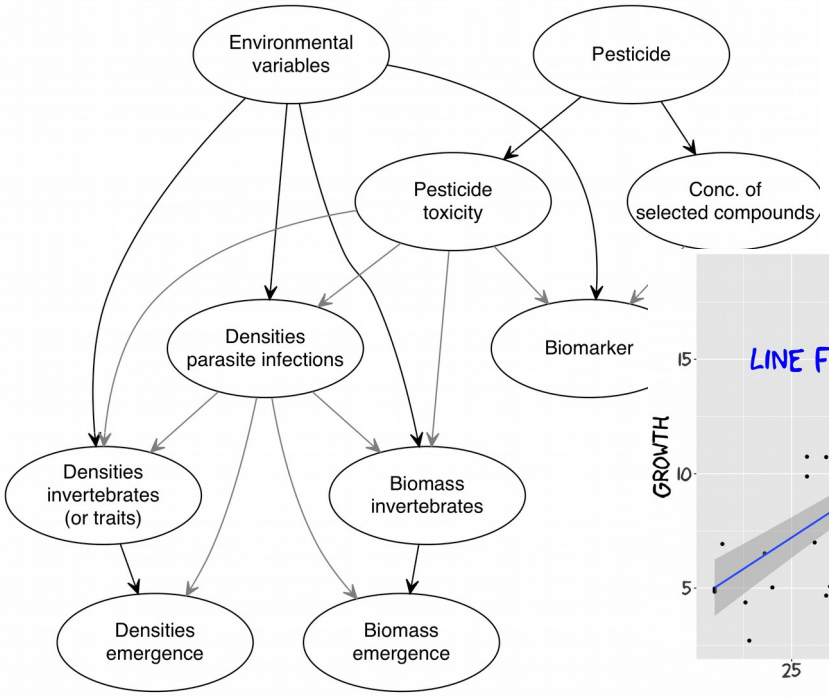
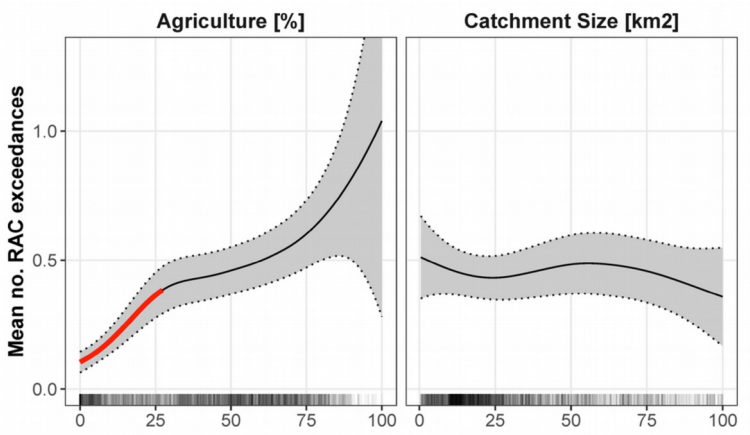
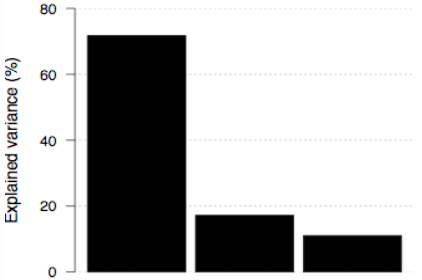
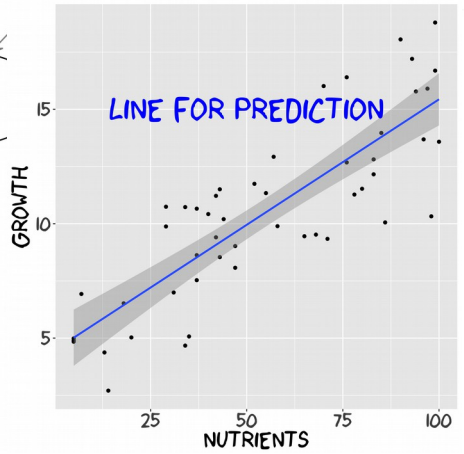
Advanced analysis of ecological data

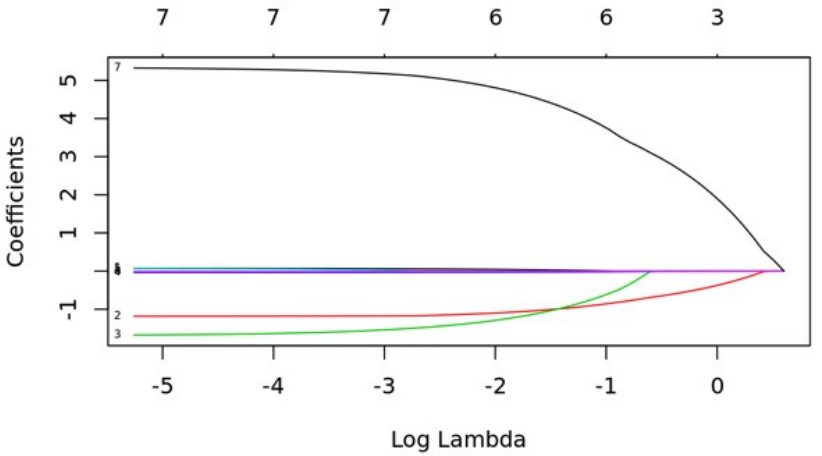
SEFS, Zagreb 2019











Ralf B. Schäfer, Andreas Scharmüller, Moritz Link

UNIVERSITÄT KOBLENZ · LANDAU

Short intro: Ralf Schäfer

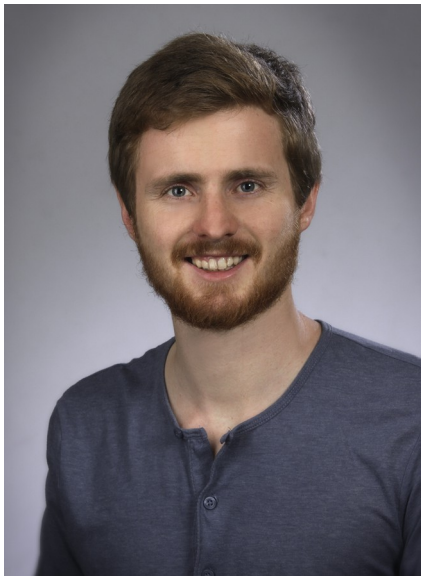
- Prof for Quantitative Landscape Ecology @UKL
- Phd @ UFZ, Leipzig; Postdoc @ RMIT, Australia
- Teaching: Data analysis; GIS; Environmental Modelling; Environmental Philosophy
- Current research projects related to:
  - Community ecology of freshwater invertebrates and microorganisms
  - Response of freshwater ecosystems to different (anthropogenic) stressors (e.g. pollution)
  - Trophic linkages between aquatic & terrestrial systems
- Primarily field studies/experiments and data analyses/ modelling

