Y) = \mu$
$\text{Var}(Y) = \mu + \frac{\mu^2}{\kappa}$
$\log(\mu) = \beta X$
$\varepsilon = y - \mu$



$f_\theta(y)$ is the probability density (or mass) function for Y given the parameter θ (which completely depends on β and represents the link in the table on the previous slide). a, b and c are known functions, and ϕ is the scaling or dispersion parameter (as introduced before). All specific distributions on the slide are representatives of this family of distributions, for details how to specify the functions a, b and c see the extract from Wood (2017) in the literature on OpenOLAT (or directly Wood 2017: 104). Note that the first derivate from $b(\theta)$ provides $E(Y)$ and that the second derivative multiplied with $a(\phi)$ yields to $\text{Var}(Y)$ (for details see Wood 2017: 102-105). For a slightly different description/notation refer to Dobson & Barnett (2018: 49-64). A very distilled overview on the theoretical background is provided in Fox & Weisberg (2019: 272-275). Finally, an example, i.e. the calculation of derivatives for the binomial distribution, is provided in Hilbe 2017: 63-65.

You should notice that the Gaussian GLM (with identity link) is equivalent to the linear regression model that we discussed intensively. In fact, the linear model is a special case of the GLM.

The negative binomial distribution represents an important distribution that in several cases fits ecological count data better than the poisson distribution, because it extends the poisson distribution with an additional parameter. However, the negative binomial should not always be used when the poisson distribution does not match (i.e. is overdispersed, which is defined later). We will discuss this in more detail later.

The visualisation of the specific GLMs was motivated by a blog: http://www.sumsar.net/blog/2013/10/how-do-you-write-your-model-definitions/ and the diagrams are available in R: https://github.com/rasmusab/distribution_diagrams

Dobson A.J. & Barnett A.G. (2018) An introduction to generalized linear models, Fourth edition. CRC Press, Taylor & Francis Group, Boca Raton.
Hilbe J.M. (2017) Logistic regression models. CRC Press, S.l.
Wood S.N. (2017) Generalized additive models: an introduction with R, Second edition. CRC Press/Taylor & Francis Group, Boca Raton.

Data type and GLM specification

Response variable	Error distribution	Canonical link function	Alternative link functions
Continuous positive and negative values	Gaussian/Normal	Identity	Log, Inverse
Counts	Poisson	Log	Identity, Sqrt
Counts with over-dispersion	Negative Binomial, Quasi-Poisson	Log Log	As per Poisson
Proportions (no. successes/total trials)	Binomial	Logit	Probit, Cauchit, Log, Complementary Log-Log
Binary (male/female, alive/dead)	Binomial (Bernoulli)	Logit	As per Binomial
Proportions or binary with overdispersion	Quasi-Binomial	logit	As per Binomial
Time to event (germination, death)	Gamma	Inverse	Inverse, Identity, Log

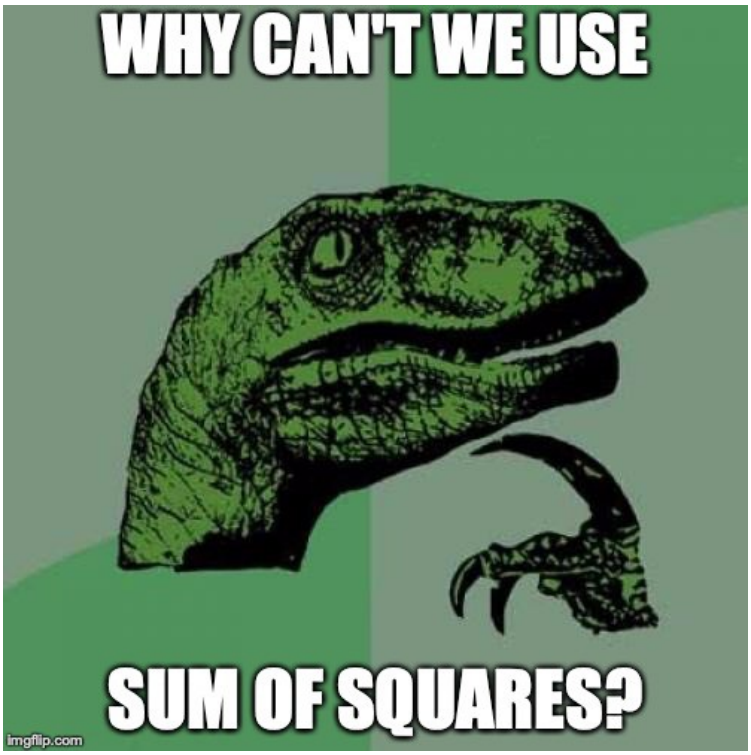
21

Buckley 2015: 137

Continuous positive data are similar to time to event data and can be modelled accordingly.

