Generalized linear model

Contents

Learning targets

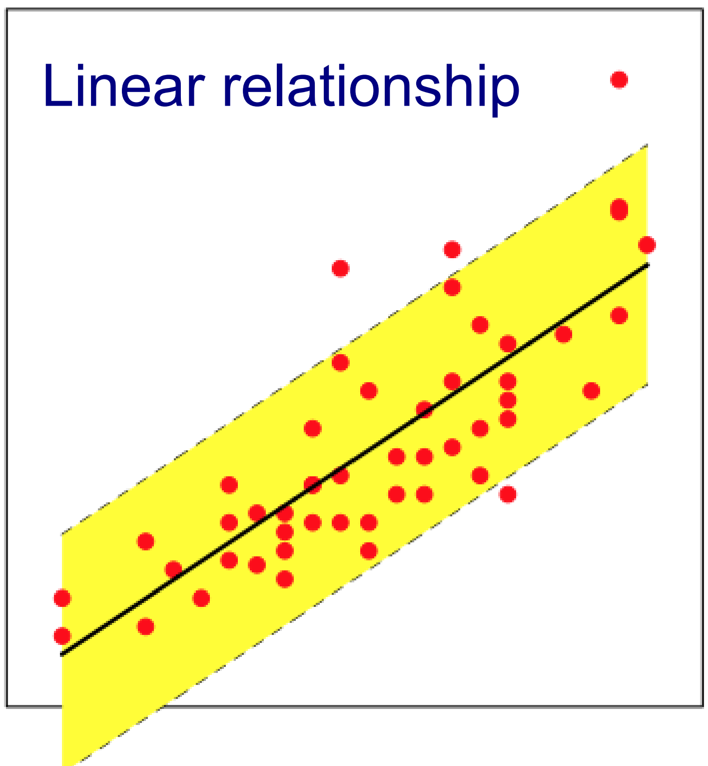
- Explaining and applying generalized linear models
- Describe the specifics of GLMs regarding model selection and model assumptions