www.landscapecology.uni-landau.de

@LandscapEcology

Short intro: Andreas Scharmüller

- PhD student Quantitative Landscape Ecology
- Environmental Sciences + Ecotoxicology
- Teaching:
  - Statistics, GIS
- Research:
  - Effects and distribution of pesticides in freshwaters
  - Ecotoxicology
- R programming:
  - Package author: standartox (in preparation)
  - Package contributions: webchem



www.landscapecology.uni-landau.de @andschar

Short intro: Moritz Link

- PhD student, Quantitative Landscape Ecology
- M.Sc. Ecotoxicology @ Uni Koblenz Landau
- Teaching:
  - Course assistence in multivariate statistics
- Research:
  - Data analysis
  - Surface water monitoring
  - Ecosystem services of aquatic fungi

link@uni-landau.de

@LandscapEcology

Course Organisation

10:00-10:15 Short intro & course organisation, Software preparation

10:15-12:00 Model selection for (G)LMs

13:00-14:30 Generalized additive models (GAMs)

15:00-16:30 Structural equation models (SEMs)

16:30-16:45 Course evaluation

Course material:
https://github.com/andreasLD/workshop-sefs11

Course structure: intro – demo – hands on exercises

Part I: Model selection in (G)LMs

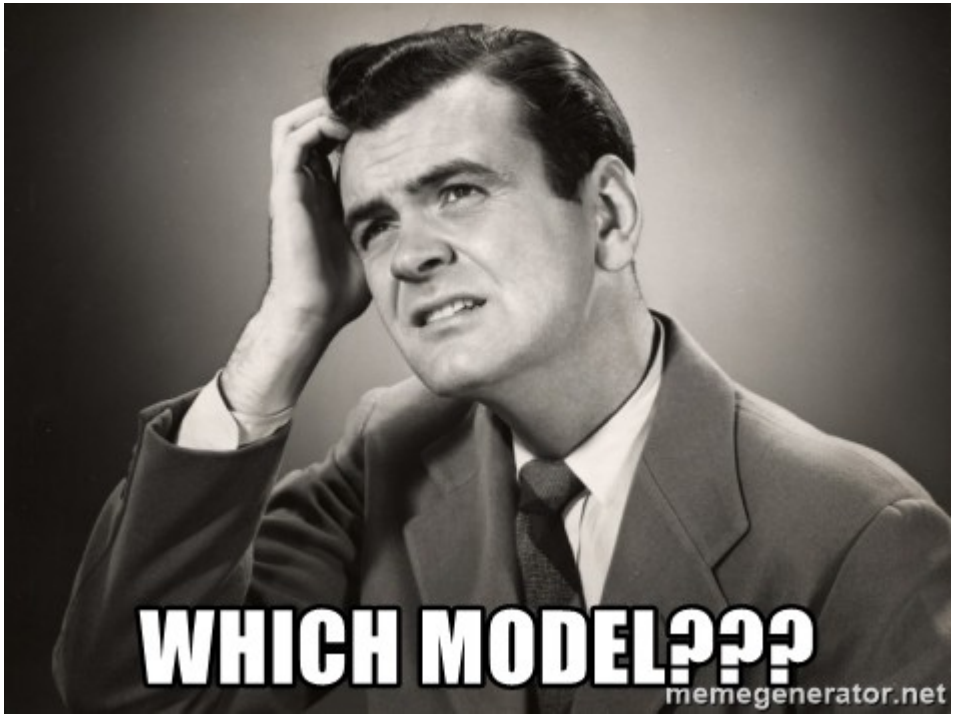
Model 1: *Diversity ~ pH, Temp., NO_3, TU, Cond.*

Model 2: *Diversity ~ TU, Temp., NO_3*

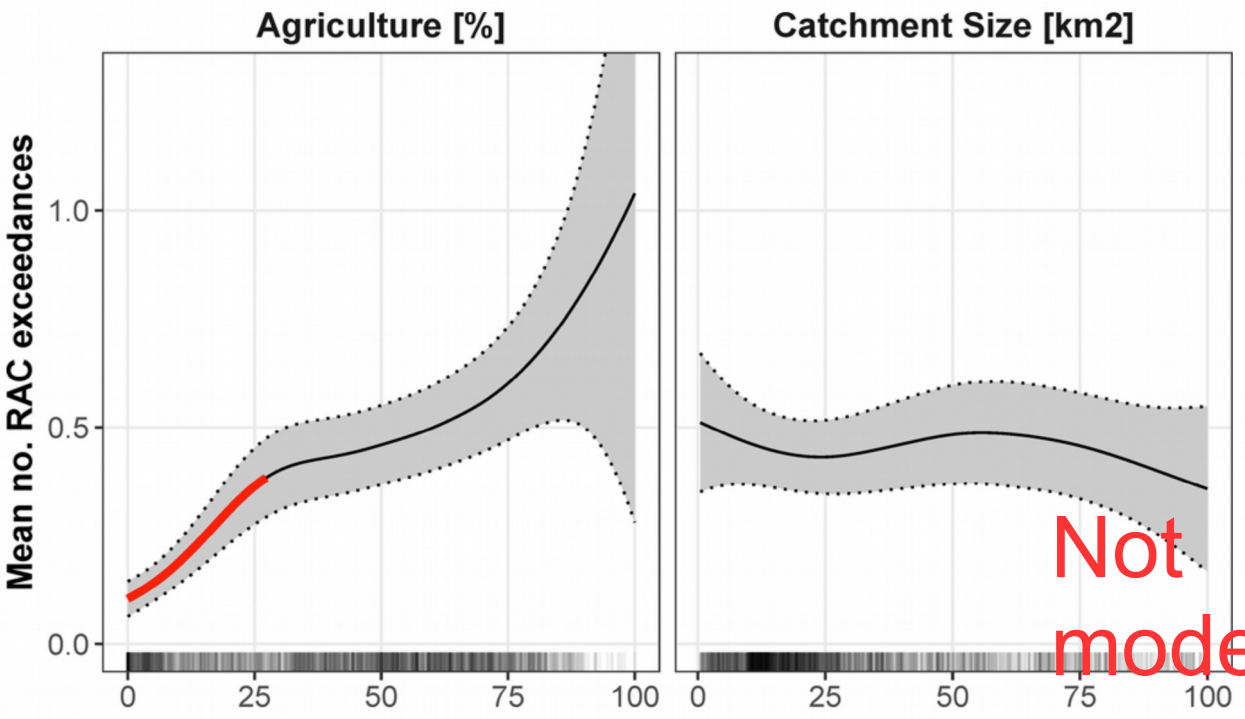
Model 3: *Diversity ~ Temp., pH, TU, NO_3*

