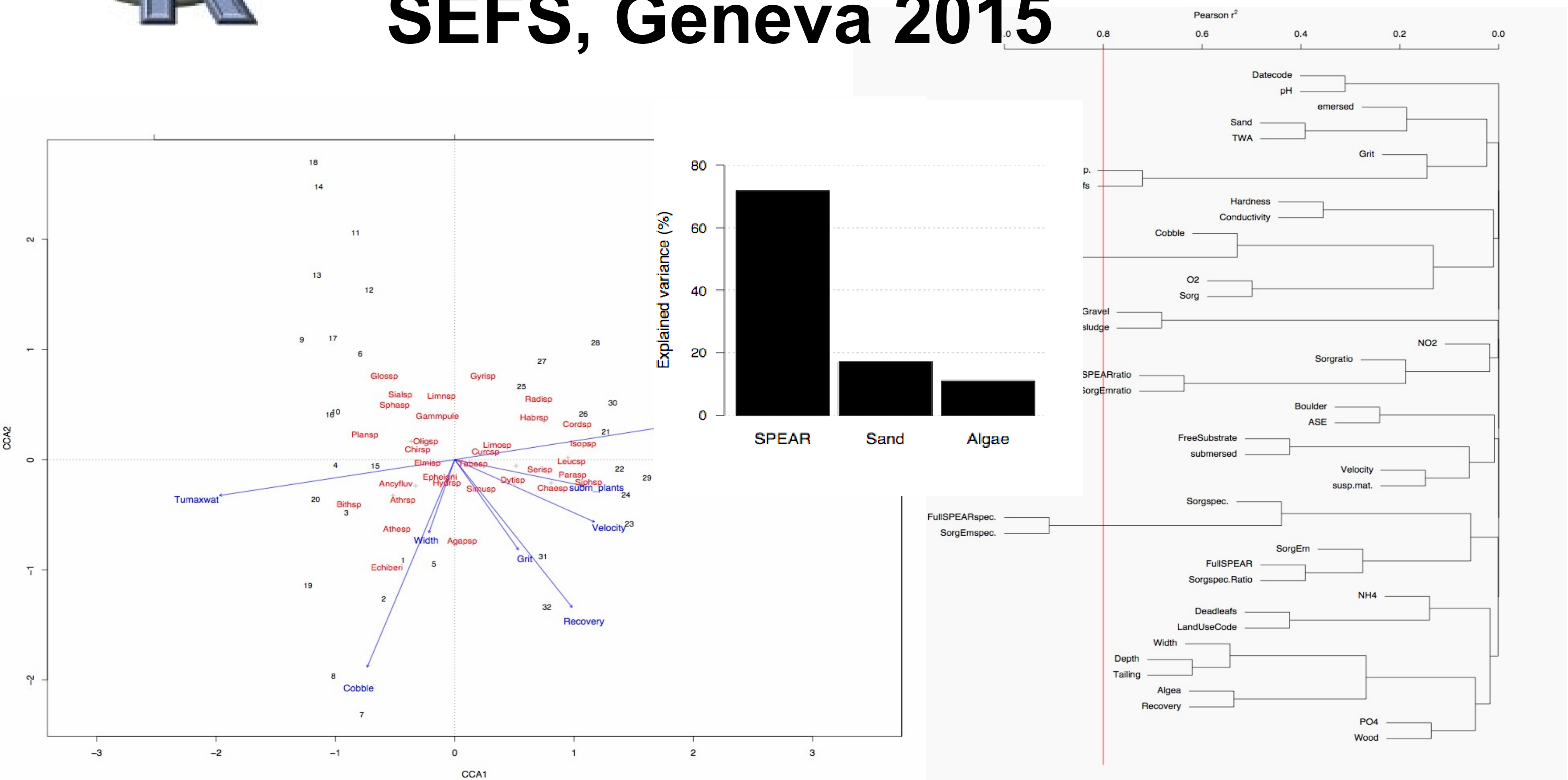
Data analysis in freshwater ecology using R

SEFS, Geneva 2015



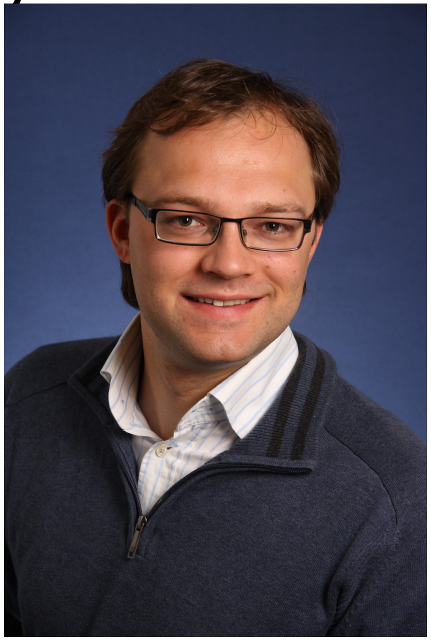
Ralf B. Schäfer, Eduard Szöcs, Avit Bhowmik

Short intro: Ralf Schäfer

- Assistant Professor for Quantitative Landscape Ecology
- Phd @ UFZ, Leipzig; Postdoc @ RMIT, Australia
- Teaching:
  - Statistics
  - GIS
  - Modelling
  - Aquatic Ecotoxicology
- Research:
  - Effects of toxicants on structure and functions
  - Modelling (Spatial, Statistical, Traits)
  - Trophic linkages between aquatic & terrestrial systems

www.landscapecology.uni-landau.de @LandscapEcology

Short intro: Eduard Szöcs



- PhD student Quantitative Landscape Ecology
- Environmental Sciences + Ecotoxicology
- Teaching:
  - Statistics
- Research:
  - Statistical Ecology - Eco(toxico)logical Statistics
  - Effects and distribution of pesticides in freshwaters
- R programming:
  - Author/Co-Author of 3 CRAN packages (taxize, webchem, rspear)
  - Other packages on github (restax, esmisc)
  - Minor contribitions to other pkgs (e.g. vegan)

edild.github.io

@EduardSzoecs

Short intro: Avit Kumar Bhowmik

- PhD student, Quantitative Landscape Ecology
- M.Sc. Geo.Tech. @ Erasmus Mundus
- Teaching:
  - GIS
  - Spatial and Geo Statistics
- Research:
  - Spatial Ecology
  - Climate
- Tools and software:
  - ATRIC: Stream threshold selection and riparian corridor delineation
  - SSTP: Spatially shifting temporal points

www.avitbhowmik.webs.com @LandscapEcology

Course Organisation

9:00-9:15 Short intro & course organisation, Software preparation

9:15-11:00 Linear and Generalised linear model

11:15-12:15 continued; Ordination (I)