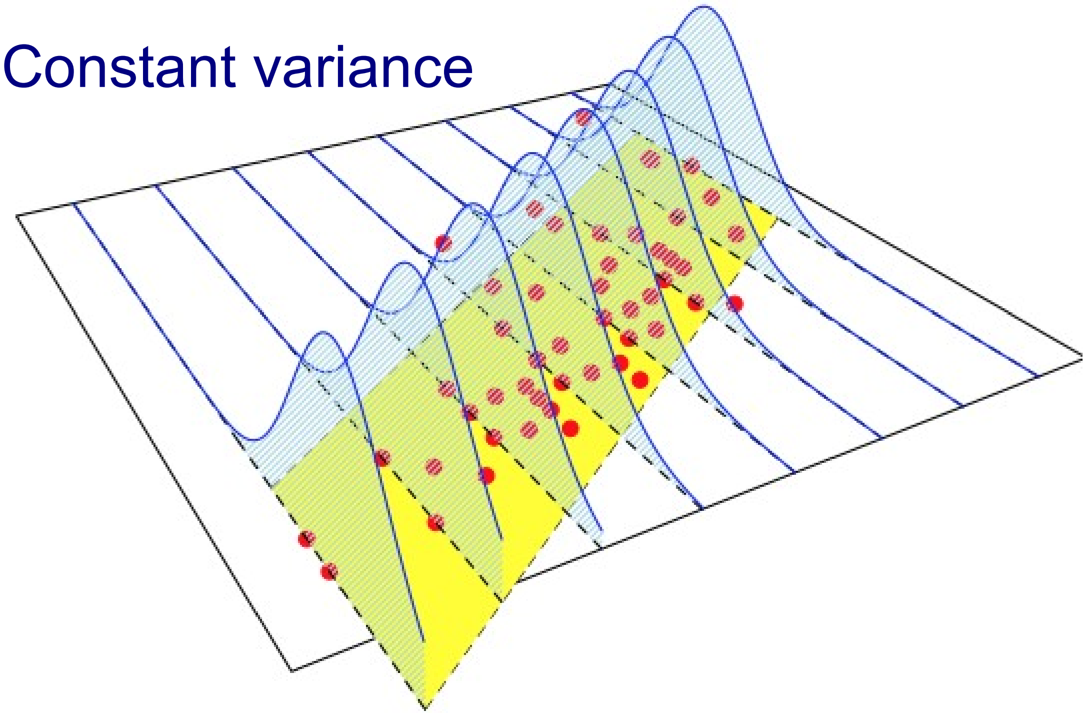
Learning targets and study questions

- 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



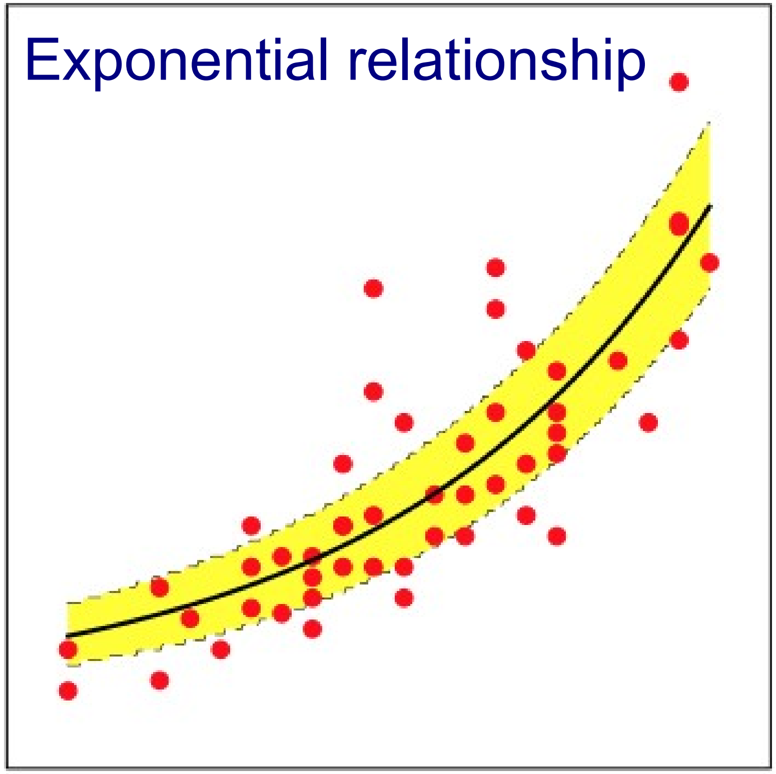


