Generalized linear model

Contents

Learning targets

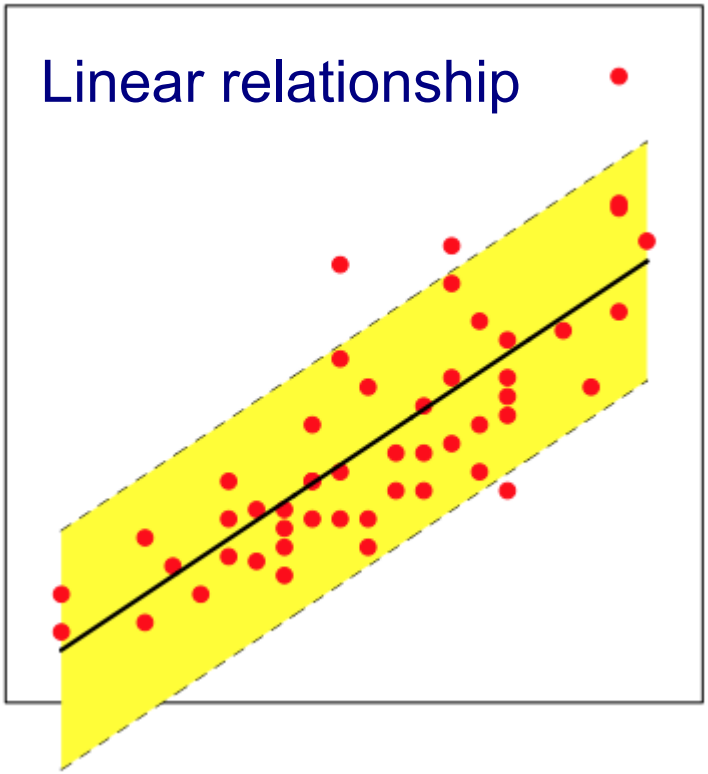
- Explaining and applying generalized linear models
- Describe the specifics of GLMs regarding model selection and model assumptions

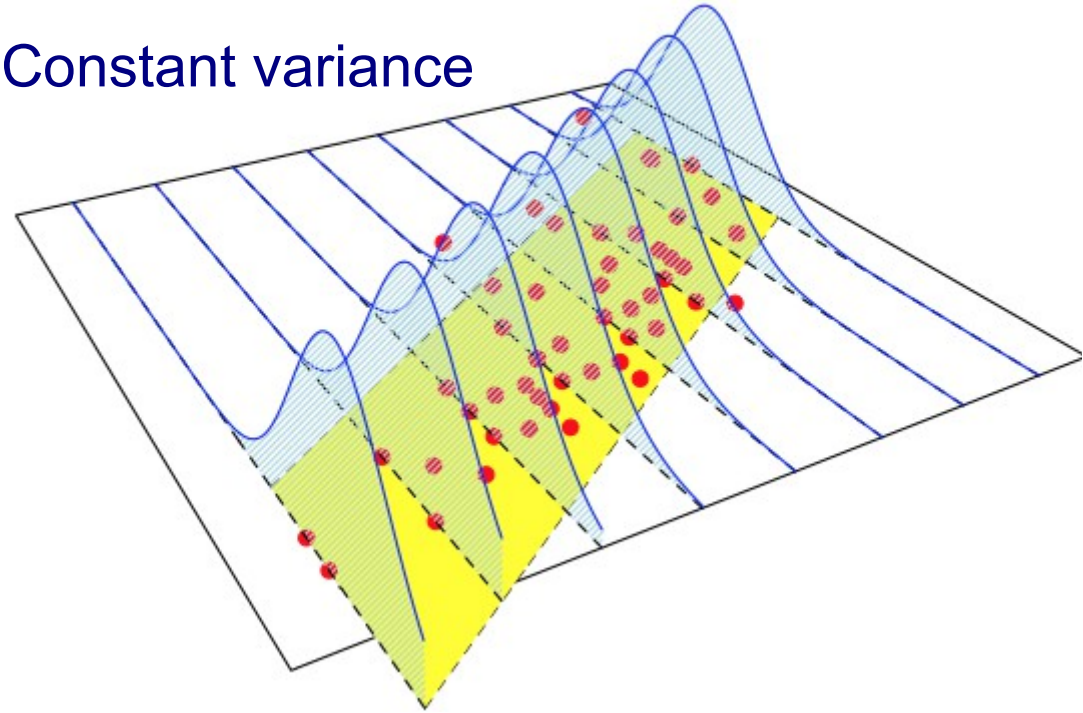
Learning targets and study questions

- Explaining and applying generalized linear models
  - When should you use a GLM?
  - Outline differences in the model structure between a simple linear model and a GLM.
  - Describe typical error distribution and link functions that you would use for modeling a) species abundances and b) fraction of surviving organisms.
- 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?