⋮

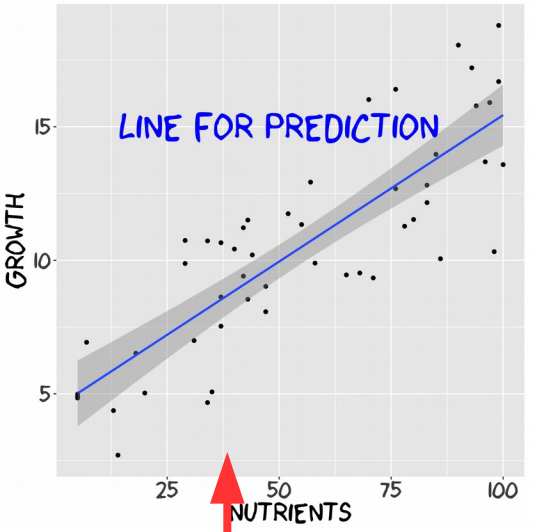
Model n: *Diversity ~ TU, pH*



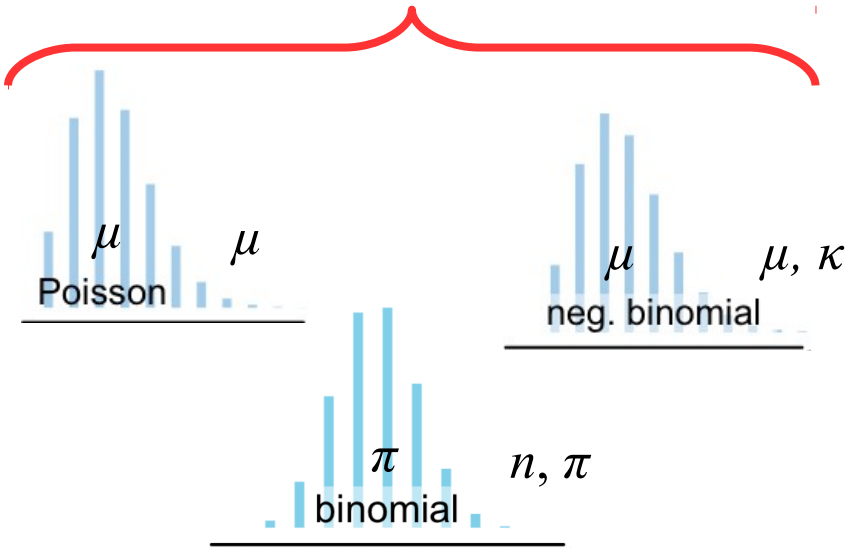
Part II: GAMs



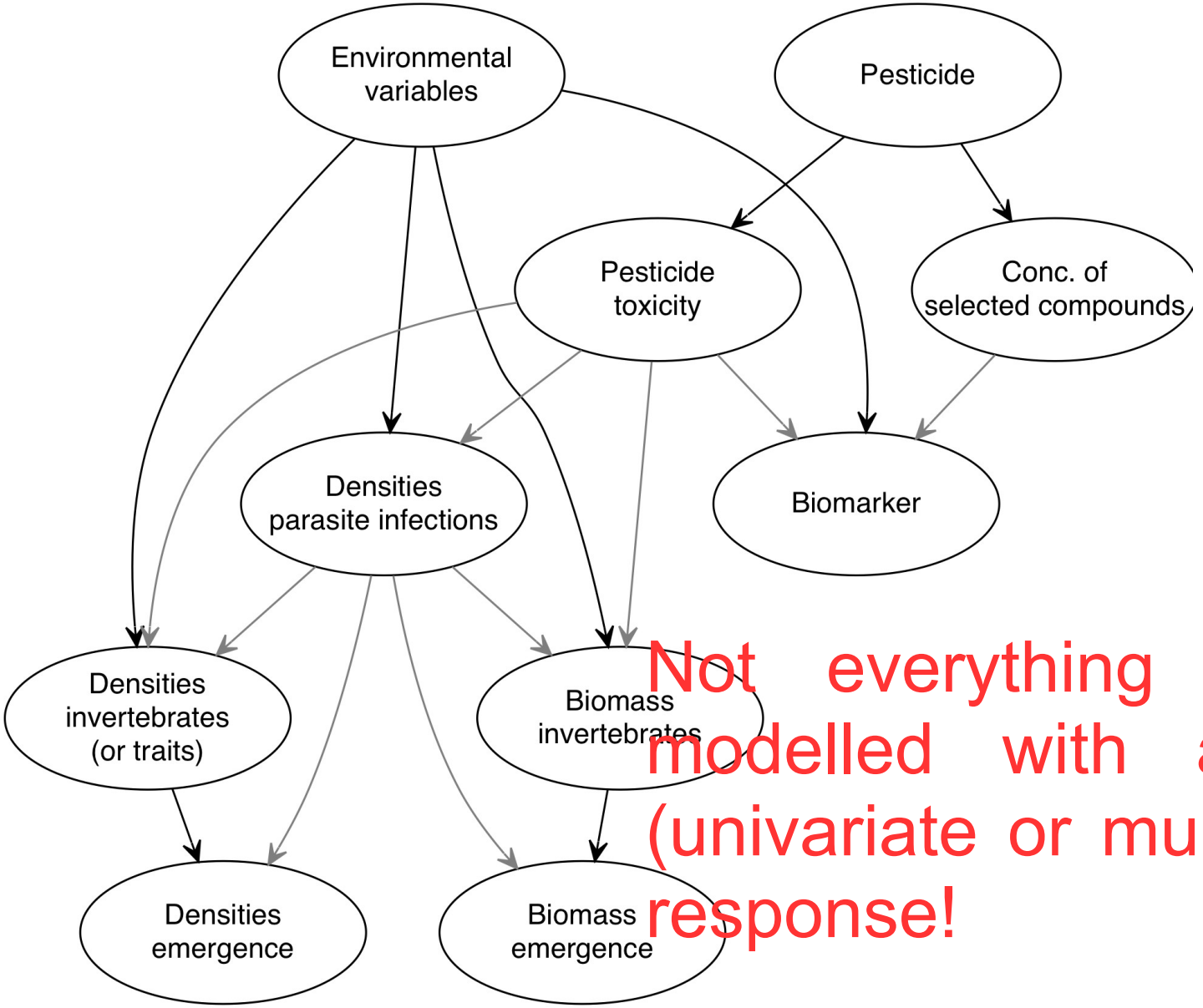
Szöcs et al. 2017 *Environ. Sci. Technol.*



Not everything can be modelled with a linear or generalized model!