Deviance: Goodness of fit for GLM

- GLMs minimize Deviance (D) instead of Sum of Squares in simple linear regression model
- Deviance derived by *Maximum Likelihood Estimation* (MLE)



22

Why can't we use sum of squares?

Example for logistic regression

Remember: For (simple) linear model, we determine coefficients by minimizing residual sum of squares (RSS):

$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - \left(b_0 + b_1 x_i\right)\right)^2 \Leftrightarrow \arg\min_{b_0, b_1} \sum_{i=1}^{n} e_i \Leftrightarrow \arg\min_{b_0, b_1} \sum \mathrm{RSS}$$

Application of concept to (simple) logistic regression:

$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} e_i \Leftrightarrow \arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - \frac{1}{1 + e^{-\left(b_0 + b_1 x_i\right)}}\right)^2$$

→ No simple algebraic matrix solution as for linear model
→ Solution based on maximum likelihood estimation

23

Check out the slides for the second session for the algebraic matrix solution for the linear model, in case you need to revisit.

Generalized linear model

Contents

1. Case study: Logistic regression
2. Extending the linear model
3. Definition of the GLM
4. **Deviance and Likelihood**
5. Model selection and diagnostics

Classical textbooks devoted to GLMs are McCullagh & Nelder (1989, Generalized Linear Models, 2nd edition, London: Chapman and Hall) and McCulloch & Searle (2001, Generalized, Linear, and Mixed Models. John Wiley & Sons).

Maximum Likelihood Estimation (MLE)

Method of estimating model parameters that maximise the likelihood function given the data. In other words, model parameters are estimated that have the highest likelihood to produce the sample data.

General likelihood function: $L(\theta|y) = f_\theta(y)$

Search for maximum: $\arg\max_\theta L(\theta|y) = \arg\max_\theta f_\theta(y)$

The ML estimate $\hat{\theta}$ is defined as: $\hat{\theta} \in \{\arg\max_\theta L(\theta|y)\}$

<u>How to identify the maximum likelihood estimate?</u>

→ Set first derivative equal to zero: $\frac{\partial L(\theta|y)}{\partial \theta} = 0$

We give the likelihood function for a continuous probability distribution. For a discrete probability distribution, it would be: $L(\theta|y) = P_\theta(Y = y)$.

Note the difference between probability and likelihood. Probability describes the plausibility of a random (future) outcome, given model parameters. For example, in the context of the binomial distribution, the probability P of the random future outcome y (number of successes), representing a realisation from a random variable Y, given the parameters n = 10 (number of trials) and p = 0.3 (probability of success in a single trial) is given by the probability mass function:

$$P(Y = y|n, p) = \binom{n}{y} p^y (1-p)^{n-y} = \binom{10}{y} 0.3^y (0.7)^{10-y}$$

Conversely, the likelihood L describes the plausibility of model parameters given specific outcomes that have occurred. Given that the data are produced, we know the number of trials in the context of the binomial model (e.g. n = 15) and the number of successes (e.g. y = 10). Hence, our likelihood function is:

$$L(p|y, n) = \binom{15}{10} p^{10} (1-p)^5$$

Example: Likelihood calculation

Research question: What is the probability p of an insect to be killed at a specific chemical concentration?

Study: Laboratory test with 15 insects. 10 died.

$$L(p|y,n) = \binom{15}{10} p^{10}(1-p)^{5}$$

p = ?; probab. of death in single trial (i.e. insect)
n = 15; no. of trials (i.e. insects)
y = 10; no. of deaths

→ Estimate parameter p with MLE

$$\hat{p} \in \{\arg\max_{p} L(p|x)\}$$

Log likelihood simplifies derivation

$$\log L(p|y,n) = \log\left(\binom{15}{10} p^{10}(1-p)^{5}\right) = \log 3003 + 10 \log p + 5 \log(1-p)$$

$$\frac{\partial L(\theta|x)}{\partial \theta} = 0 \Rightarrow \frac{\partial \log L(\hat{p}|y,n)}{\partial p} = 0$$