Extending the linear model: Motivation

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



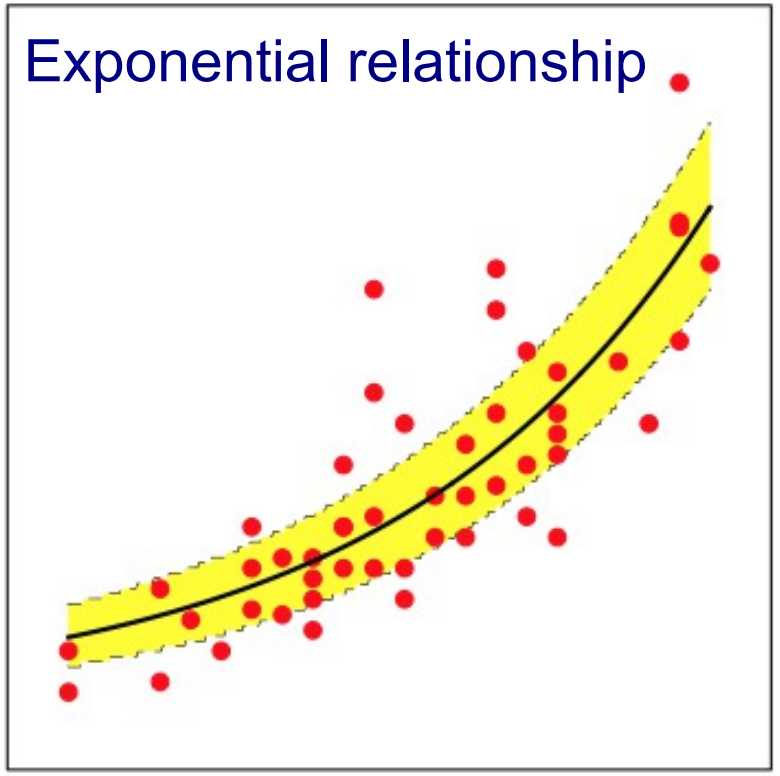


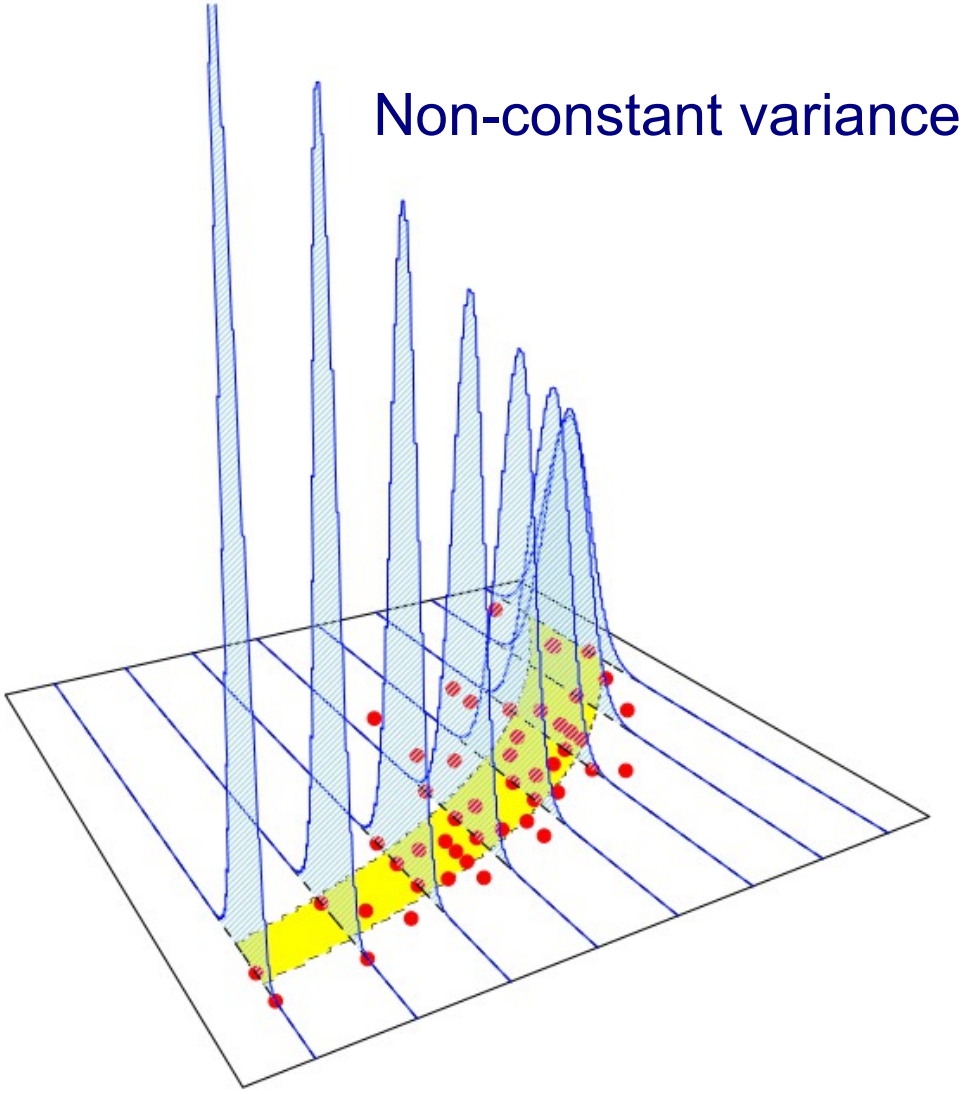
- For ecological data, the relationship with response variable is often not linear and variance not constant

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

Extending the linear model: Motivation

- Example: Accelerating loss in ecosystem functioning with increasing toxicity

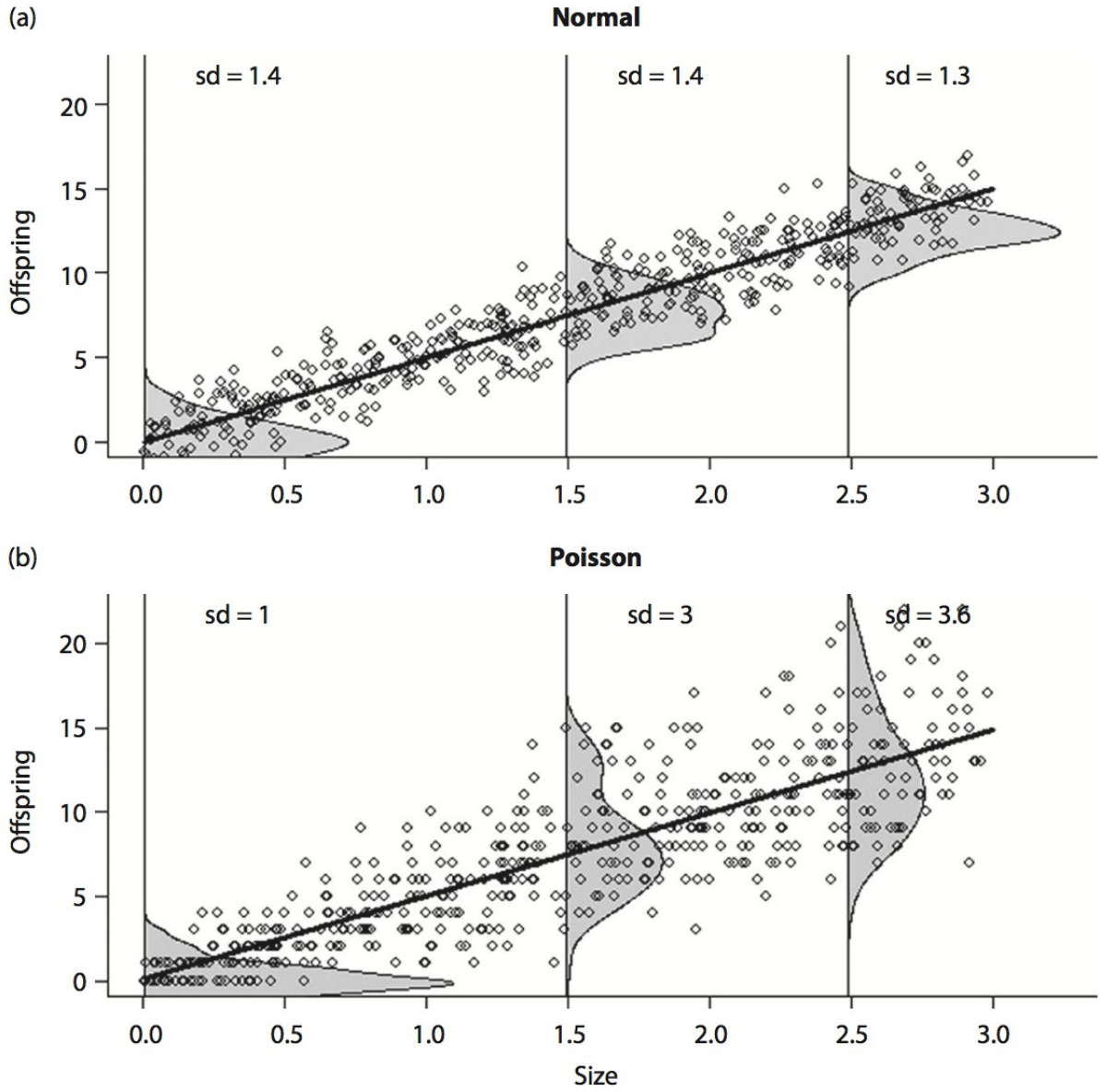




http://freakonometrics.hypotheses.org/tag/poisson

Extending the linear model: Motivation

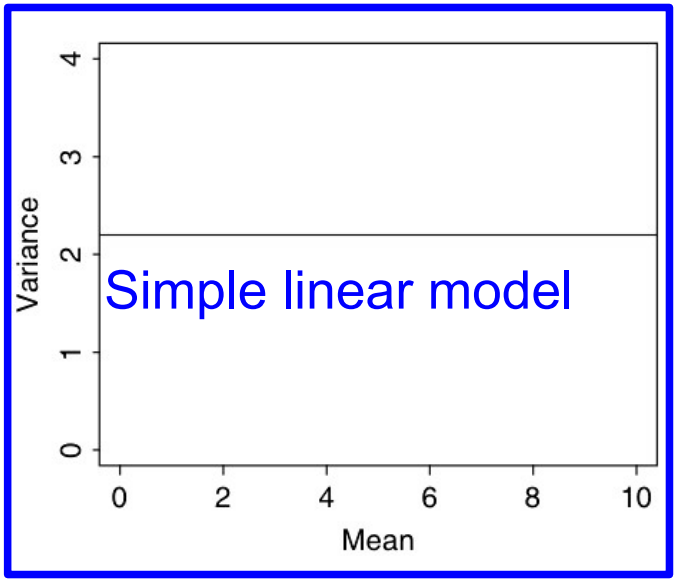
- Example: Increasing variability in number of offsprings with increasing body size of individuals