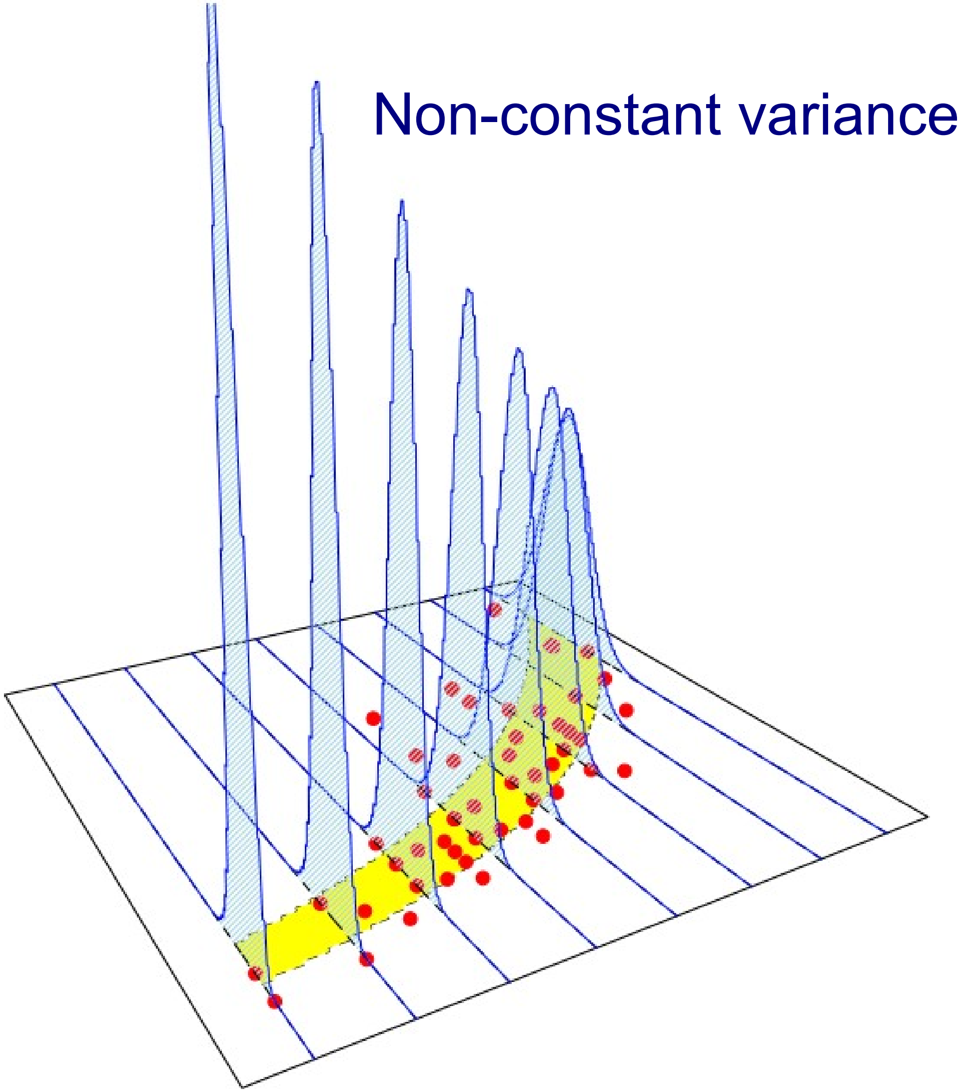
- For ecological data, the relationship with response variable is often not linear and variance not constant