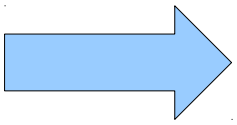
Part III: SEMs



Not everything can be modelled with a single (univariate or multivariate) response!

Your background

- 95% beginner or intermediate R users and data analysts
- ~70% have used (G)LMs, 20% lack knowledge
- 50% don't know GAMs, 40% have a little bit of prior knowledge
- > 80% don't know SEMs

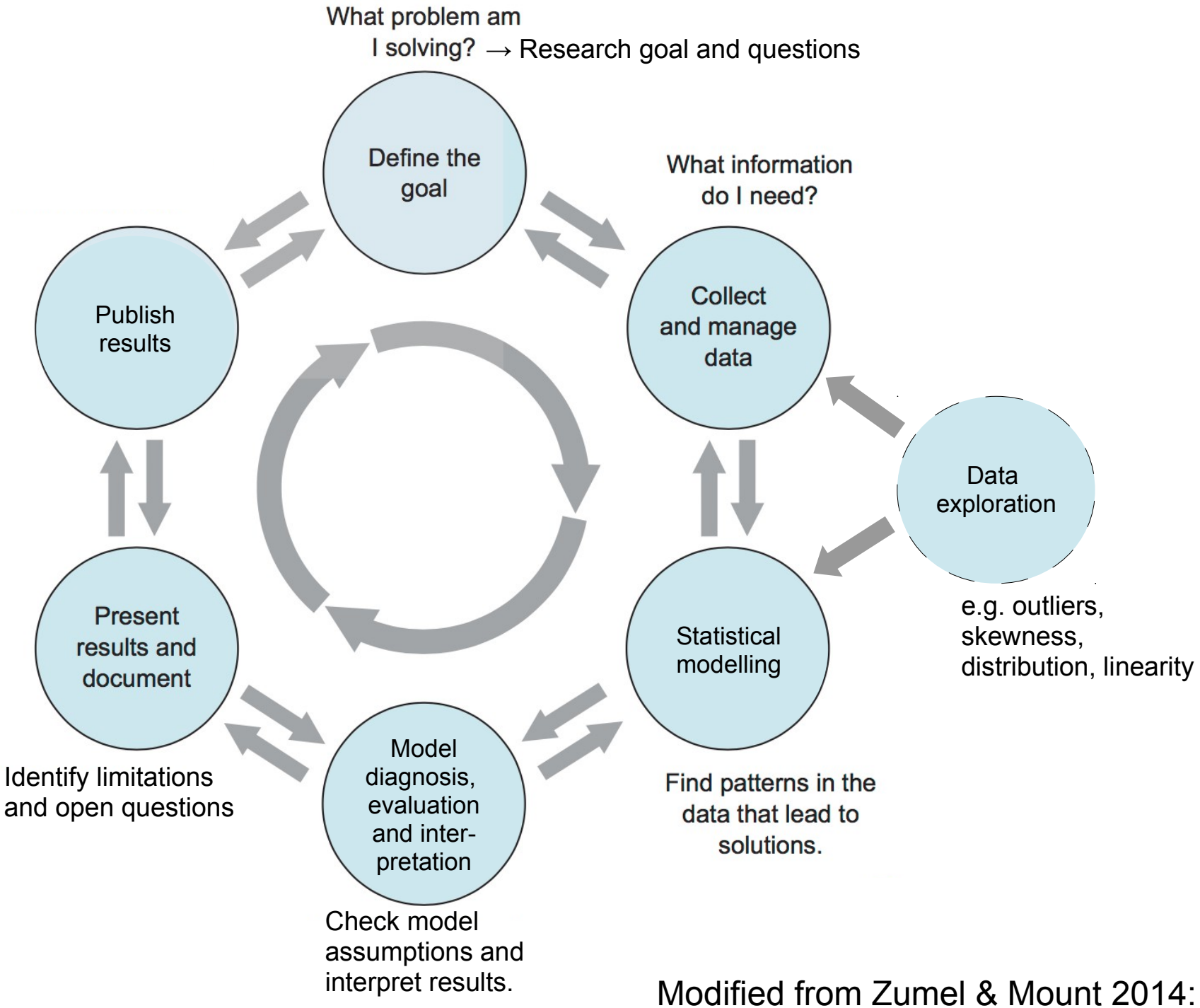


Very brief intro into (G)LMs, more extensive intro into GAMs and SEMs

Block I

Model selection for Linear and Generalised Linear models

Data analysis cycle



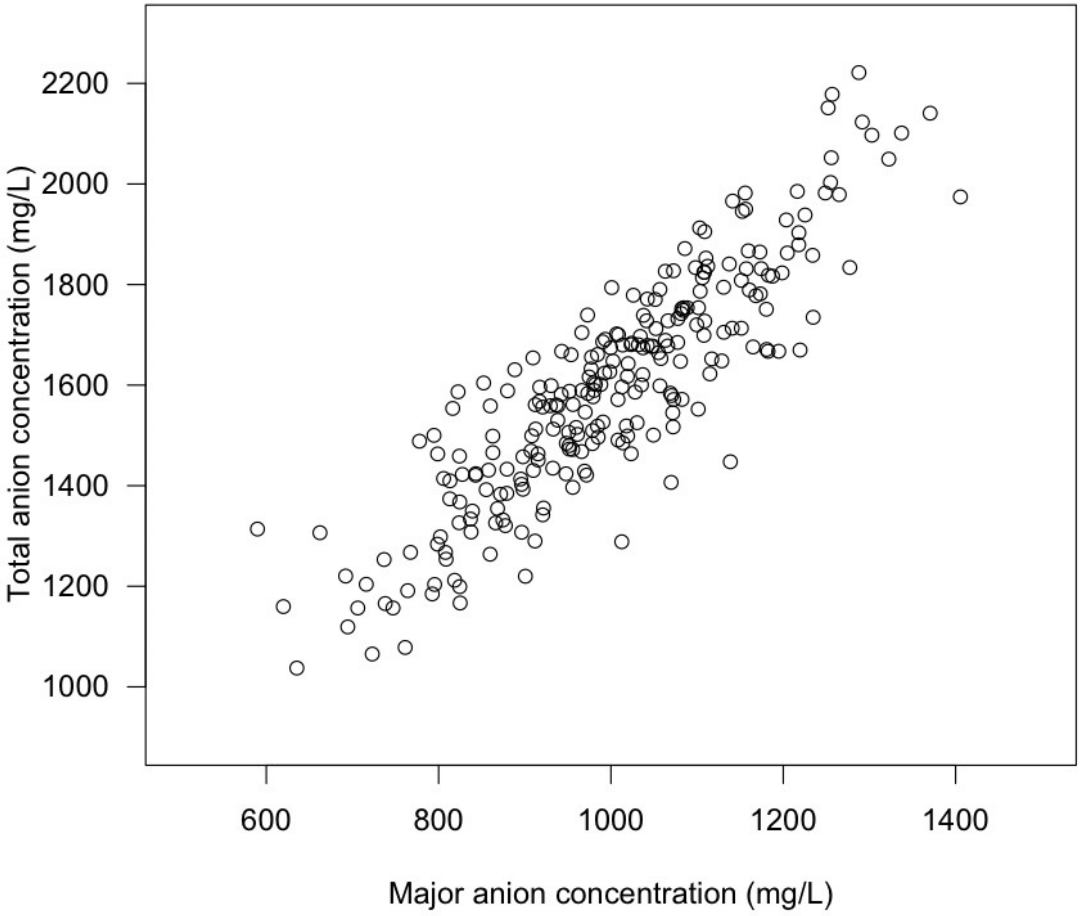
Modified from Zumel & Mount 2014: 6

Case study: Water concentrations

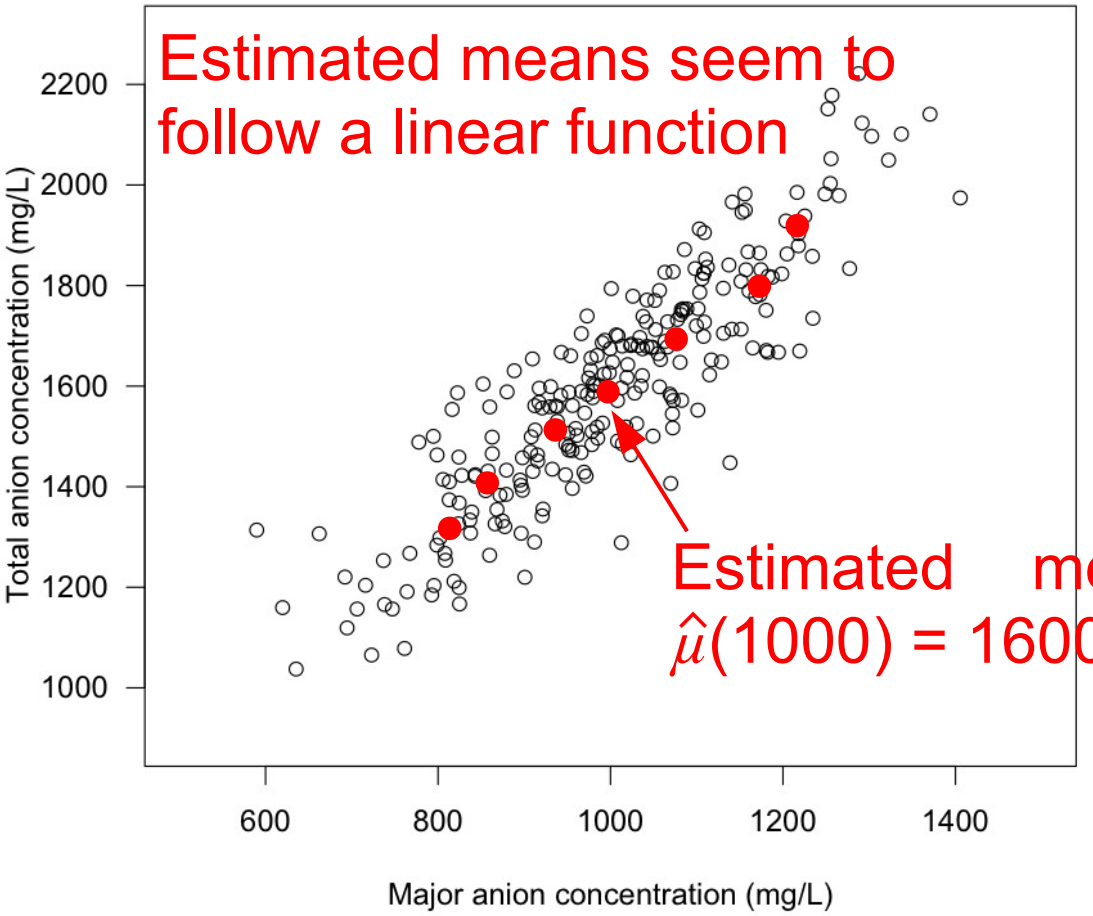
Research question: Can we predict the total anion concentration in water from the concentration of major anions (Cl^-, SO_4^{2-}, PO_4^{3-})?