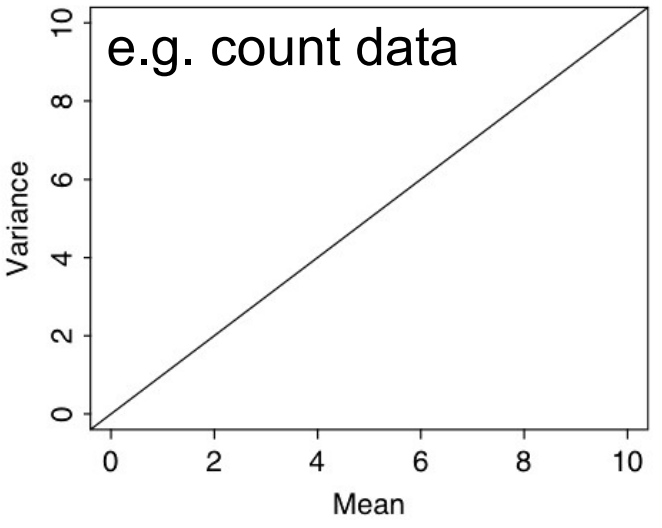


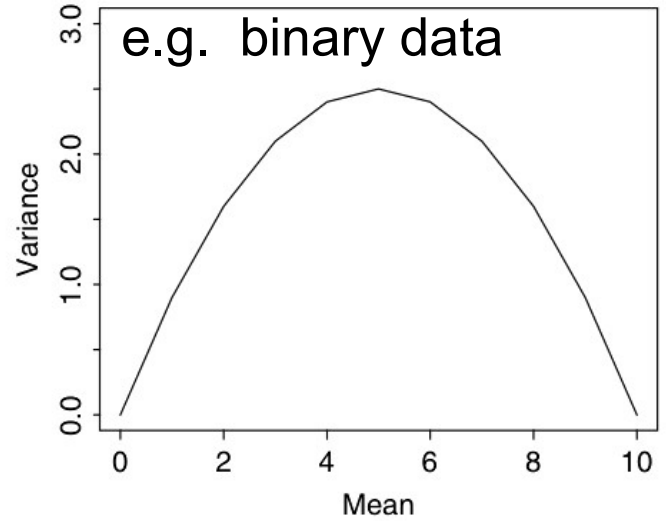
taken from Buckley 2015: 134

Modelling the mean-variance relationship

Model extension: Variance can be expressed as a function of the mean!





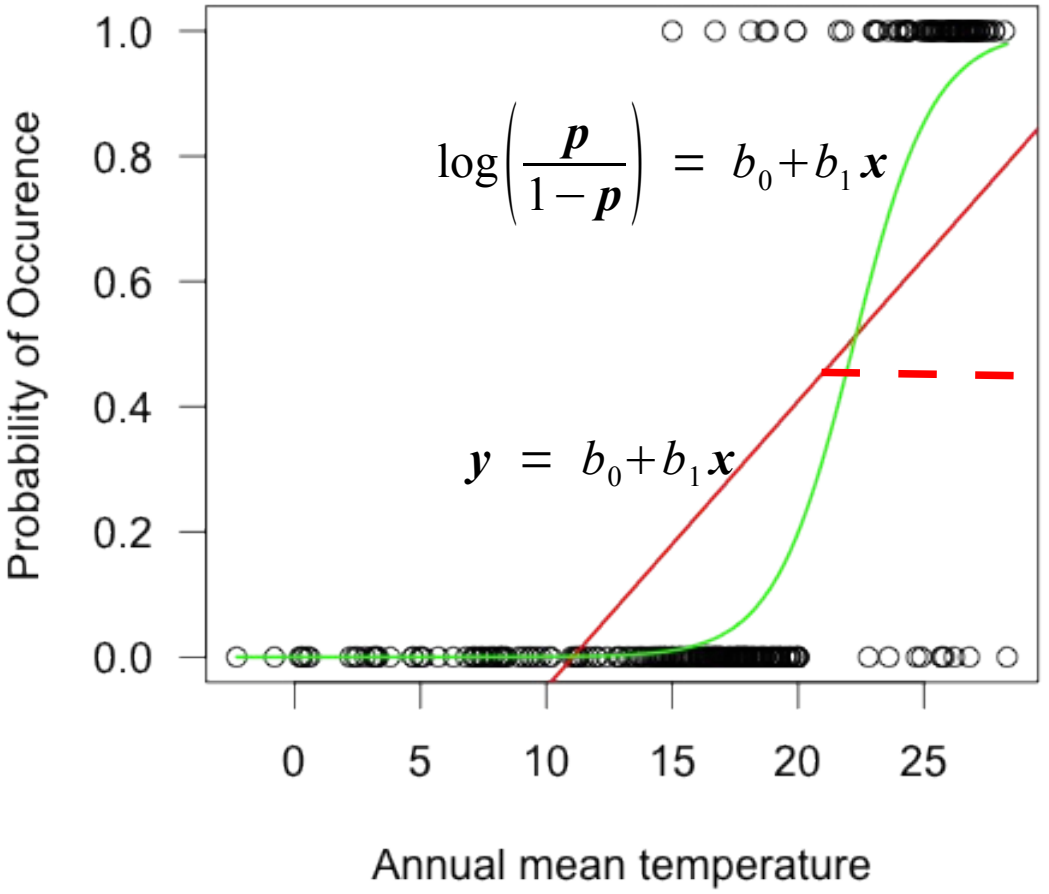


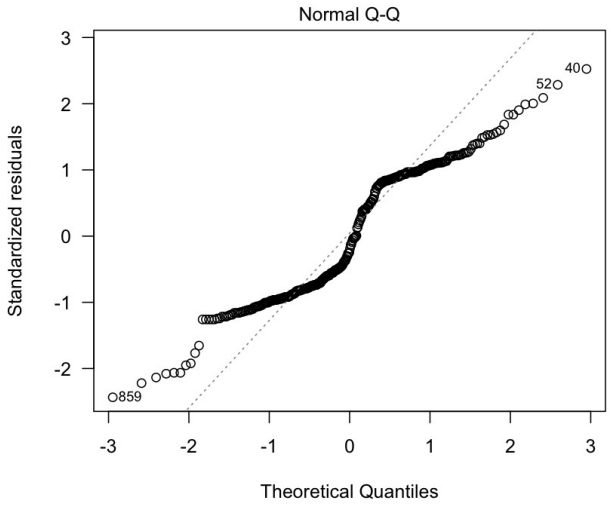
taken from
Crawley 2007: 511

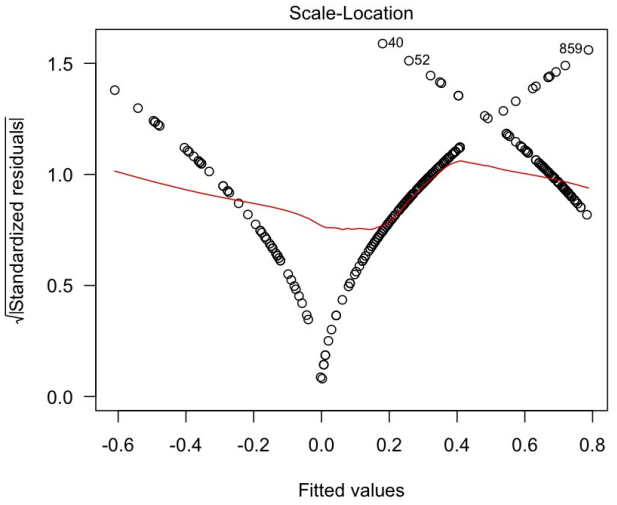
Generalised linear models (GLMs)

Models non-constant error variance by expressing the variance as a function of the mean and introduces non-normal error distribution (residuals)