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

Extending the linear model: Motivation

- Example: Accelerating loss in ecosystem functioning with increasing toxicity

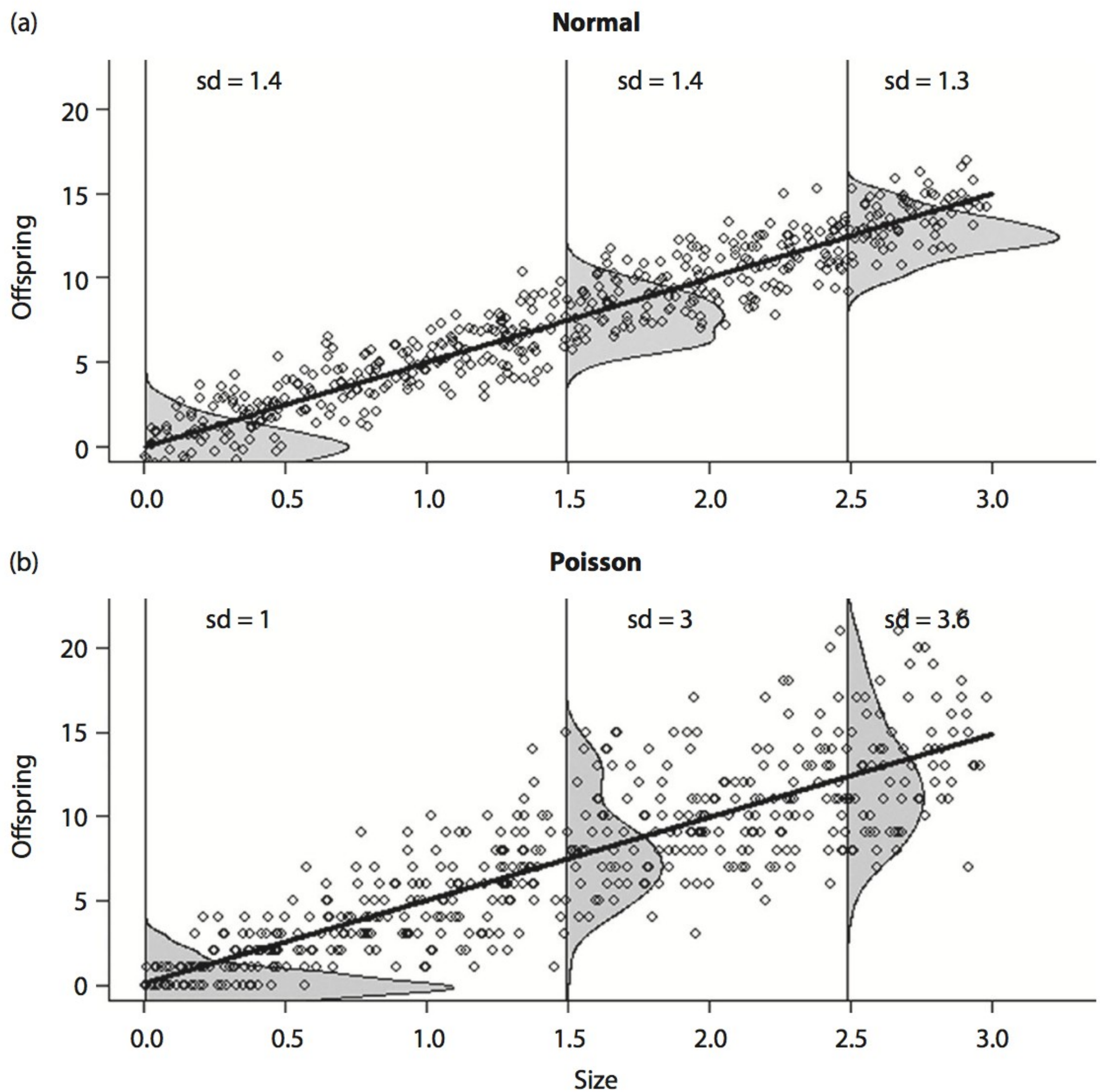