Study: Samples of total anion and major anion concentrations from 250 streams.

https://upload.wikimedia.org/wikipedia/commons/1/11/Water_resources%2C_taking_a_water_sample.jpg



Predicting Y from X with a linear function



We assume a linear function of the true population $\mu(X)$:

$\mu(X) = \beta_0 + \beta_1 X$ from which follows that: $Y = \beta_0 + \beta_1 X + \varepsilon$

β_0 and β_1 : regression coefficients

(Simple) Linear regression model

Assuming that the true relationship is a linear function of the form $Y = \beta_0 + \beta_1 X + \varepsilon$, we can use sample data to obtain estimates of β_0 and β_1, denoted as $\hat{\beta}_0 = b_0$ and $\hat{\beta}_1 = b_1$, and subsequently predict $\hat{Y}$:

$\hat{Y} = b_0 + b_1 X$ for realisations of X we can rewrite this to:

$\hat{y}_i = b_0 + b_1 x_i$

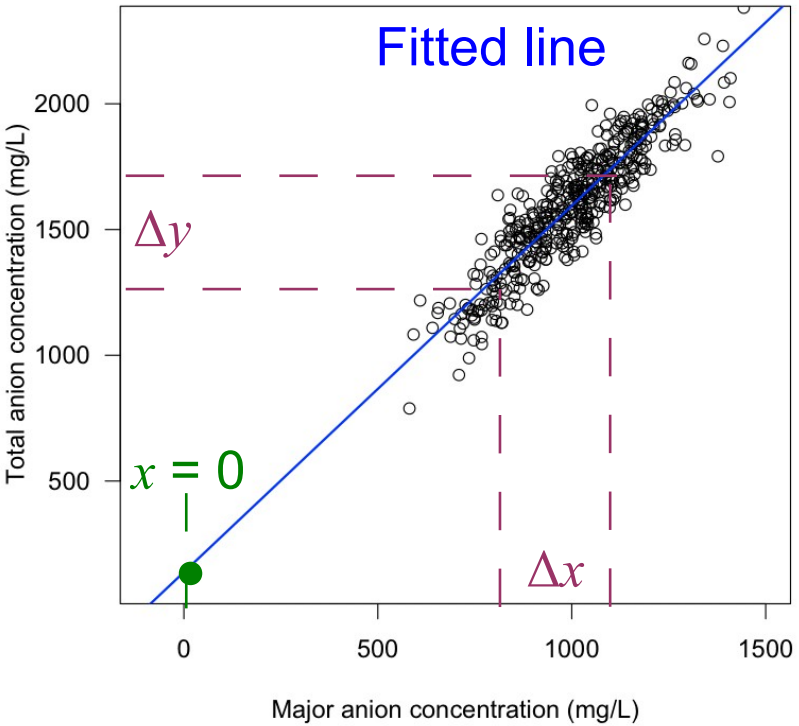
What are β_0 and β_1?