13:15-14:45 Ordination (II)

15:00-16:45 Spatial autocorrelation in linear models

16:45-17:00 Course evaluation

Course material:
https://github.com/EDiLD/sefs9_Rworkshop

Course structure: intro – demo – hands on exercises

Block I

Linear and Generalised Linear model

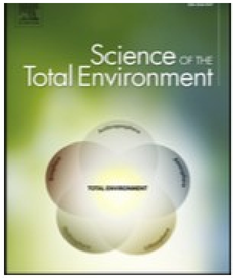
Case study: Which variables explain microbial leaf decomposition in streams?

Assumption: Linear relationship

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv



Organic matter breakdown in streams in a region of contrasting anthropogenic land use

K. Voß *, D. Fernández, R.B. Schäfer

Quantitative Landscape Ecology, Institute for Environmental Science, University of Koblenz-Landau, Fortstraße 7, D-76829 Landau, Germany

HIGHLIGHTS

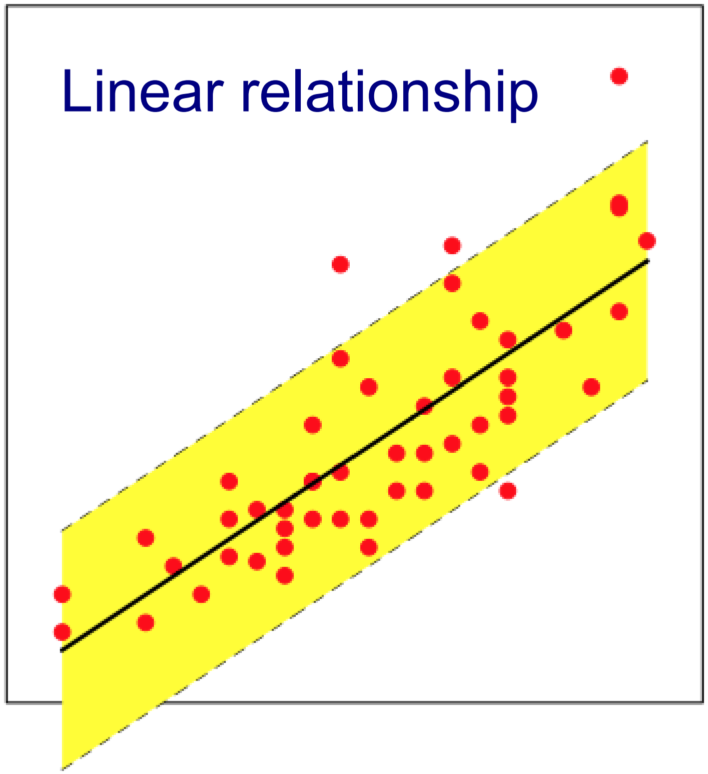
- Investigated land use effects on organic matter breakdown
- Only microbial breakdown differed between land use types
- Tree cover correlated with invertebrate-mediated breakdown
- pH correlated with microbial breakdown
- Land use insufficient to distinguish differences in breakdown

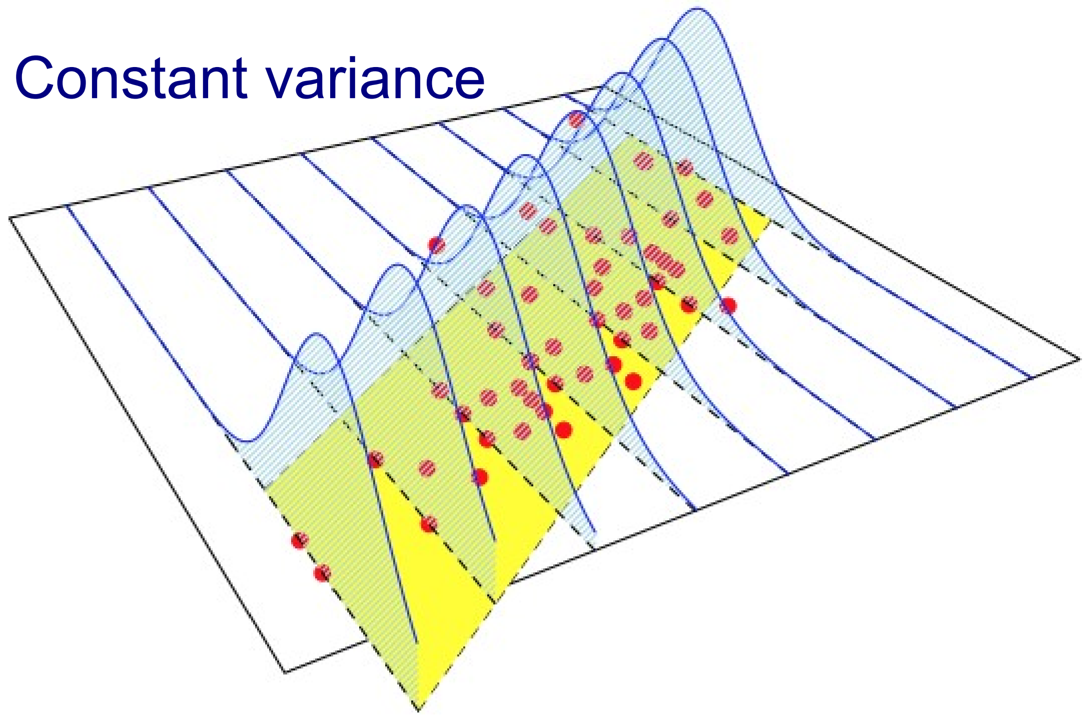
Linear regression model

- Bivariate relationship:

$$Y_i = \alpha + \beta_1 X_i + \epsilon_i, \quad \text{with} \quad \epsilon \sim N(0, \sigma^2)$$