Simple linear regression vs. logit model

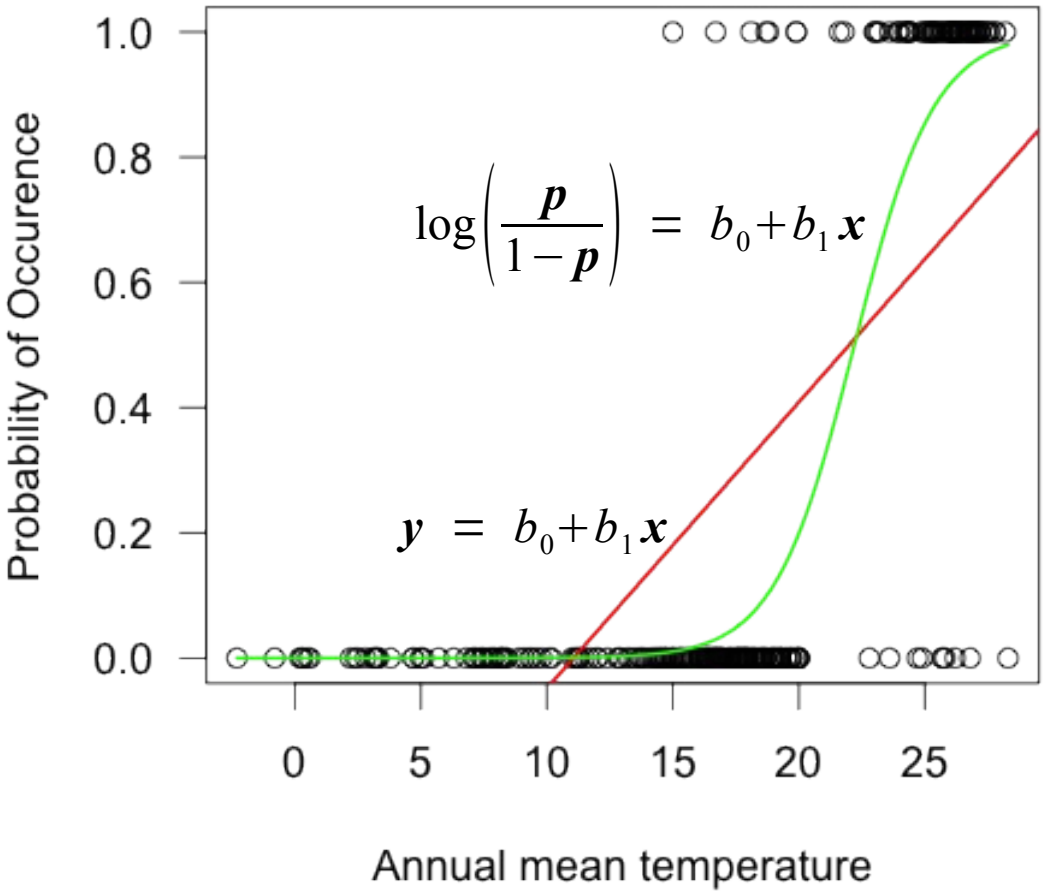






Parameters in logistic regression

Simple linear regression vs. logit model



Parameters of logit model:
b_0 = -15.9, b_1 = +0.72

Which *x* relates to p = 0.5?

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

Calculate *p* = 0.1 and *p* = 0.9

Generalized linear model

Contents

1. Learning targets and the need for GLMs
2. **Specification of the GLM**
3. Model selection and diagnostics

Comparison of LM and GLM

Simple linear model: $\boldsymbol{y} = b_0 + b_1\boldsymbol{x} + \boldsymbol{error}$

Generalised linear model:

1. Linear predictor: $\boldsymbol{\eta} = b_0 + b_1\boldsymbol{x}$
2. Link function: $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ with $E(\boldsymbol{Y}) = \boldsymbol{\mu}$
3. Error distribution of response: $\text{var}(\boldsymbol{Y}) = \phi V(\boldsymbol{\mu})$

Error distribution with related variance function and typical link function

Family (error structure)	Link	Variance function
normal	$\eta = \mu$	1
poisson	$\eta = \log\mu$	μ
binomial	$\eta = \log(\mu/(n-\mu))$	$\frac{\mu(n-\mu)}{n}$
Gamma	$\eta = \mu^{-1}$	μ^2
inverse. gaussian	$\eta = \mu^{-2}$	μ^3

modified from Crawley 2007: 511

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

taken from Buckley 2015: 137

Deviance: Goodness of fit for GLM

- GLMs minimize Deviance instead of Sum of Squares in simple linear regression model
- Deviance derived by maximum likelihood estimation (MLE)

Relation between error distribution, variance function *V(μ)* and Deviance

Family (error structure)	Deviance	Variance function
normal	$\sum (y-\hat{y})^2$	1
poisson	$2\sum y \ln(y/\mu) - (y-\mu)$	μ
binomial	$2\sum y \ln(y/\mu) + (n-y)\ln(n-y)/(n-\mu)$	$\frac{\mu(n-\mu)}{n}$
Gamma	$2\sum (y-\mu)/y - \ln(y/\mu)$	μ^2
inverse. gaussian	$\sum (y-\mu)^2/(\mu^2 y)$	μ^3

y = observations
$\hat{y}$ = fitted values for y
μ = fitted values using maximum likelihood
n = binomial denominator

taken from Crawley 2007: 516

Generalized linear model

Contents

1. Learning targets and the need for GLMs
2. Specification of the GLM
3. **Model selection and diagnostics**

Model selection for GLM

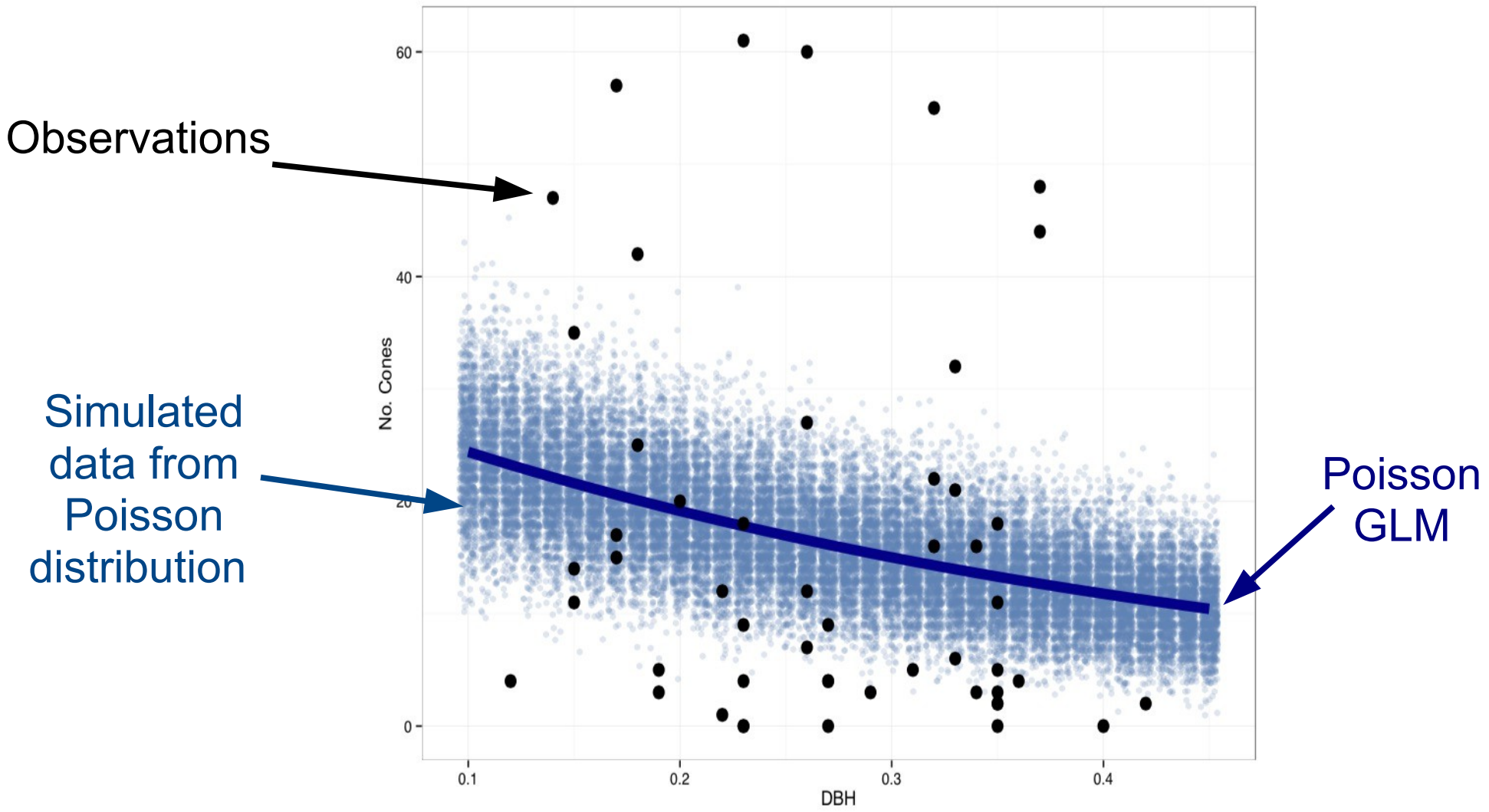
- Same methods as for multiple linear regression model
- Best subset and multi-model averaging
- Hypothesis-based stepwise model selection:
  - Wald test for individual regression coefficients
  - Log-likelihood ratio test for complete model comparison
- Information-theoretic stepwise model selection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