26

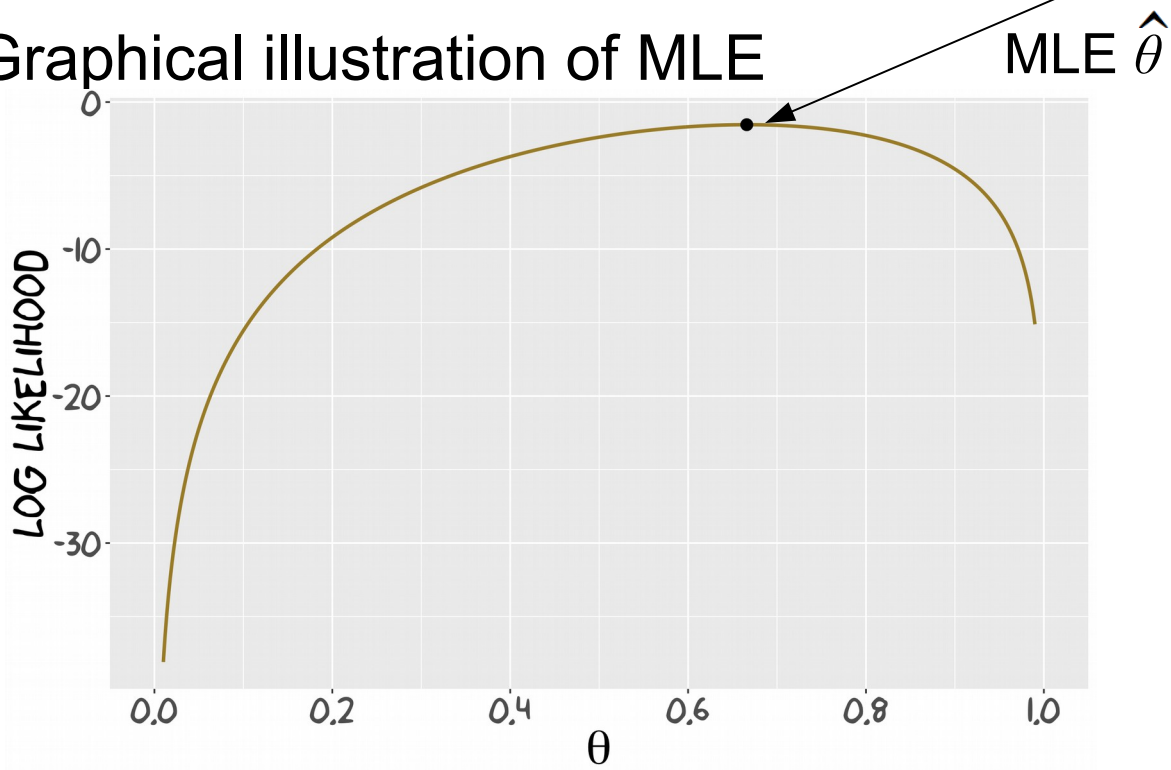
We use the probability mass function of the binomial distribution to calculate the respective likelihood function.

Example: Likelihood calculation

$$\frac{\partial \log L(\hat{p}|y,n)}{\partial p} = 0 \Leftrightarrow \frac{\partial(8+10\log\hat{p}+5\log(1-\hat{p}))}{\partial p} = 0 \Leftrightarrow$$

$$\frac{10}{\hat{p}}+\frac{5}{(\hat{p}-1)} = 0 \Leftrightarrow \frac{5}{(\hat{p}-1)} = -\frac{10}{\hat{p}} \Leftrightarrow 5\hat{p} = -10\hat{p}+10 \Leftrightarrow \hat{p} = \frac{2}{3}$$

Graphical illustration of MLE



A detailed treatment of likelihood, maximum likelihood and the GLM is beyond the scope of this course, but can be found in Dobson & Barnett (13-23).

Dobson A.J. & Barnett A.G. (2018) An introduction to generalized linear models, Fourth edition. CRC Press, Taylor & Francis Group, Boca Raton.