- Assumptions





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

Linear regression model

- Bivariate relationship (simple regression):

$$Y_i = \alpha + \beta_1 X_i + \epsilon_i, \quad \text{with} \quad \epsilon \sim N(0, \sigma^2)$$

- Relationship between explanatory variables *X* (predictors) and a response variable *Y* (multiple regression):

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_m X_{m,i} + \epsilon_i, \quad \text{with} \quad \epsilon \sim N(0, \sigma^2)$$

Steps of multiple regression

1. Transform variables if necessary (check range, distribution)
2. Check explanatory variables for multicollinearity: Omit variables or adjust regression method

Data preparation

3. Search for best-fit model

Modelling

4. Check best-fit model with model diagnostics
5. Validate model using cross-validation or a validation sample
6. Determine relevance of individual explanatory variables

Model evaluation

Multicollinearity

- Strong correlation between explanatory variables (graphical inspection or correlation matrix)
- Can lead to wrong estimates of the regression coefficients (betas) and non-significant terms in the model, while the overall F-test indicates a highly significant model
- Scatterplots and Variance inflation factors (VIFs) can aid in identifying variables with high multicollinearity, but can not suggest what to do
- Strategies to deal with multicollinearity: Omission of variables from the model or adjust regression method (e.g. ridge regression, principal component regression).

How to identify the best-fit model?

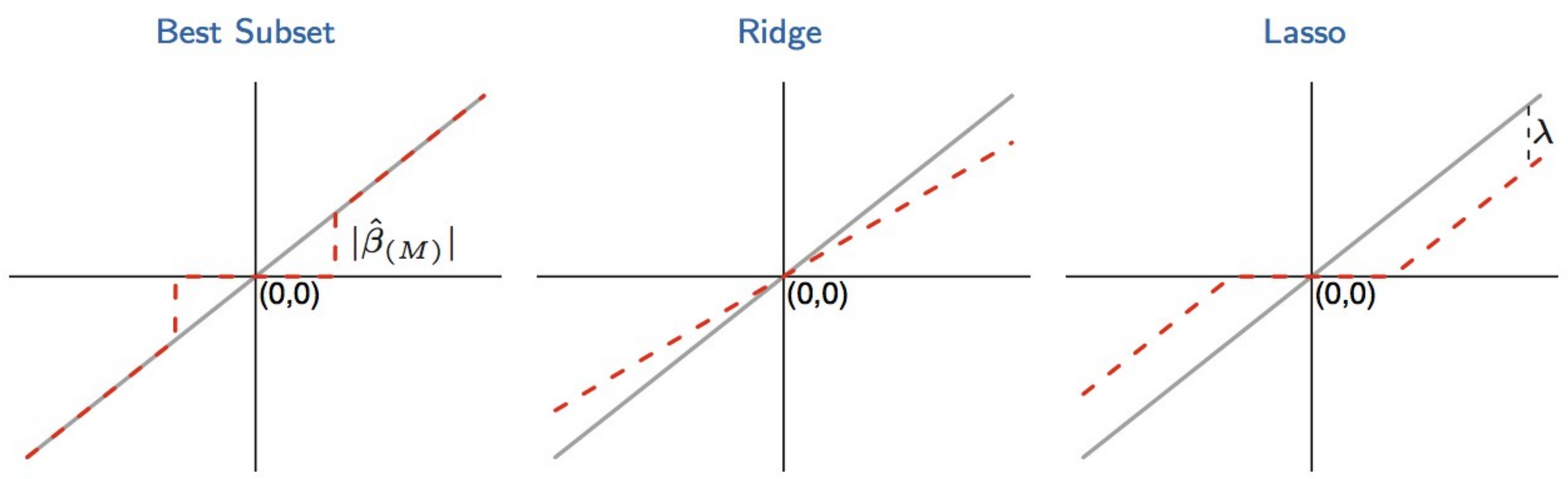
- Model selection – selection of variable subset:
  - (1) comparison of all possible models
  - (2) comparison of selected models (e.g. based on expert knowledge*)*
  - (3) stepwise variable selection

- Goodness of fit measures:
  - (1) Information theoretic (AIC, BIC)
  - (2) Explained variance (r^2 or adj. r^2)
  - (3) Hypothesis testing (Model variance or t-test for variables)

Issues: Overfitting, multiple testing → *p*-value inflation
Contrary to common belief, information theoretic approach has problems similar to hypothesis testing (cf. Murtaugh 2014, Taylor & Tibshirani 2015)

How to identify the best-fit model?

- Variable subset selection binary: retains or omits variable → prediction error can be high
- Shrinkage methods more continuous: Ridge regression and LASSO
  → set constraint on βs