http://freakonometrics.hypotheses.org/tag/poisson

Extending the linear model: Motivation

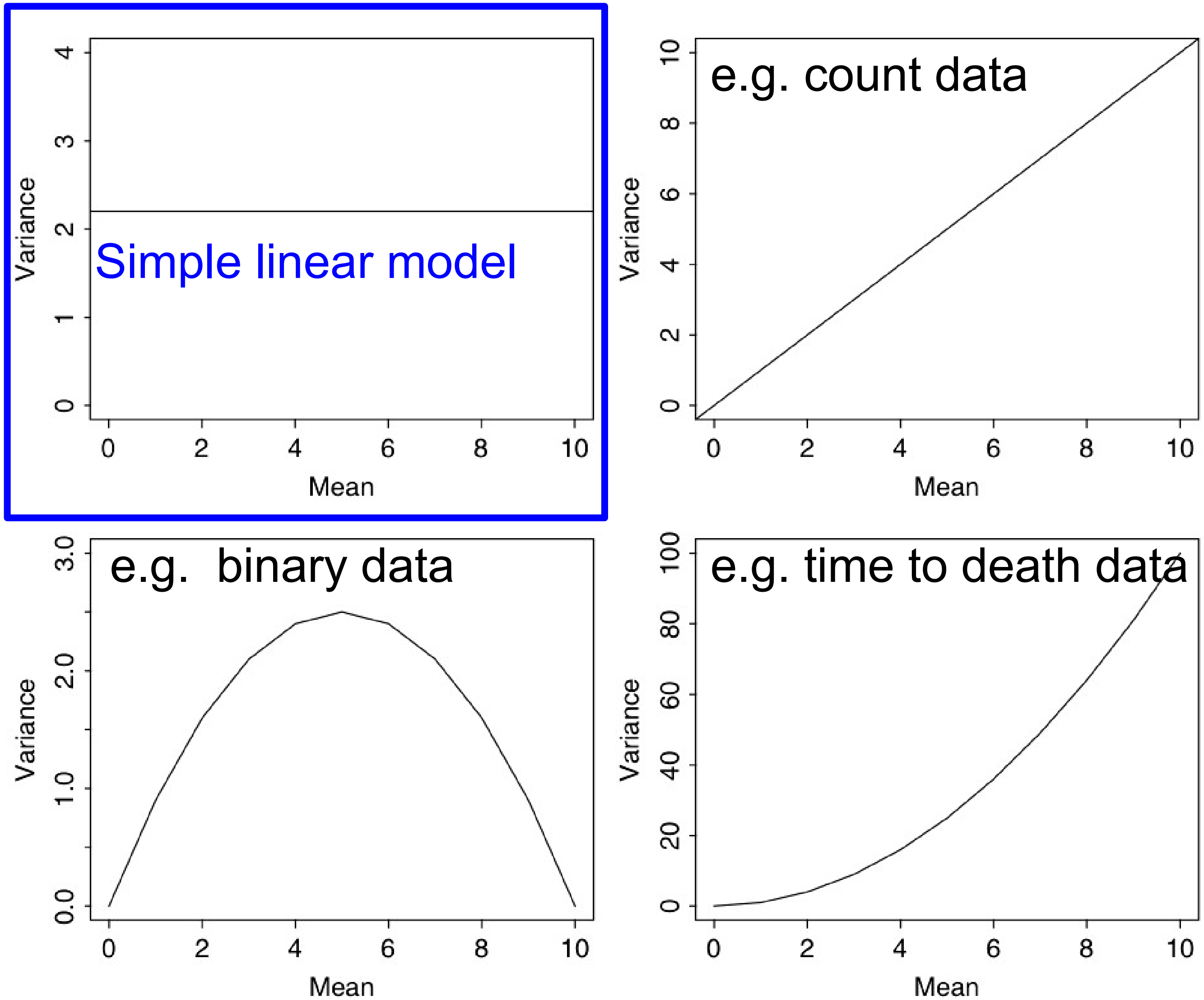
- <u>Example:</u> Increasing variability in number of offsprings with increasing body size of individuals