MLE and logistic regression

Probability P for simple logistic regression

$$P(y_i=0|x_i, b_0, b_1) = 1-\frac{1}{1+e^{-(b_0+b_1 x_i)}}$$

$$P(y_i=1|x_i, b_0, b_1) = \frac{1}{1+e^{-(b_0+b_1 x_i)}}$$

Likelihood L of any observation (x_i, y_i) in simple logistic regression:

$$L(b_0, b_1|y_i, x_i) = \left(\frac{1}{1+e^{-(b_0+b_1 x_i)}}\right)^{y_i}\left(1-\frac{1}{1+e^{-(b_0+b_1 x_i)}}\right)^{1-y_i}$$

Likelihood L of all n observations (x, y) is:

$$L(b_0, b_1|y, x) = \prod_{i=1}^{n}\left(\frac{1}{1+e^{-(b_0+b_1 x_i)}}\right)^{y_i}\left(1-\frac{1}{1+e^{-(b_0+b_1 x_i)}}\right)^{1-y_i}$$

28

In the previous example, we chose the parameter that maximised the likelihood of obtaining the data for a single chemical concentration. If we have multiple concentrations (here observation pairs (x_i, y_i), this becomes more complicated and we illustrate this for the case of simple logistic regression.

MLE and GLMs

log Likelihood L of all n observations (x, y) is:

$$\log L(b_0, b_1|y, x) = \sum_{i=1}^{n} \left(-\log\left(1+e^{-(b_0+b_1 x_i)}\right)+y_i(b_0+b_1 x_i)\right)$$

General: log Likelihood for GLMs

$$\log L(\theta, \phi|y) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$$

Solution to MLE based on complex matrix algebra and different algorithms, for details see: Dobson & Barnett (2018: 71-76), Wood (2017: 105-107) and Hilbe (2017: 51-62)

A detailed treatment of likelihood, maximum likelihood and the GLM is beyond the scope of this course, but can be found in Dobson & Barnett (2018: 13-23) and more comprehensively in Millar (2011). The authors also provide an example for a MLE calculation for a case study (pp. 154-155).

Practically, the fitting is achieved by an iterative re-weighted least squares (IRWLS) algorithm.

Dobson A.J. & Barnett A.G. (2018) An introduction to generalized linear models, Fourth edition. CRC Press, Taylor & Francis Group, Boca Raton.

Hilbe J.M. (2017) Logistic regression models. CRC Press, S.l.

Millar R.B. (2011) Maximum likelihood estimation and inference: with examples in R, SAS, and ADMB. Wiley, Chichester, West Sussex.

Wood S.N. (2017) Generalized additive models: an introduction with R, Second edition. CRC Press/Taylor & Francis Group, Boca Raton.

Deviance: Goodness of fit for GLM

Definition of residual deviance D:

$$D(y|\hat{\mu}) = -2(\log L_m - \log L_s) = -2[\log L(\hat{\mu}, \phi|y) - \log L(y, \phi|y)]$$

L_m is the maximised likelihood of our current model that is nested in a saturated model with maximised likelihood L_s (i.e. a model fitting the data as closely as possible)

Deviance can generally be used for model comparison

log likelihood and Information criteria

The information-theoretic criteria AIC and BIC also rely on the log likelihood $\log L$:

$$\text{AIC} = -2\log L + 2p$$

$$\text{BIC} = -2\log L + \ln(n)p$$

n = sample size

p = parameters in model

30

The deviance for two models is also called log likelihood ratio statistic. It follows approximately a χ^2 distribution (through multiplication with 2).

The saturated model contains as many parameters as are necessary to fully describe all observations y. Hence, a low deviance means that our model of interest has a similar explanatory power, typically with only a few parameters (for details see Dobson & Barnett (2018: 86-88). Higher deviance means lower fit of a model.

For an overview on different deviances see the extract from Wood (2017: 104) on OpenOLAT or with more explanation Dobson & Barnett (2018: 88-92). The deviance for the Gaussian model, i.e. the linear model, is equivalent to RSS.

Several pseudo-R^2 measures have been developed for the different GLMs, but they are beyond the scope of our course.

Dobson A.J. & Barnett A.G. (2018) An introduction to generalized linear models, Fourth edition. CRC Press, Taylor & Francis Group, Boca Raton.