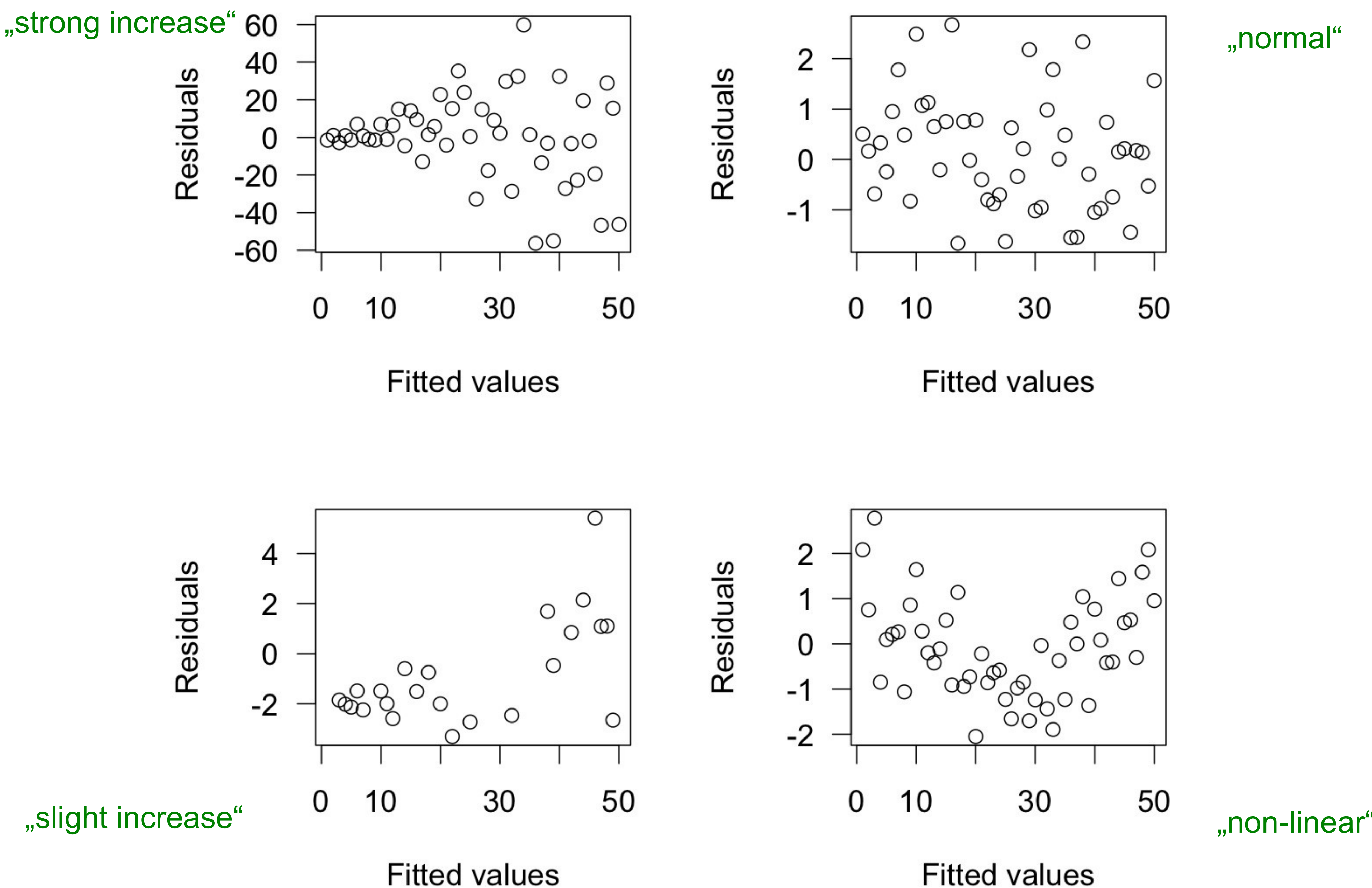
Hastie, Tibshirani & Friedman, 2009 Elements of statistical learning: 71

Diagnostics for the linear model

Check model assumptions:

- Normality of residuals
- Independence of residuals
- Linearity
- Homogeneity of residual variance

Diagnostics for the linear model



Diagnostics for the linear model

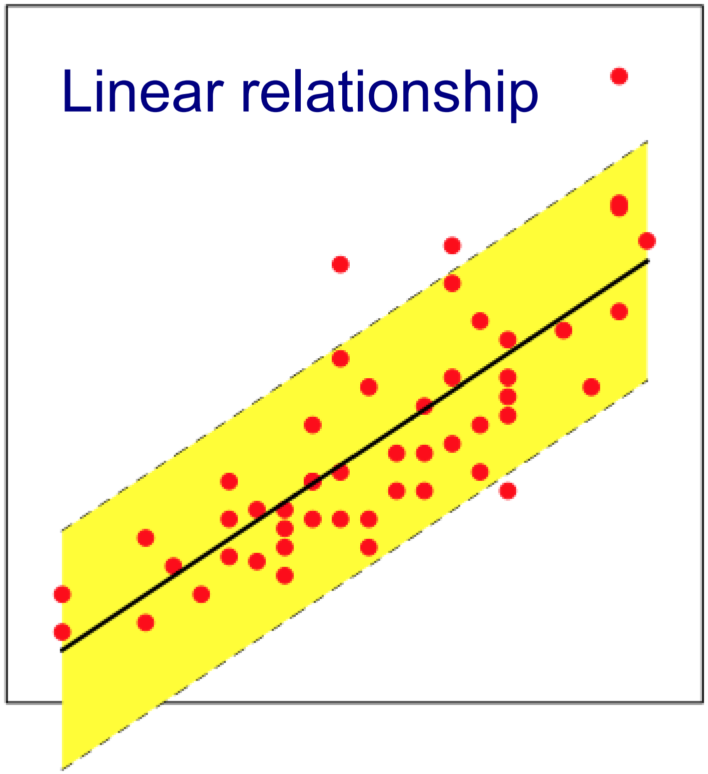
Check model assumptions:

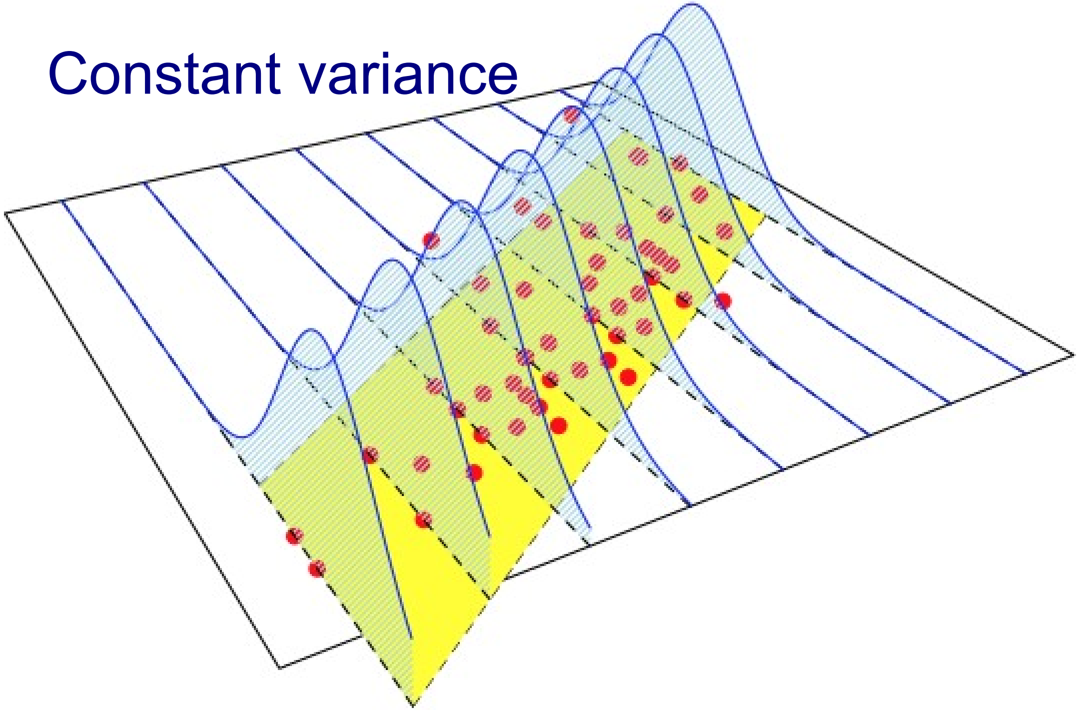
- Normality of residuals
- Independence of residuals
- Linearity
- Homogeneity of residual variance