GLM assumptions and diagnostics

Assumptions

- Independence of observations
  - Temporal- or spatial autocorrelation: GLMMs (see Bolker 2009)
- 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 (graphical diagnostics and measures e.g. dfbetas, Cooks distance)
- Assumed mean-to-variance relationship matches data (no over- or underdispersion) (graphical diagnostics with *q*-*q* plot randomized quantile residuals and calculation of dispersion parameter)

Overdispersion



Fix: Use appropriate error distribution or quasi-likelihood estimation of mean-to-variance relation (e.g. quasibinomial)

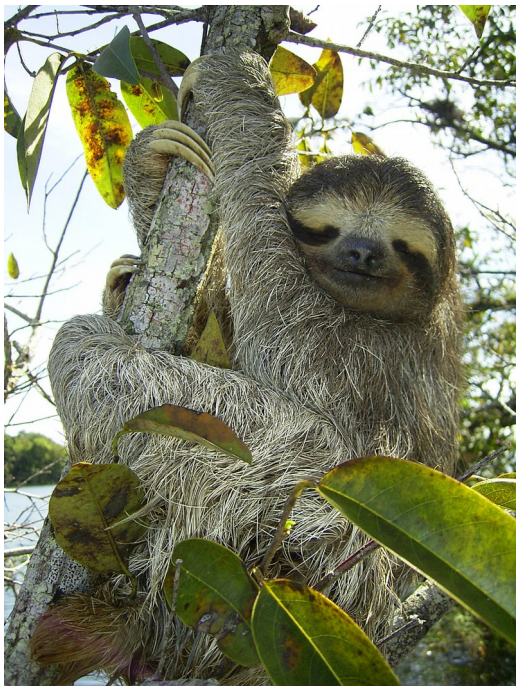
Demonstration and Exercise

For the demonstration we will work with a data set on the Southern Corroboree frog. This data is contained in the DAAG package (frogs).

Research question:

Which environmental parameters have the highest explanatory power for the occurrence of the frog?

Source: ABC Natural History Unit
http://www.abc.net.au/science/scribblygum/june2004/frog.htm



<u>Exercise:</u>

Identify the variables with the highest explanatory power for the occurrence of the *Bradypus sp*.

Generalized linear model

Contents

1

Learning targets

- Explaining and applying generalized linear models
- Describe the specifics of GLMs regarding model selection and model assumptions

2

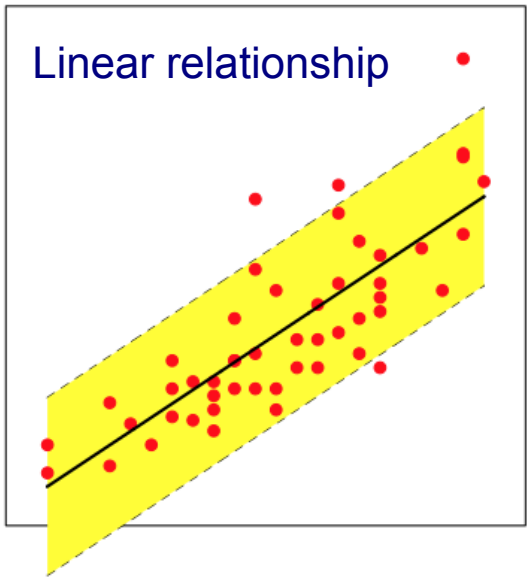
Learning targets and study questions

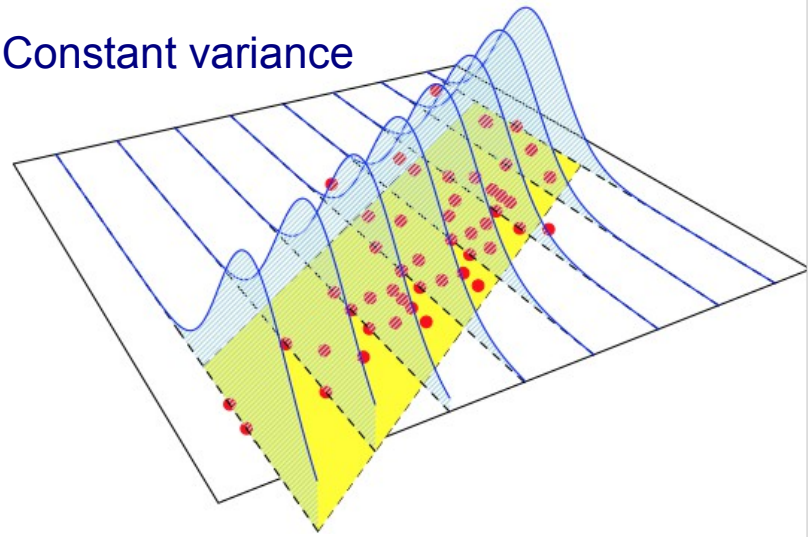
- Explaining and applying generalized linear models
  - When should you use a GLM?
  - Outline differences in the model structure between a simple linear model and a GLM.
  - Describe typical error distribution and link functions that you would use for modeling a) species abundances and b) fraction of surviving organisms.
- 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?

3

Extending the linear model: Motivation

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





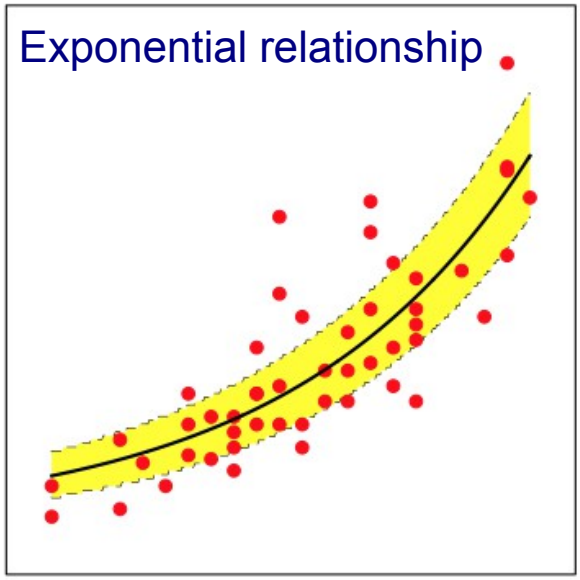
- For ecological data, the relationship with response variable is often not linear and variance not constant