Generalized linear model

Contents

1. Case study: Logistic regression
2. Extending the linear model
3. Definition of the GLM
4. Deviance and Likelihood
5. **Model selection and diagnostics**

31

Model selection for GLM

- Same methods as for multiple linear regression model
- Best subset and multi-model averaging
- Hypothesis-based stepwise model selection:
  - Wald test for individual regression coefficients
  - Log-likelihood ratio test for complete model comparison
- Information-theoretic stepwise model selection (e.g. AIC, corrected AIC, BIC)
- Post-selection shrinkage and LASSO

The Wald test should not be trusted for small (e.g. n = 10) sample sizes (Agresti 2007: 13). The Log-likelihood ratio test represents a more robust alternative.

For details on model comparison see Wood 2017: 108-110.

Agresti A. (2007) An introduction to categorical data analysis, 2nd ed. Wiley-Interscience, Hoboken, NJ.

GLM assumptions and diagnostics

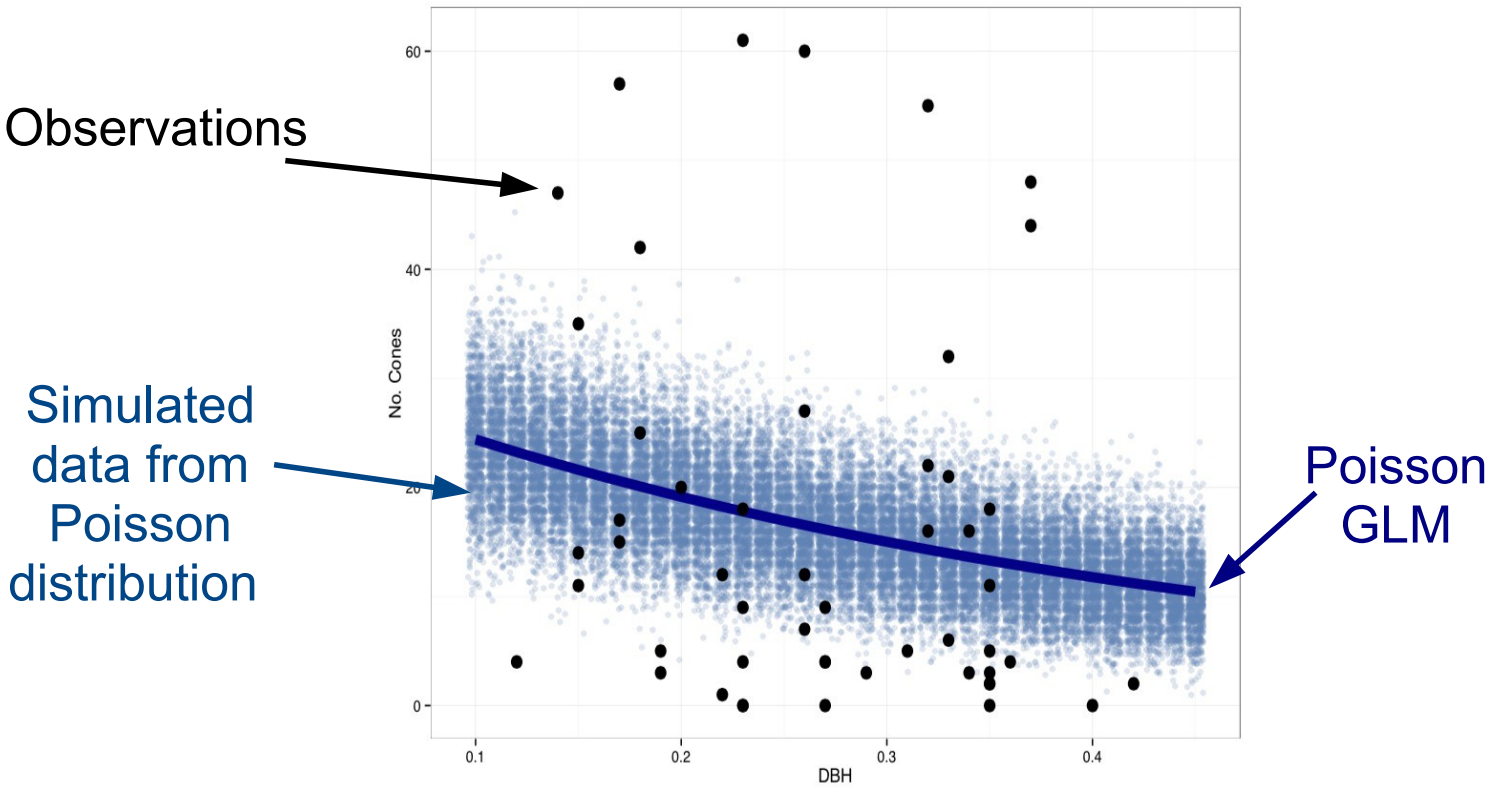
- Assumptions & tools from linear model largely apply to GLM
- Independence of observations
  - For spatiotemporal autocorrelation: GLMMs (see Bolker 2009)
- Assumed mean-variance relationship matches data (no over- or underdispersion in poisson and binomial GLM) (→ check with dispersion parameter and graphical diagnostics)