taken from Buckley 2015: 134

Modelling the mean-variance relationship

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

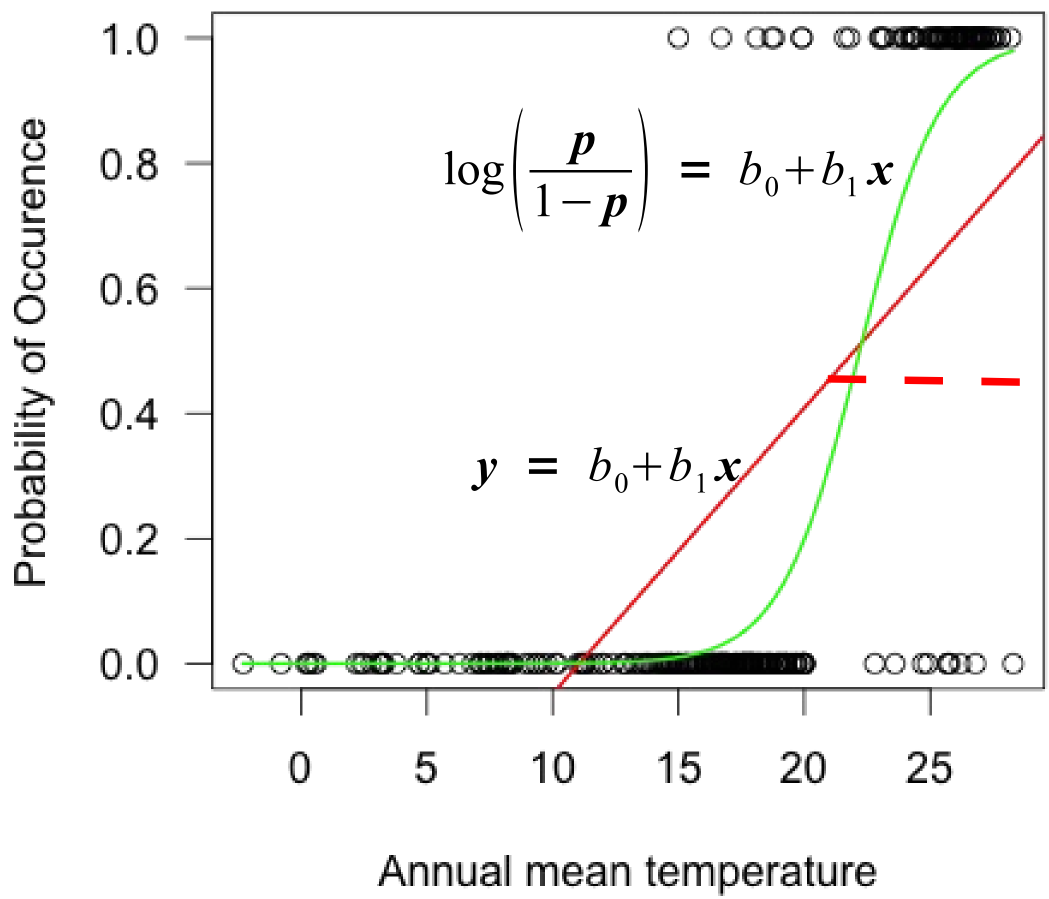


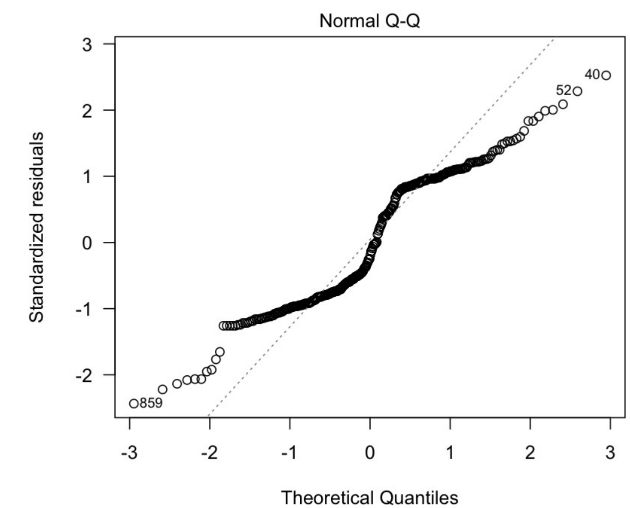
taken from
Crawley 2007: 511

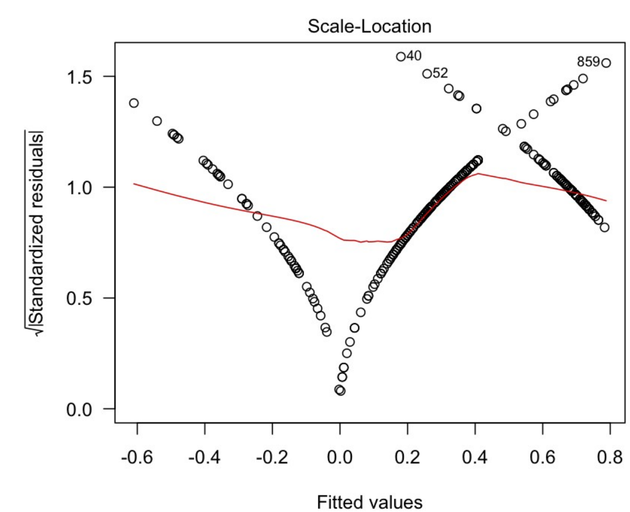
Generalised linear models (GLMs)