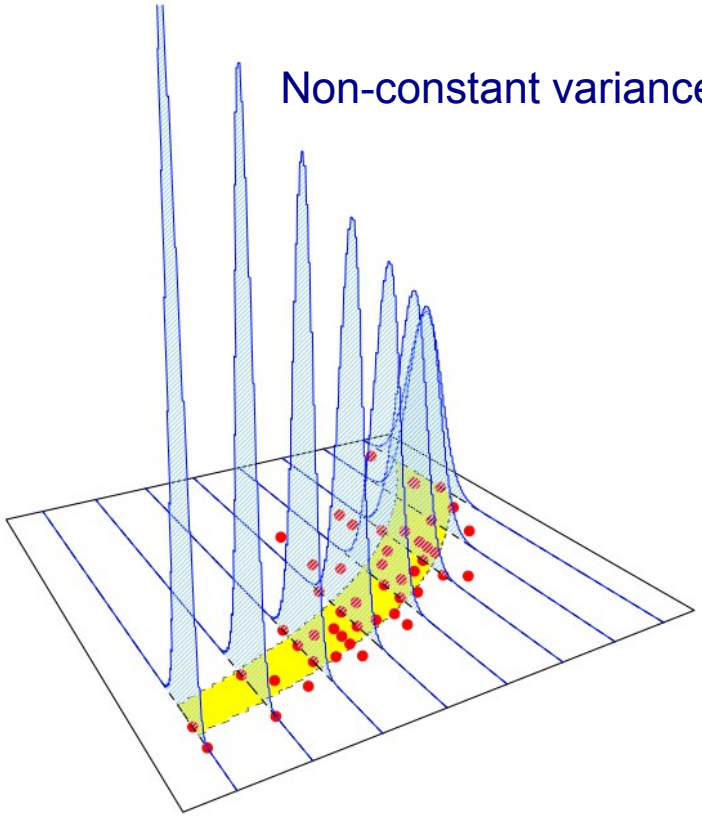
4 http://freakonometrics.hypotheses.org/tag/poisson

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

Extending the linear model: Motivation

- Example: Accelerating loss in ecosystem functioning with increasing toxicity



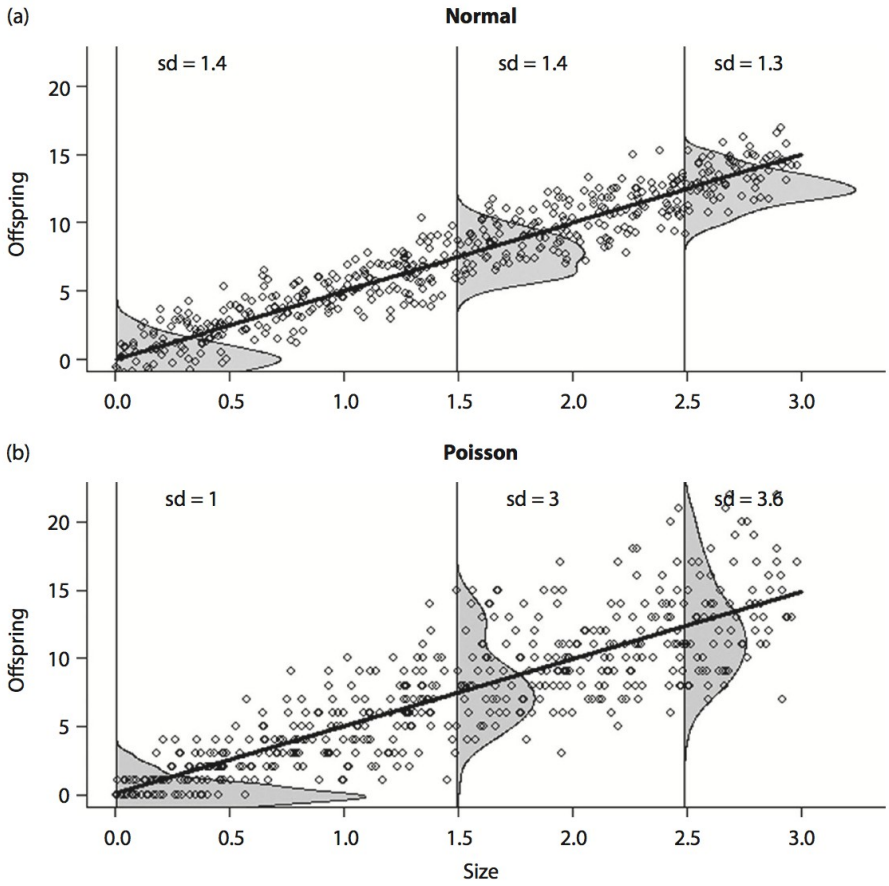


5 http://freakonometrics.hypotheses.org/tag/poisson

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

Extending the linear model: Motivation

- <u>Example:</u> Increasing variability in number of offsprings with increasing body size of individuals



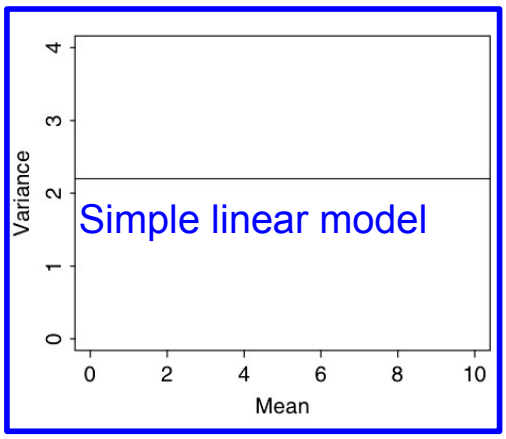
taken from Buckley 2015: 134

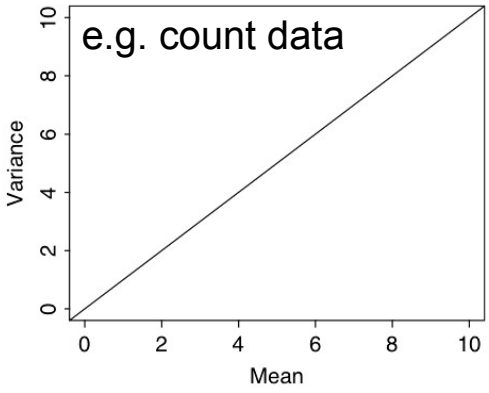
Buckley, Y.M. (2015): Generalized linear models in: Fox G.A., Negrete-Yankelevich S. & Sosa V.J. Eds: Ecological statistics: contemporary theory and application. Oxford University Press, Oxford. p. 132-148

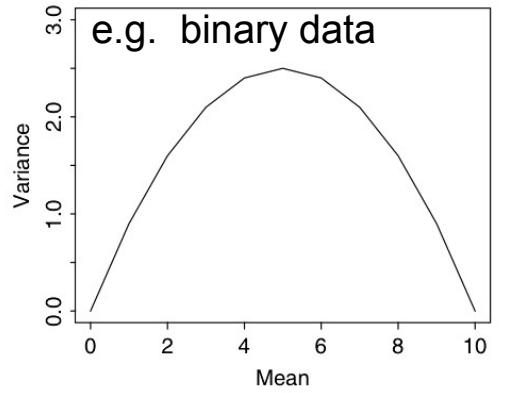
6

Modelling the mean-variance relationship

Model extension: Variance can be expressed as a function of the mean!