Check for leverage points, outliers and influential points

Extending the linear model

- Assumption of linear relationship and constant variance often violated for ecological data

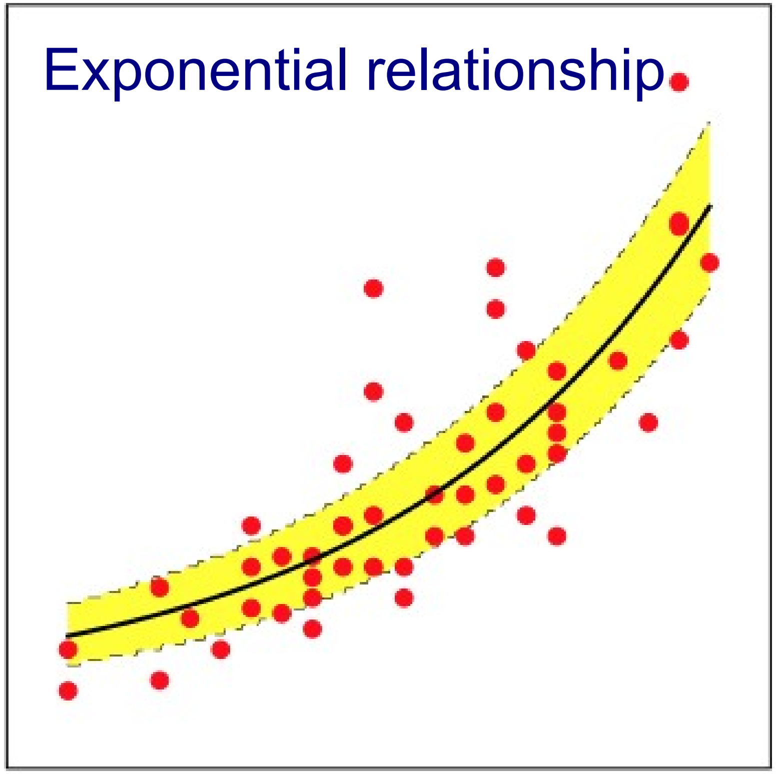


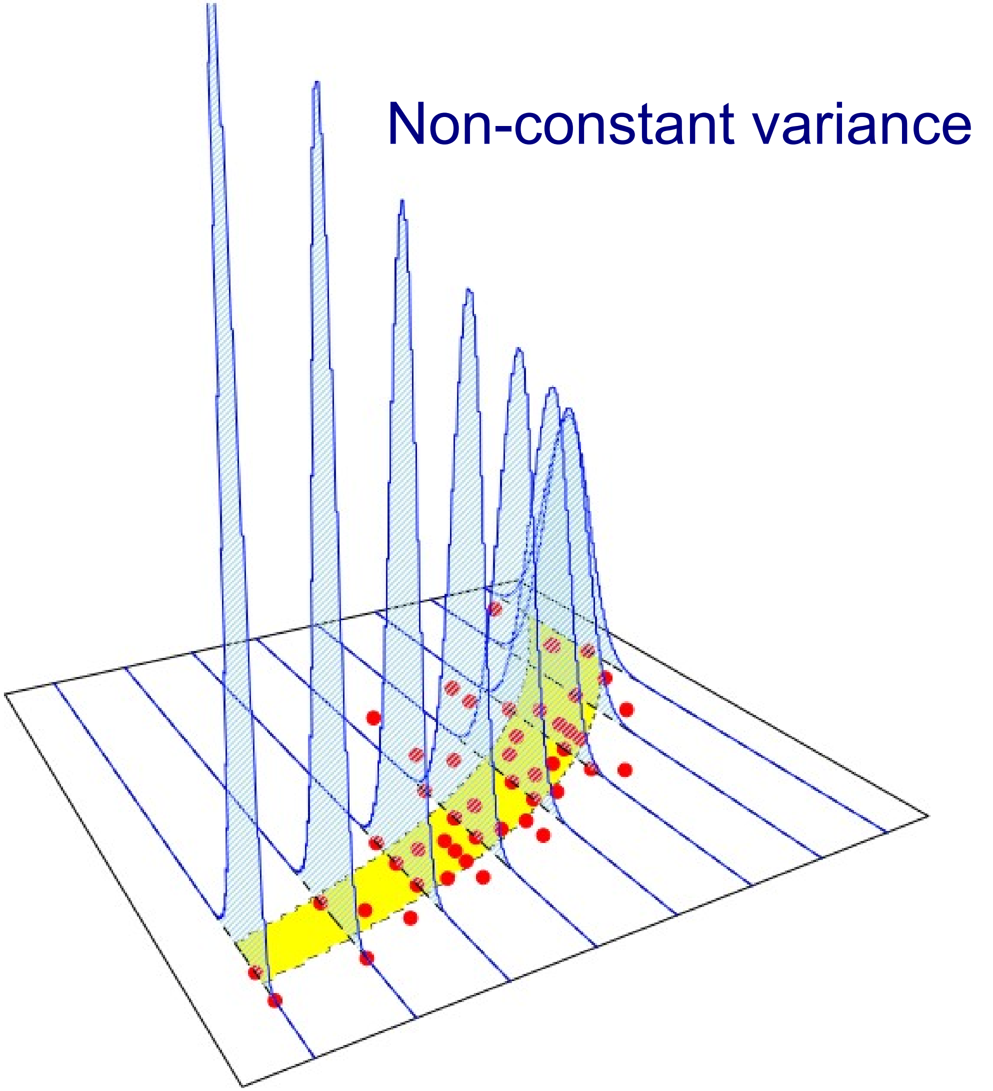


http://freakonometrics.hypotheses.org/tag/poisson

Extending the linear model

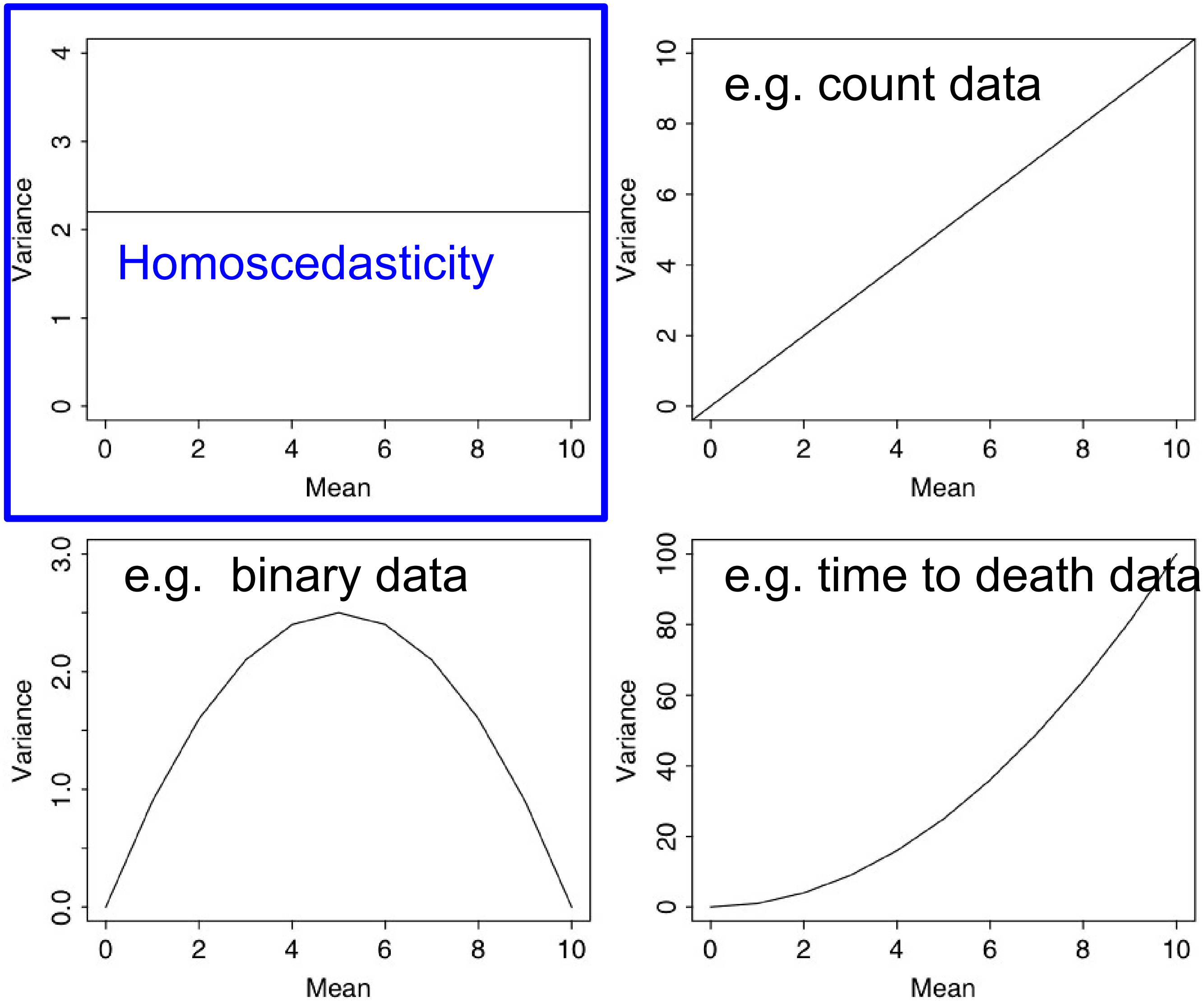
- Example: Exponential loss of ecosystem functioning with decline in biodiversity





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

Extending the linear model



taken from Crawley 2007: 511

- Variance is non-constant (Heteroscedasticity), but can be expressed as a function of the mean

Generalised linear model (GLM)

Comparison of model structures

L $Y_i = \alpha + \beta_1 X_i + \epsilon_i, \quad \text{with} \quad \epsilon \sim N(0, \sigma^2)$