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

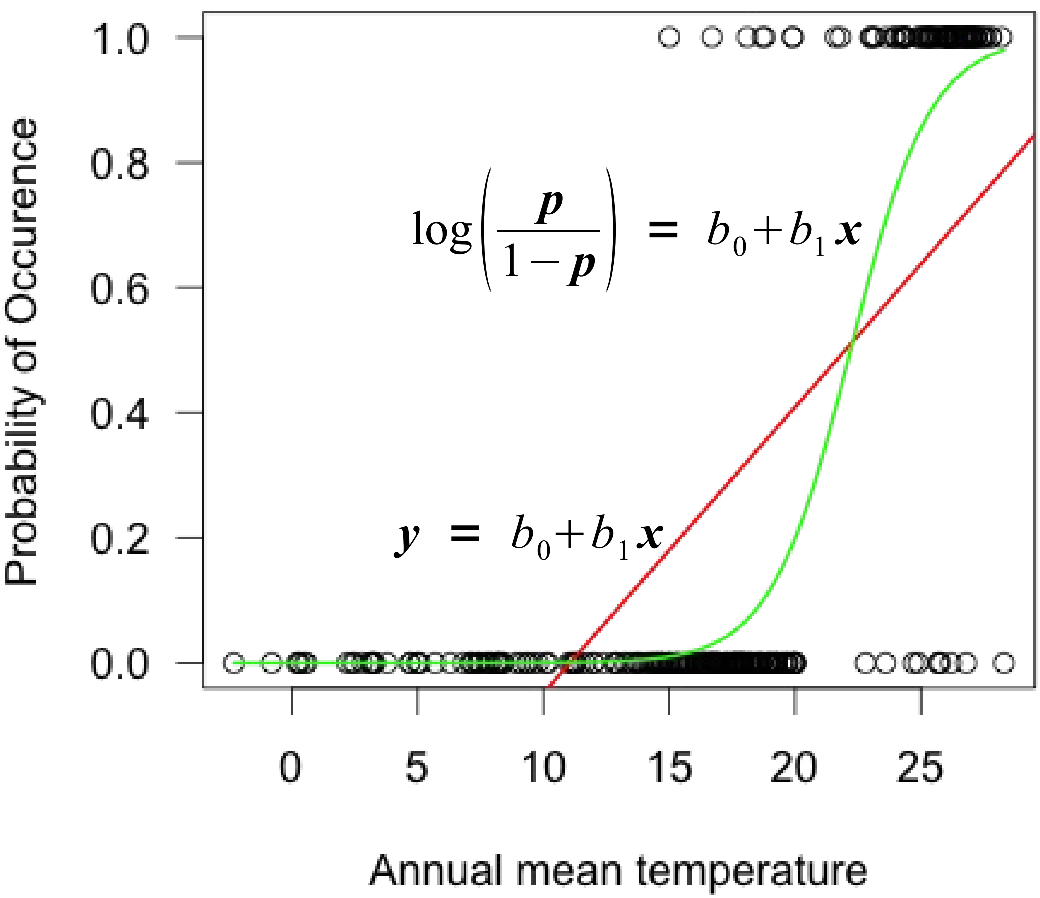






Parameters in logistic regression

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?