$\beta_0 = E(Y|X = 0)$ "intercept"

$\beta_1 = \frac{dy}{dx}$ "slope"

$b_0 = 138$

$b_1 = 1.5$



What is the optimal regression line?

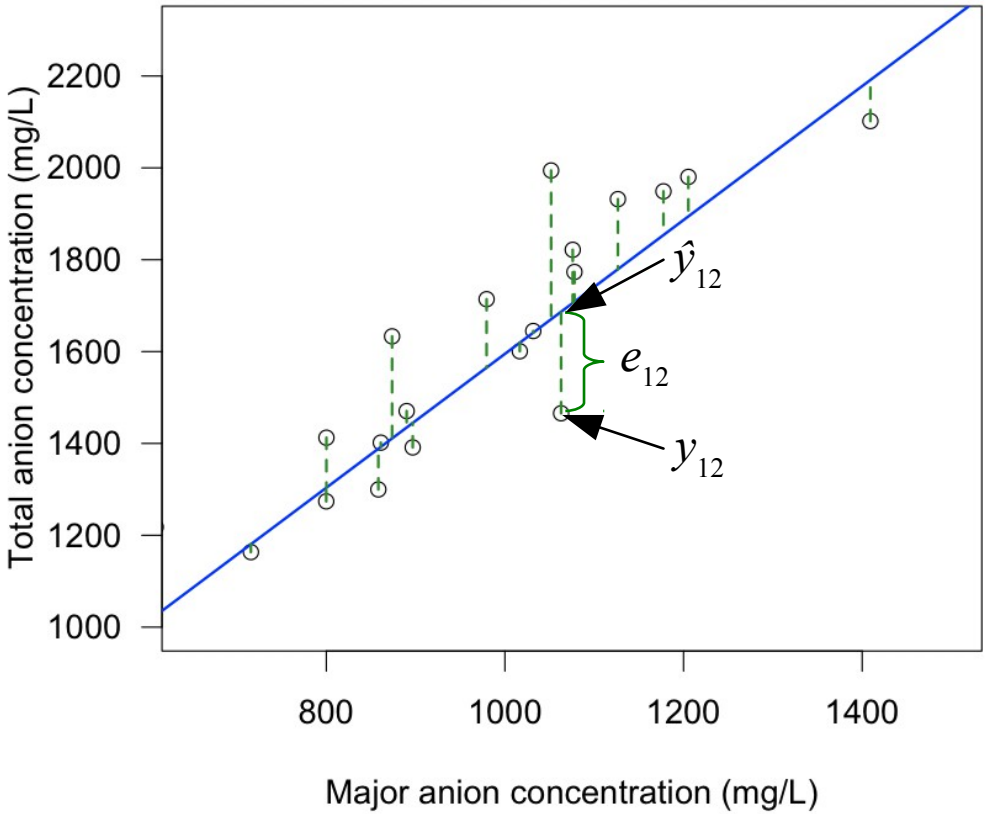
We defined for prediction: $\hat{Y} = f(X)$ and $Y = f(X) + \varepsilon$

$\Rightarrow Y = \hat{Y} + \varepsilon \quad \Leftrightarrow \quad \varepsilon = Y - \hat{Y}$

For sample data ($i = 1,2,3, \ldots, n$) and the regression model, we defined: $\hat{y}_i = b_0 + b_1 x_i$

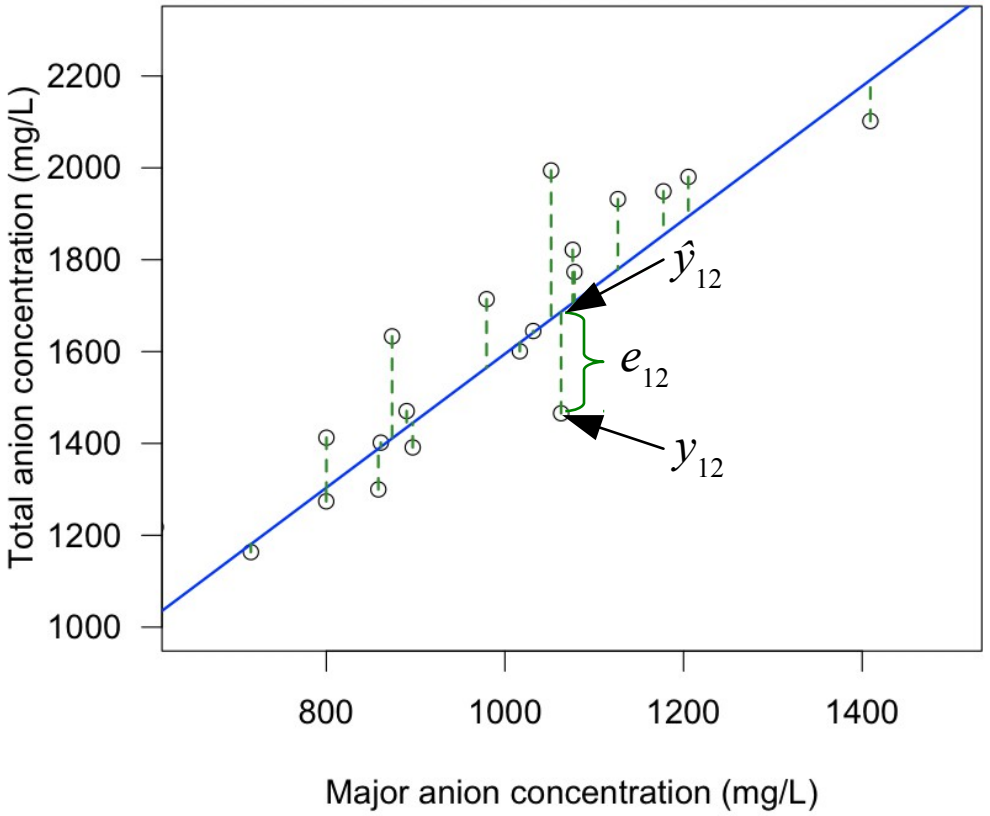
We define the residual e_i as: $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$

Example for observation i = 12



What is the optimal regression line?