33

The first step in model diagnosis of binomial and poisson models should be checking of the mean-variance relationship, i.e. whether this matches the assumed dispersion parameter of 1. Overdispersion means that the mean-variance relationship in the binomial or poisson model is higher than assumed (and the opposite applies to underdispersion). As a rule of thumb, over- or underdispersion is indicated when the dispersion parameter approaches or exceeds 2 or 0.5, respectively (Logan 2010: 493). The dispersion parameter is calculated as the residual deviance divided by the degrees of freedom. The calculation and graphical methods are discussed in the R demonstration.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., and White, J.S.S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol* 24, 127–135.

Overdispersion

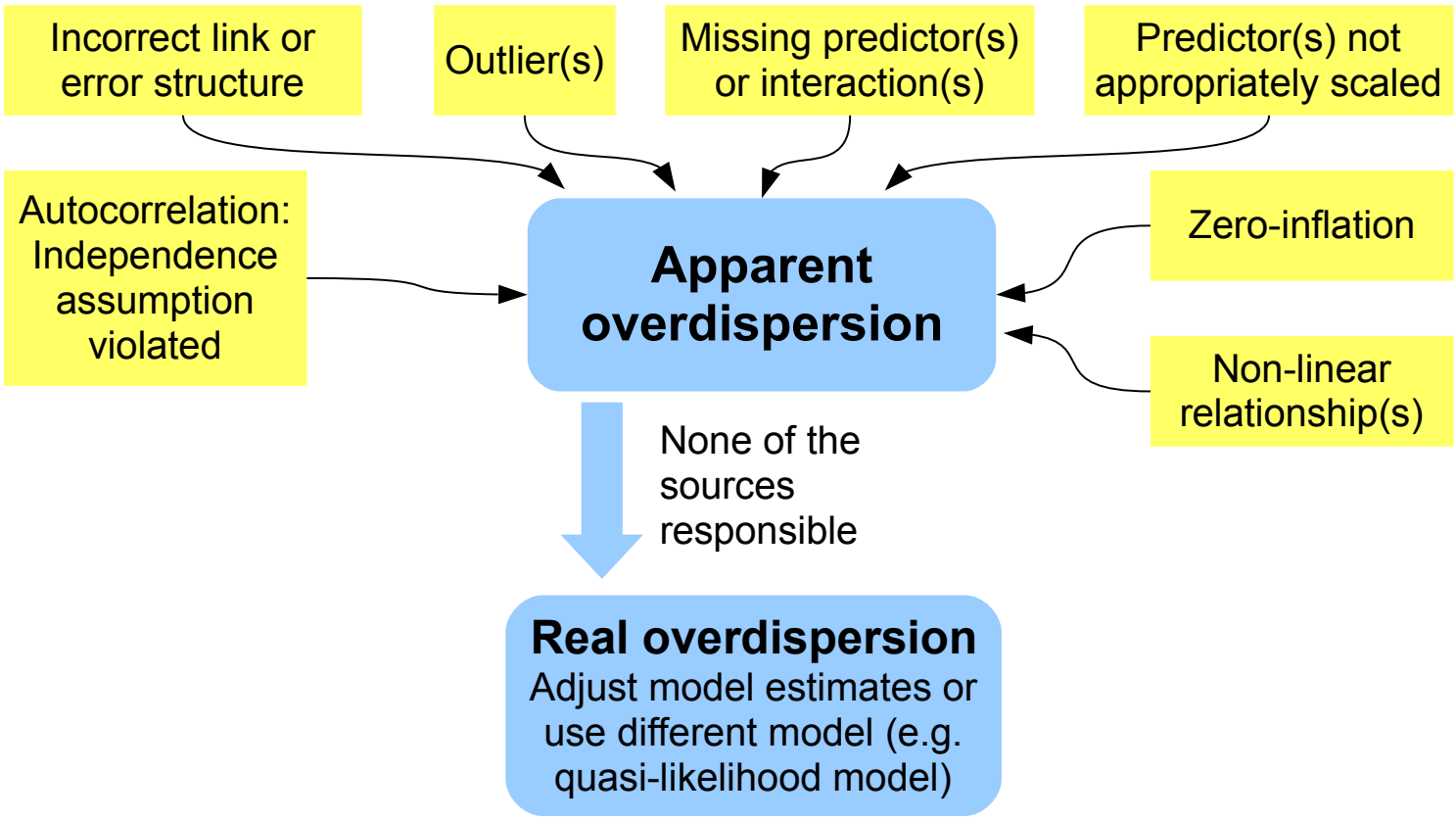


→ Too narrow standard errors (and in turn *p*-values or CIs) and too high estimates for regression coefficients

34

How to deal with overdispersion?

First step: Check for potential source of overdispersion



35

modified from Zuur et al. 2013: 20

Note that for beginners over- or underdispersion is often due to the selection of an incorrect error distribution or link function. Thus, their plausibility should be checked first. Most other sources should be self-explanatory. If predictors are not scaled appropriately, this can be cured by transforming predictors (e.g. if a variable exhibits a squared relationship with a response but enters the model non-squared, this may lead to overdispersion). Zero-inflation means that the data contain many zeros, which is often the case for ecological data. For further details on how to diagnose and deal with overdispersion see Zuur et al. (2013: 19-29) and Hilbe (2017: 319-352).

Quasi-likelihood models such as the quasi-binomial and quasi-poisson model only assume a specific mean-variance relationship but treat the distribution of the response as unknown. They estimate the scale parameter ϕ from the data. For further background on quasi-likelihood models see Wood (2017: 113-115) and Fox (2015: 431-432).

Overdispersion is very common, particularly for count data. Therefore, alternative models such as the quasipoisson and the negative binomial model represent frequently used alternatives. How to decide between the two is discussed in Ver Hoef & Boveng (2007). For an overview on 1) how to adjust model estimates and 2) alternative models in the case of binomial overdispersion see Hilbe (2017: 342).