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

Calculate p = 0.1 and p = 0.9

Generalized linear model

Contents

Comparison of LM and GLM

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

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: $\mathrm{var}(\boldsymbol{Y}) = \phi V(\boldsymbol{\mu})$

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

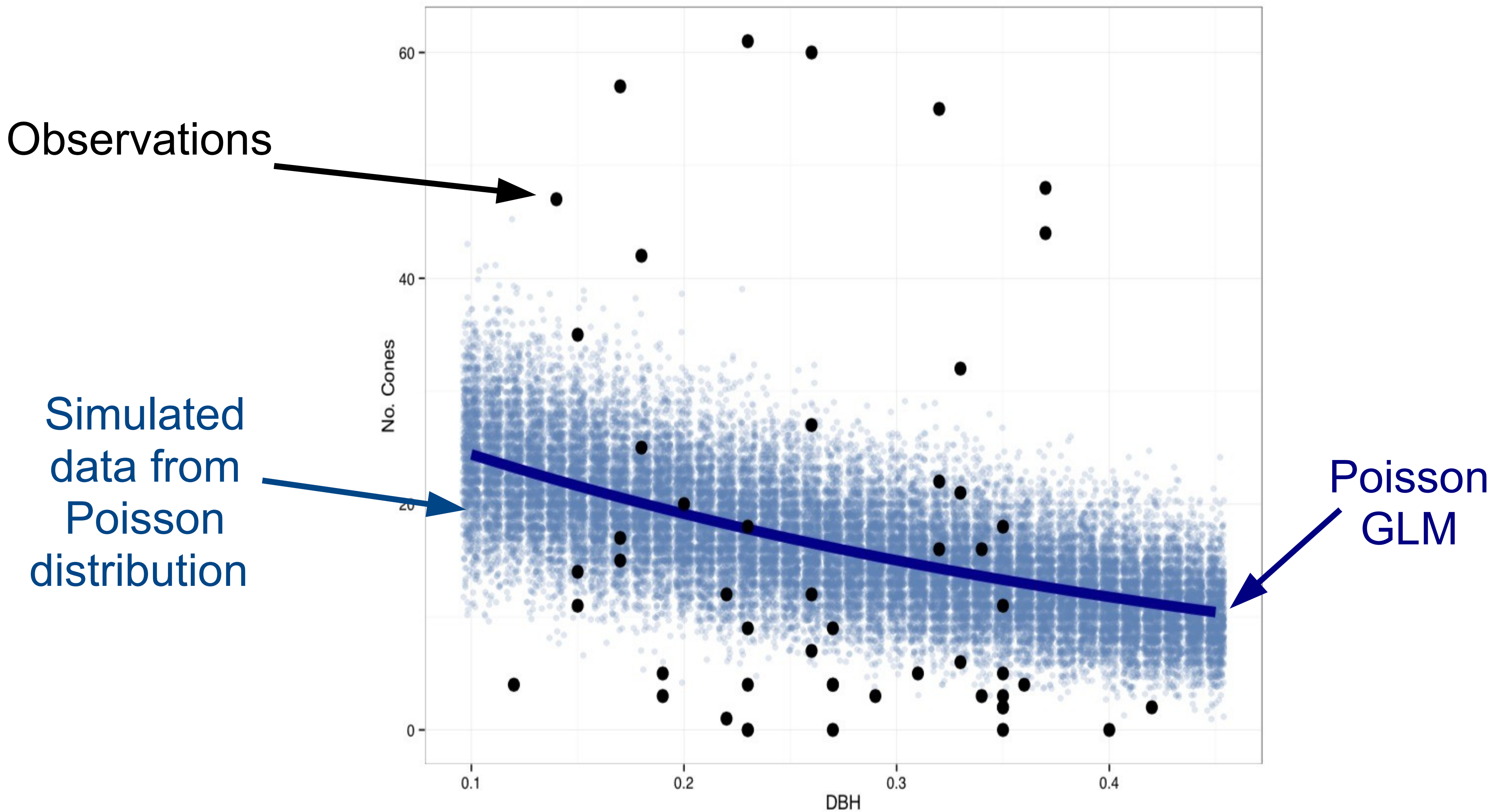
- Same methods as for multiple linear regression model
- Best subset and multi-model averaging
- Hypothesis-based stepwise model selection:
  - Wald test for individual regression coefficients
  - Log-likelihood ratio test for complete model comparison
- Information-theoretic stepwise model selection (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*.