Example for observation i = 12



Optimal line minimises the Residual Sum of Squares (RSS):

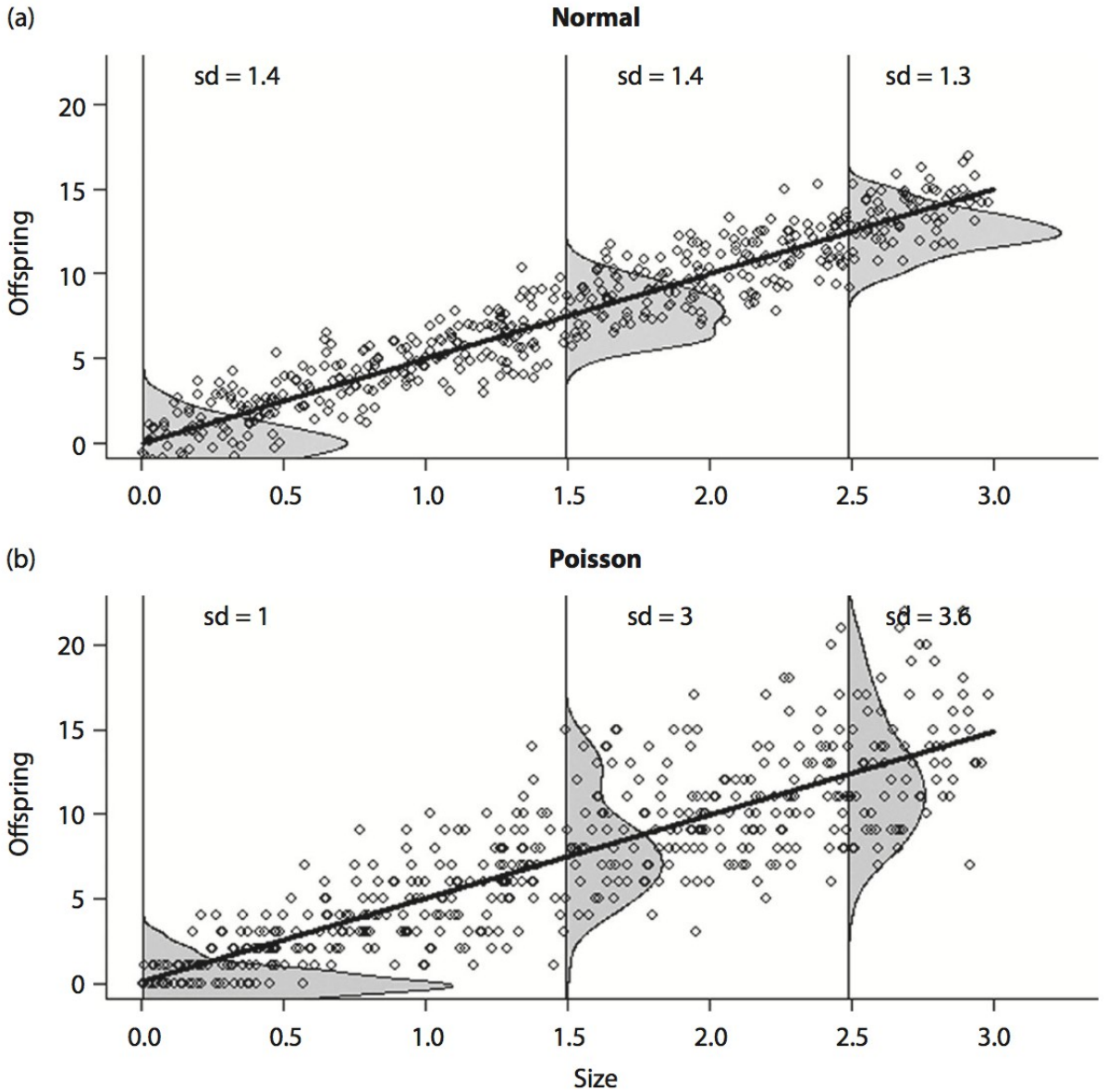
$$\text{RSS} = e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2$$

$$= (y_1 - (b_0 + b_1 x_1))^2 + (y_2 - (b_0 + b_1 x_2))^2 + \ldots + (y_n - (b_0 + b_1 x_n))^2$$

$$= \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \Rightarrow \text{Find} \arg\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

Extending the linear model: Motivation

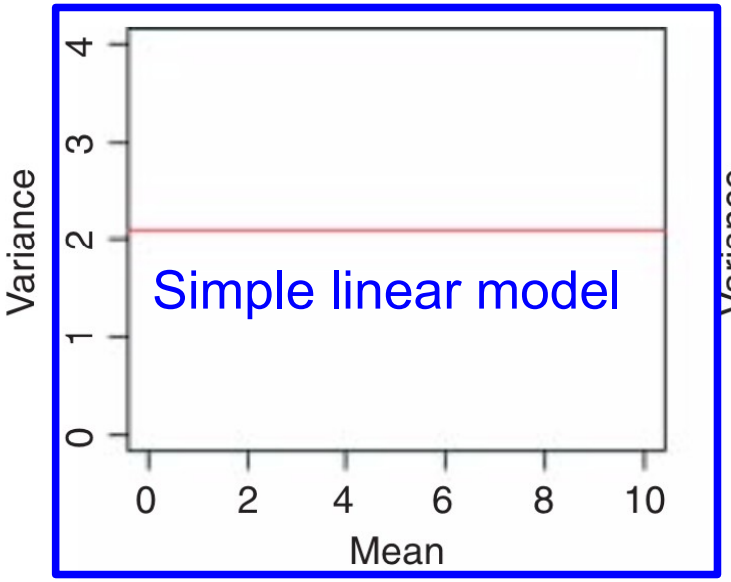
Example: Increasing variability in number of offsprings with increasing body size of individuals