Hilbe J.M. (2017) Logistic regression models. CRC Press, S.l.

Ver Hoef J.M. & Boveng P.L. (2007) Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology* 88, 2766–2772.

Wood S.N. (2017) Generalized additive models: an introduction with R, Second edition. CRC Press/Taylor & Francis Group, Boca Raton.

Zuur A.F., Hilbe J.M. & Ieno E.N. (2013) A beginners guide to GLM and GLMM with R: a frequentist and bayesian perspective for ecologists. Highland Statistics, Newburgh.

GLM assumptions and diagnostics

- Assumptions & tools from linear model largely apply to GLM
- Independence of observations
  - For spatiotemporal autocorrelation: GLMMs (see Bolker 2009)
- Assumed mean-variance relationship matches data (no over- or underdispersion in poisson and binomial GLM)(→ check with dispersion parameter and graphical diagnostics)
- Linear relationship between η and predictor (→ check with Component-residual plot)
  - Non-linearity: Use nonlinear or nonparametric (e.g. GAMs) regression (see Zuur 2007)
- No observation overly influential (→ check with measures e.g. Cooks distance and graphical diagnostics)

36

Note that the ordinary residuals of linear models (so-called *response residuals*) should not be used for interpretation of GLMs other than the Gaussian GLM, given that they ignore the non-constant variance that is associated with GLMs. A range of other residuals have been introduced such as *Pearson residuals* (accounting for the variance and dispersion parameter), *Deviance residuals* (accounting for deviance differences) and *Randomized quantile residuals* (resulting from the conversion to normal residuals).

Moreover, standardized and studentized versions of these residuals are available (see the session on linear model diagnostics if you forgot what these are). A complete overview on the residuals is given in Hilbe (2017: 268-284), and, more condensed, in Fox & Weisberg (2019: 418-421). Although the latter two text books recommend the use of Deviance residuals, these are more difficult to interpret than the randomized quantile residuals (for a comparison see: https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html). Details on the latter are provided by Dunn & Smyth (1996).

Dunn P.K. & Smyth G.K. (1996) Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics* 5, 236–244.