taken from Crawley 2007: 511

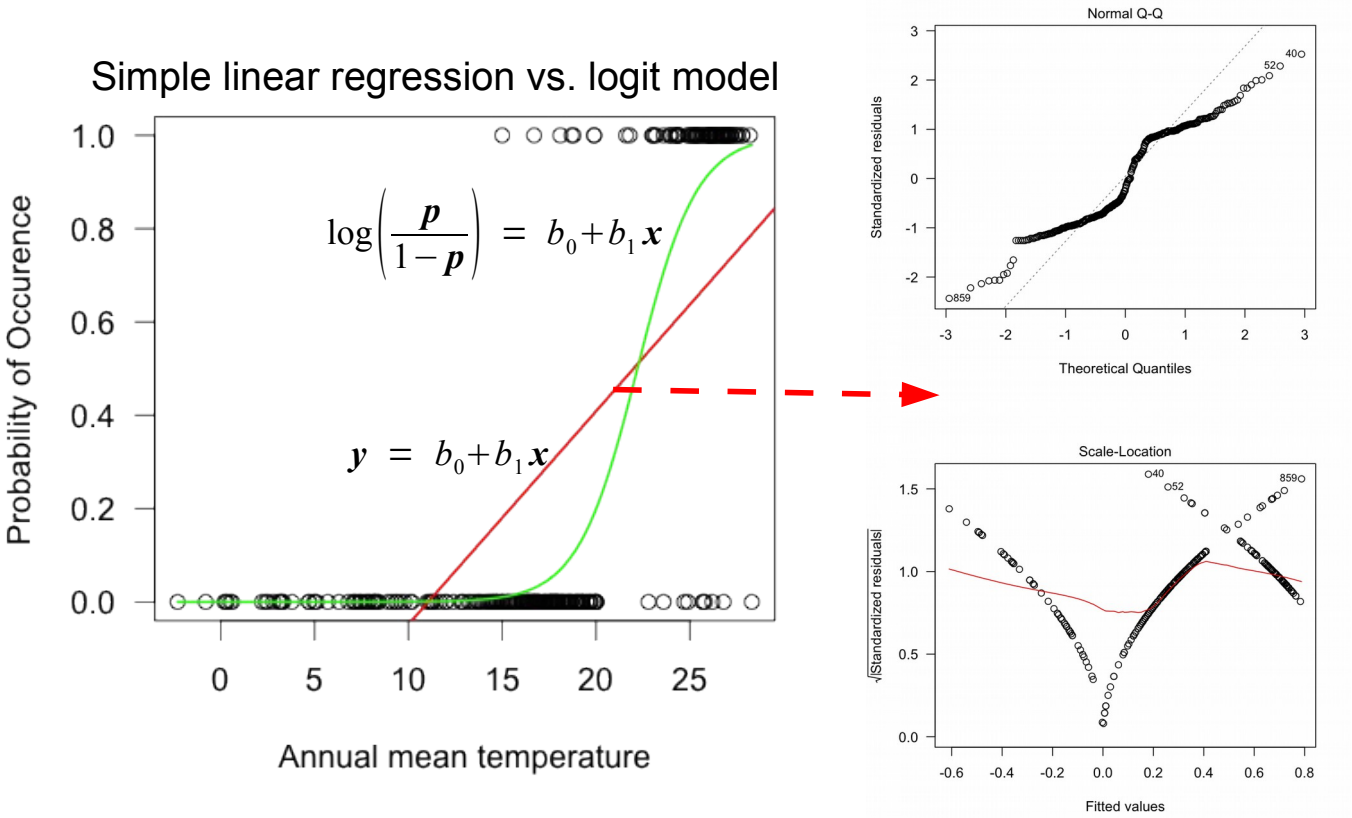
7

Examples for binary data include species presence-absence data or data from ecotoxicological experiments, where for each individual the response can be dead or alive.

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

Generalised linear models (GLMs)

Models non-constant error variance by expressing the variance as a function of the mean and introduces non-normal error distribution (residuals)



Non-normal error distribution indicates that the relationship between the explanatory variable(s) and the response variable is non-linear.

A useful tool to inspect the consequences of data from different distributions for model diagnosis is the GLM explorer: http://139.14.20.252:3838/teaching/glm_explorer/

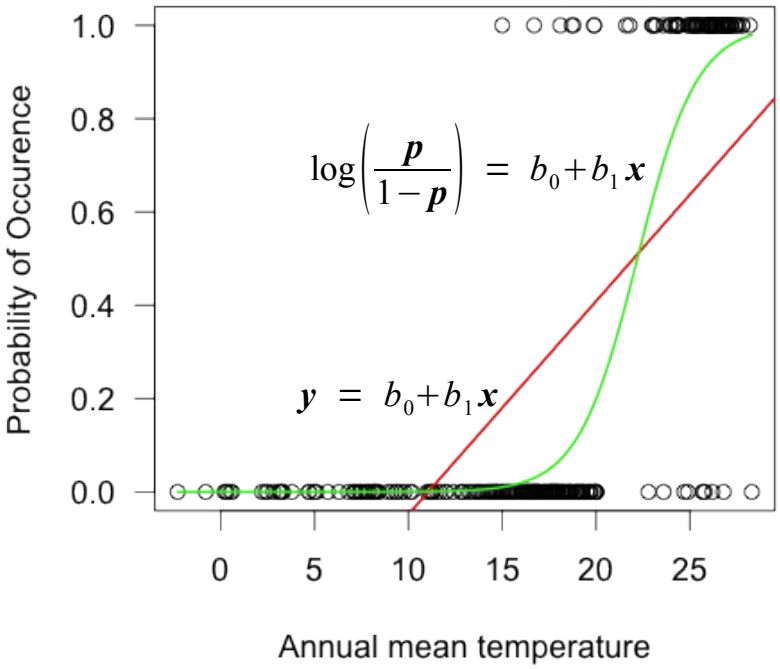
Fox (2015) gives a thorough introduction to GLMs and Fox & Weisberg (2011) features the implementation in R. In addition, Faraway (2006) gives a very readable introduction into linear models including GLMs and the example in the figure is adapted from this book (p.26).

Faraway J.J. (2006) Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. Chapman & Hall/CRC, Boca Raton, Fla. 1st edition.

Note that an updated edition has been published.

Parameters in logistic regression

Simple linear regression vs. logit model



Parameters of logit model:
b_0 = -15.9, b_1 = +0.72

Which *x* relates to p = 0.5?

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

Calculate *p* = 0.1 and *p* = 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 log normal model means that the left hand side of the equation for a given *p* can be multiplied by -1 to solve the equation for 1-*p*.

Note that *p* = 0.5 would represent the *LC50* (i.e. the median lethal effect concentration) in an ecotoxicological context.

Note also that *p* = 1 results in infinity.

Generalized linear model

Contents

1. Learning targets and the need for GLMs
2. **Specification of the GLM**
3. Model selection and diagnostics

10

Comparison of LM and GLM

Simple linear model: $\boldsymbol{y} = b_0 + b_1\boldsymbol{x} + \boldsymbol{error}$

Generalised linear model:

1. Linear predictor: $\boldsymbol{\eta} = b_0 + b_1\boldsymbol{x}$
2. Link function: $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ with $E(\boldsymbol{Y}) = \boldsymbol{\mu}$
3. Error distribution of response: $\text{var}(\boldsymbol{Y}) = \phi V(\boldsymbol{\mu})$

Error distribution with related variance function and typical link function

Family (error structure)	Link	Variance function
normal	$\eta = \mu$	1
poisson	$\eta = \log\mu$	μ
binomial	$\eta = \log(\mu/(n-\mu))$	$\frac{\mu(n-\mu)}{n}$
Gamma	$\eta = \mu^{-1}$	μ^2
inverse. gaussian	$\eta = \mu^{-2}$	μ^3

modified from Crawley 2007: 511

$E(\boldsymbol{Y})$ is the expected value, i.e. the mean. ϕ is a constant dispersion parameter. For the simple linear regression model, $g(\boldsymbol{\mu}) = \boldsymbol{\mu}$ and $var(\boldsymbol{Y}) = \phi$

The table gives the error distribution with the related variance function and a typical link function. However, alternative link functions could be used, which is elaborated later. For example, in ecological studies the gamma distribution is often used with the log link.

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): 137ff and the following papers for details:

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.

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

12

taken from Buckley 2015: 137

The negative binomial distribution represents an important distribution that in several cases fits ecological count data better than the poisson distribution given that it extends the poisson distribution with an additional parameter. However, see also this slide, if the poisson distribution does not match (is overdispersed, which is explained later).