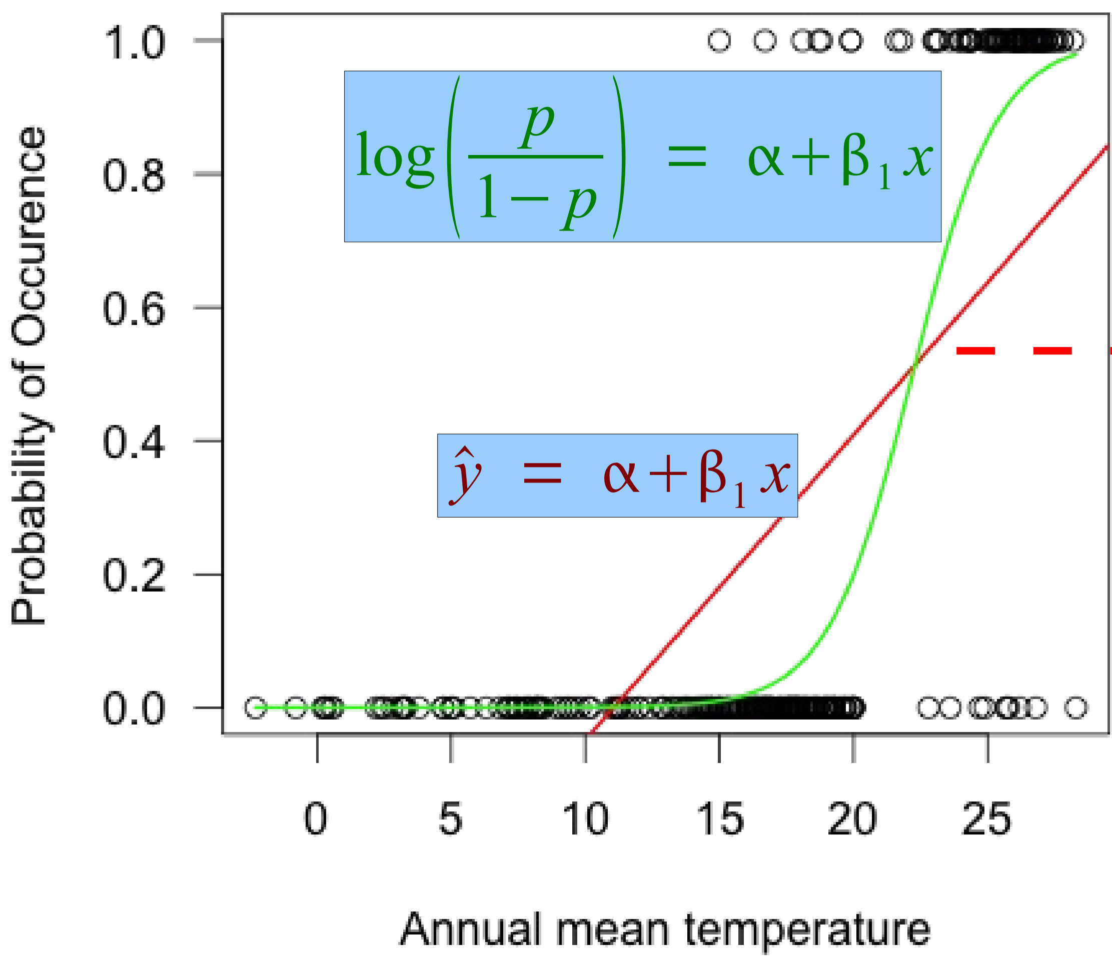
Generalised linear model:

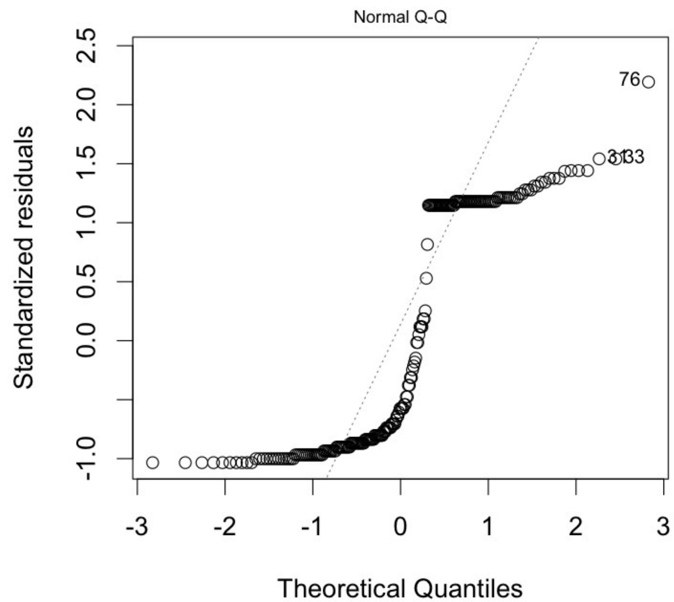
1. Linear predictor: $g(\mu) = \alpha + \beta_1 X$
2. Link function: $g(\mu) = \eta$
3. Error distribution of response variable

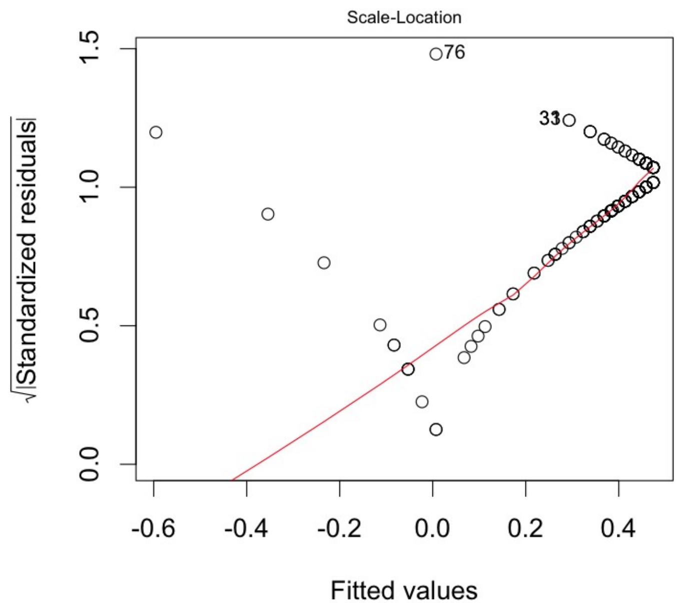
Family (error structure)	Link	Variance function
normal	$\eta = \mu$	1
poisson	$\eta = \log \mu$	μ
binomial	$\eta = \log(\mu/(n-\mu))$	$\frac{\mu(n-\mu)}{n}$
Gamma	$\eta = \mu^{-1}$	μ^2
inverse. gaussian	$\eta = \mu^{-2}$	μ^3

modified from Crawley 2007: 511

Example: Binomial GLM vs. LM







Goodness of fit for the GLM: Deviance

- GLMs minimize Deviance instead of Sum of Squares
- Deviance derived by maximum likelihood estimation (MLE)

Relation between error structure, Deviance and variance function

Family (error structure)	**Deviance**	**Variance function**
normal	$\sum (y-\bar{y})^2$	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
$\bar{y}$ = mean for y
μ = fitted values
n = binomial denominator

taken from Crawley 2007: 511

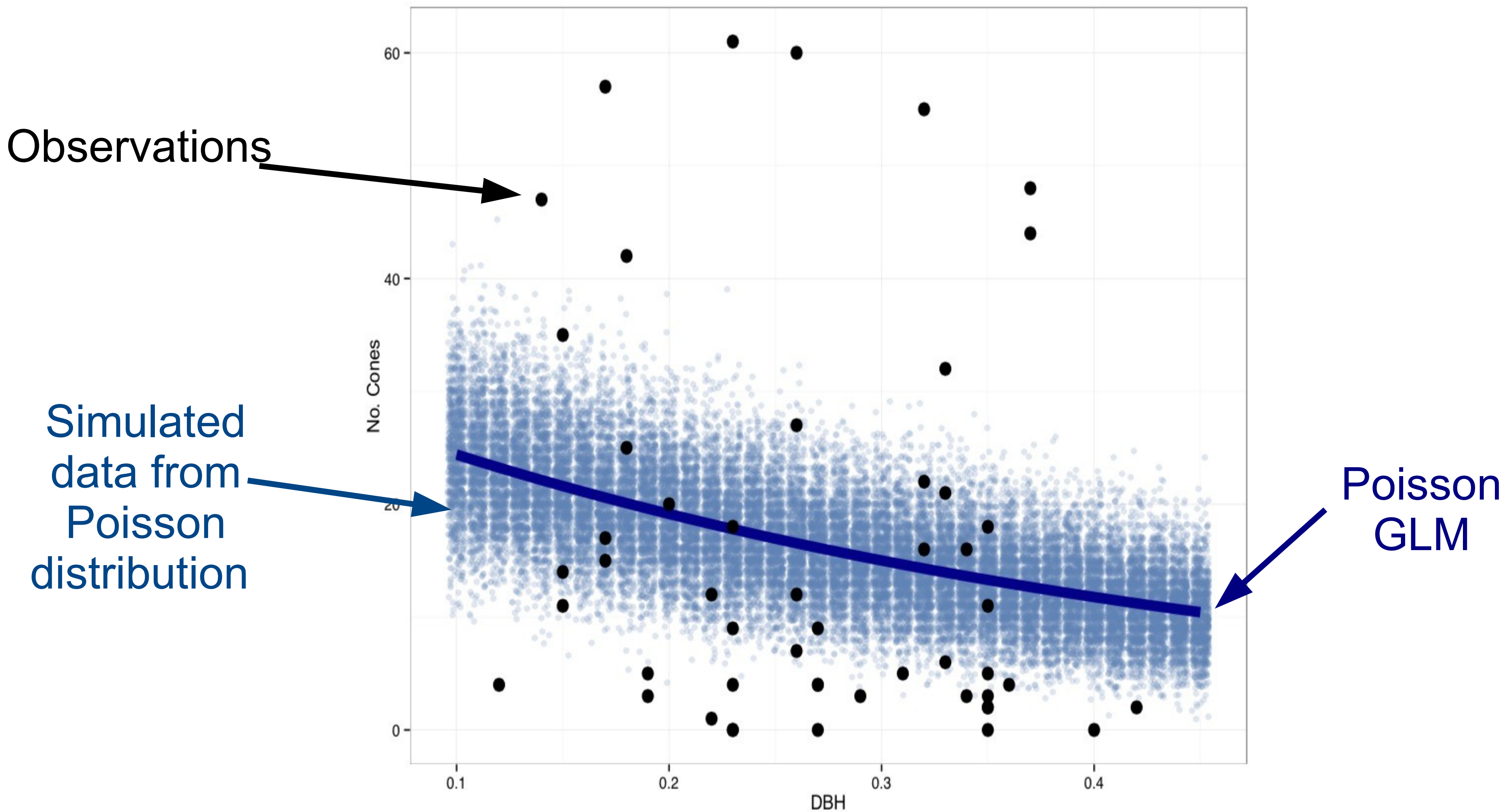
Modelling

- Follows same approaches as described for LM
- Hypothesis-based approach:
  - Wald test (*t* or *z* ratio depending on sample size) for single model parameters
  - Log-likelihood ratio tests for comparison of full and reduced model
- Information-theoretic approaches (e.g. AIC, BIC)
- Relative importance based on partitioning of Deviance

GLM assumptions and diagnostics

- Independence of observations
  - In case of temporal- or spatial autocorrelation: GLMMs (see Bolker 2009)
- Linear relationship between link function and predictor (Component-residual plot)
  - Non-linearity: Use nonlinear or nonparametric (e.g. GAMs) regression (see Zuur 2007)
- Assumed Mean-to-variance relationship holds (no over- or underdispersion) (graphical diagnostics with *q*-*q* plot and dispersion parameter)
- Checking for influential observations (graphical diagnostics and measures e.g. Cooks distance)

Overdispersion



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

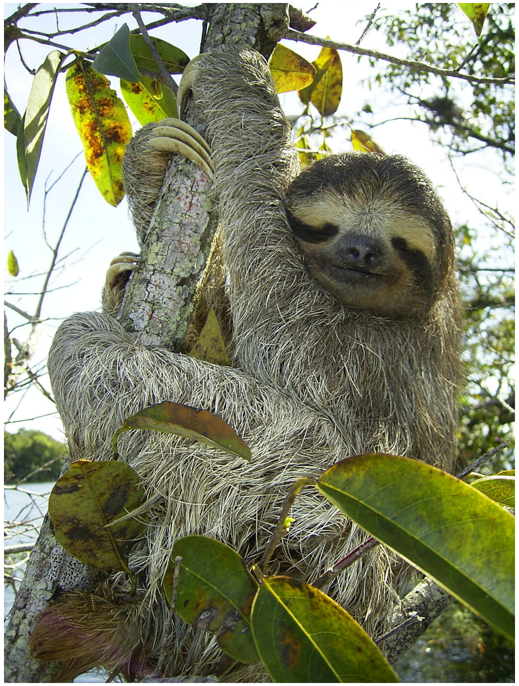
Demonstration and Exercise

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 best fit model to explain the occurrence of the *Bradypus sp*.