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

Deviance: Goodness of fit for GLM

- GLMs minimize Deviance instead of Sum of Squares in simple linear regression model
- Deviance derived by maximum likelihood estimation (MLE)

Relation between error distribution, variance function *V(μ)* and Deviance

Family (error structure)	Deviance	Variance function
normal	$\sum(y-\hat{y})^2$	1
poisson	$2\sum y\ln(y/\mu)-(y-\mu)$	μ
binomial	$2\sum y\ln(y/\mu)+(n-y)\ln(n-y)/(n-\mu)$	$\frac{\mu(n-\mu)}{n}$
Gamma	$2\sum(y-\mu)/y-\ln(y/\mu)$	μ^2
inverse. gaussian	$\sum(y-\mu)^2/(\mu^2 y)$	μ^3

y = observations
$\hat{y}$ = fitted values for y
μ = fitted values using maximum likelihood
n = binomial denominator

taken from Crawley 2007: 516

13

MLE estimates the model parameters conditional on the sample data, i.e. the parameters that have the highest likelihood to generate the observed data. Practically, the fitting is achieved by an iterative re-weighted least squares (IRWLS) algorithm. For details on the mathematical background of GLMs see Fox (2008): 402f or Dunteman & Ho. 2006: An introduction to generalized linear models. Sage Publications. 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).

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

Generalized linear model

Contents

14

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 used if the true value is very far from or close to 0. For example, if 10/20 larvae died in the control but 5/5 pupae died, the larvae to pupae odds ratio would be infinite (see logistic regression slide $p = 1$). The Log-likelihood ratio test represents a more robust alternative.

GLM assumptions and diagnostics

Assumptions

- Independence of observations
  - Temporal- or spatial autocorrelation: GLMMs (see Bolker 2009)
- 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 (graphical diagnostics and measures e.g. dfbetas, Cooks distance)
- Assumed mean-to-variance relationship matches data (no over- or underdispersion) (graphical diagnostics with *q*-*q* plot randomized quantile residuals and calculation of dispersion parameter)

Dfbetas describe the standardized change in the betas (regression coefficients) when refitting the model *n* times, while omitting each of the *n* observations once from the model).

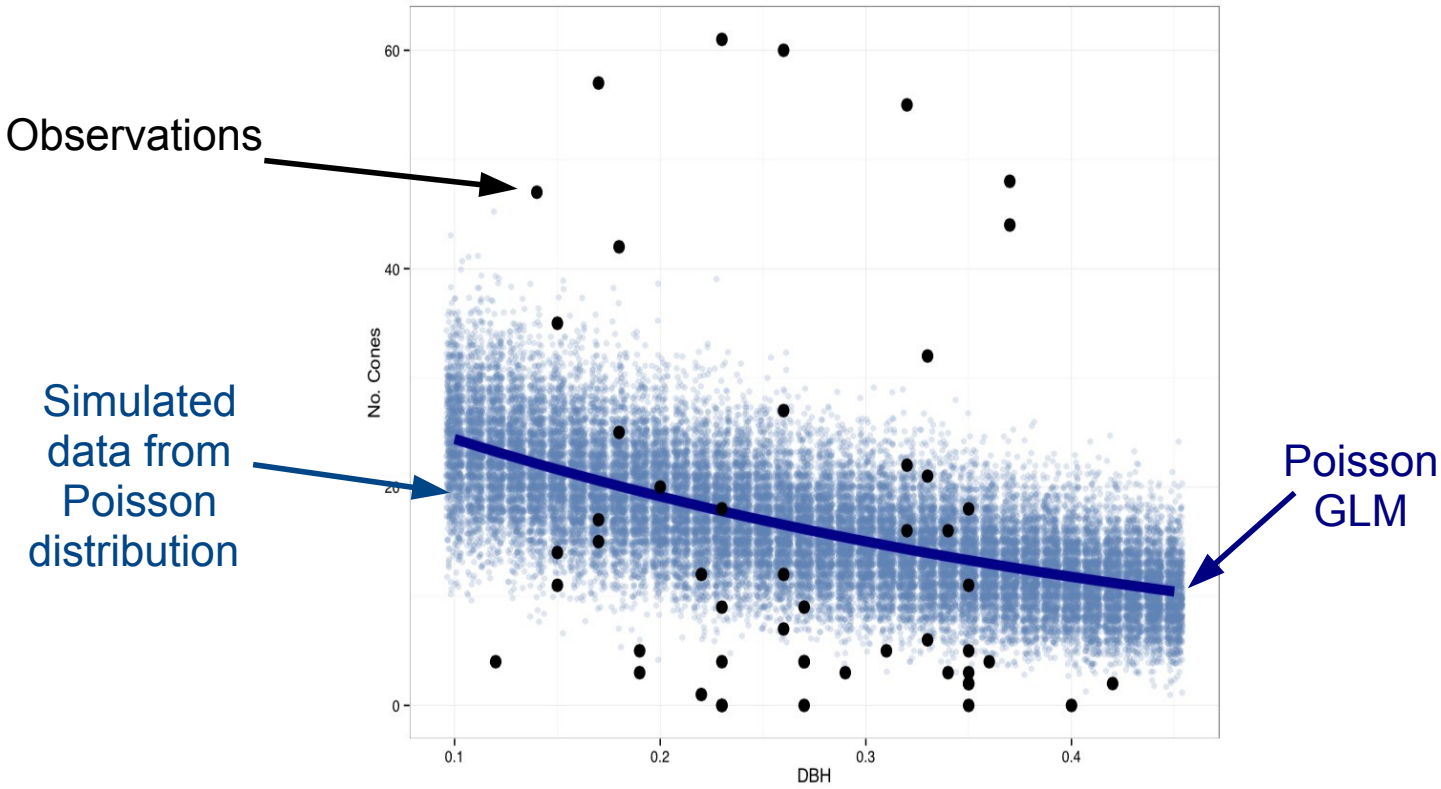
Overdispersion means that the mean-to-variance relationship in the model is higher than assumed (and the opposite applies to underdispersion). As a rule of thumb, over- or underdispersion are indicated when the dispersion parameter approaches or exceeds 2 or 0.5, respectively (Logan 2010: 493). The dispersion parameter is calculated as the residual deviance divided by the degrees of freedom. Graphical methods are discussed in the R demonstration.

Rootograms can be used to compare the fit of models such as the poisson model to other models (e.g. the negative binomial model). Details and R code are provided in: Kleiber C. & Zeileis A. (2016) Visualizing Count Data Regressions Using Rootograms. The American Statistician 70, 296–303

Cross-validation should be applied as explained for the quantification of prediction accuracy that has been discussed for the linear model.

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



Fix: Use appropriate error distribution or quasi-likelihood estimation of mean-to-variance relation (e.g. quasibinomial)

17

Note that for beginners over- or underdispersion can be due to the selection of an incorrect error distribution or link function. Thus, their plausibility should be checked first, before estimating the mean-to-variance relationship (e.g. quasipoisson).

In case of overdispersion of the Poisson GLM, the quasipoisson and the negative binomial model represent alternatives. How to decide between the two is discussed in: 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.

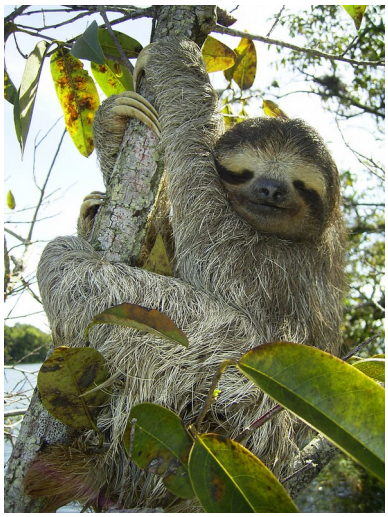
Demonstration and Exercise

For the demonstration we will work with a data set on the Southern Corroboree frog. This data is contained in the DAAG package (frogs).

Research question:

Which environmental parameters have the highest explanatory power for the occurrence of the frog?

Source: ABC Natural History Unit
http://www.abc.net.au/science/scribblygum/june2004/frog.htm



Exercise:

Identify the variables with the highest explanatory power for the occurrence of the *Bradypus sp*.

18