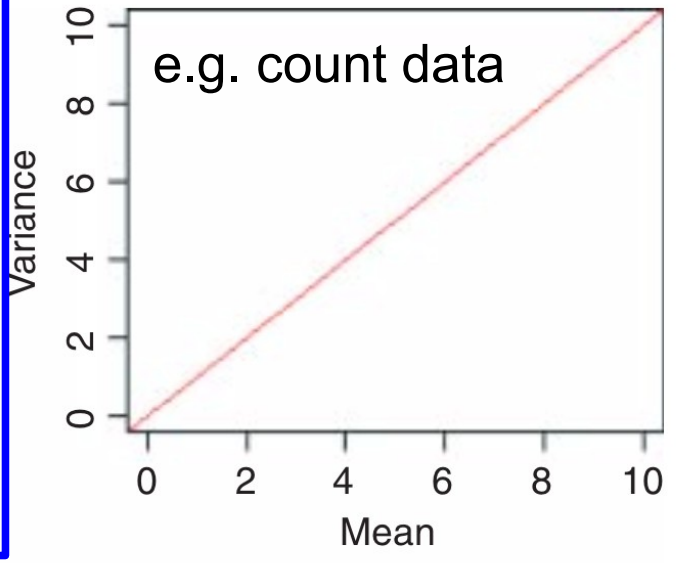


Buckley 2015: 134

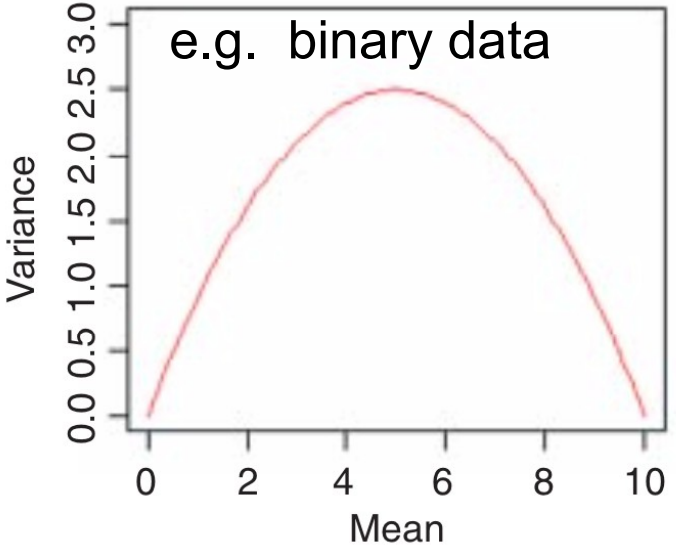
Modelling the mean-variance relationship

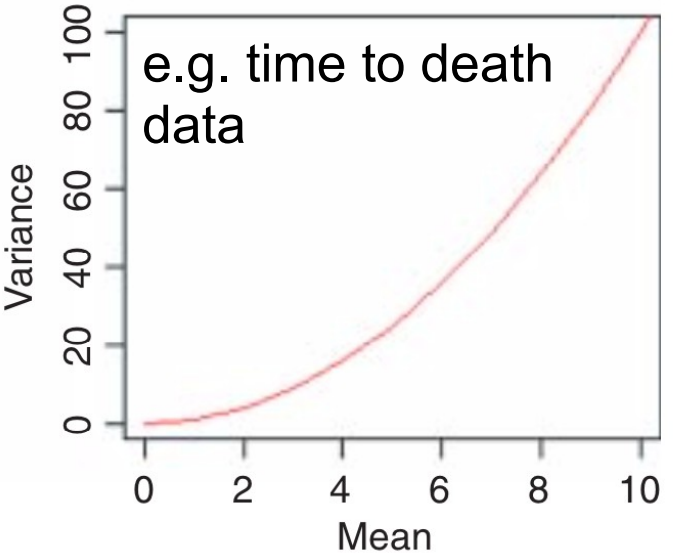
Idea: Express variance as a function of the mean!





taken from
Crawley 2012: 557





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

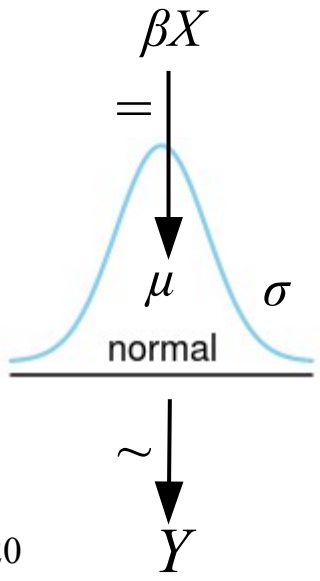
$Y \sim \text{Normal}(\mu, \sigma)$

$E(Y) = \mu$

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

$\mu = \beta X$

$\varepsilon = y - \mu$



Binomial

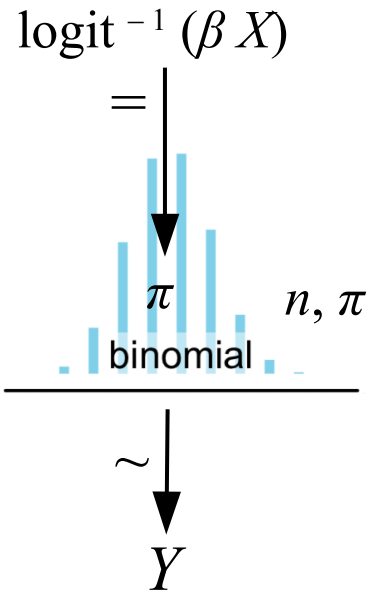
$Y \sim \text{Bin}(n, \pi)$

$E(Y) = \pi$

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

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

$\varepsilon = y - \pi$



Poisson

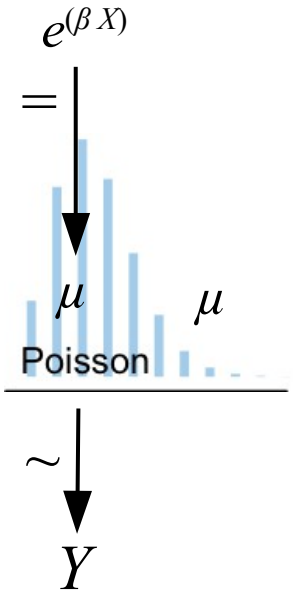
$Y \sim \text{Pois}(\mu)$

$E(Y) = \mu$

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

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

$\varepsilon = y - \mu$



Negative binomial

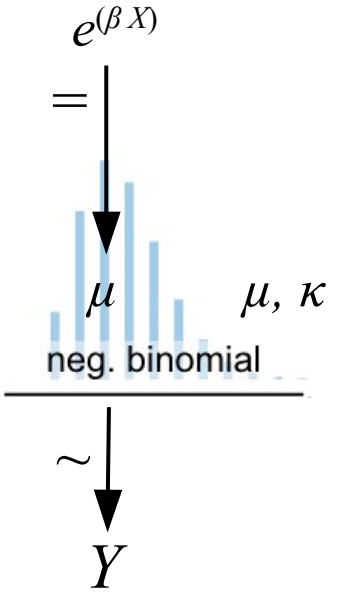
$Y \sim \text{Neg.Bin}(\mu, \kappa)$

$E(Y) = \mu$

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

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

$\varepsilon = y - \mu$

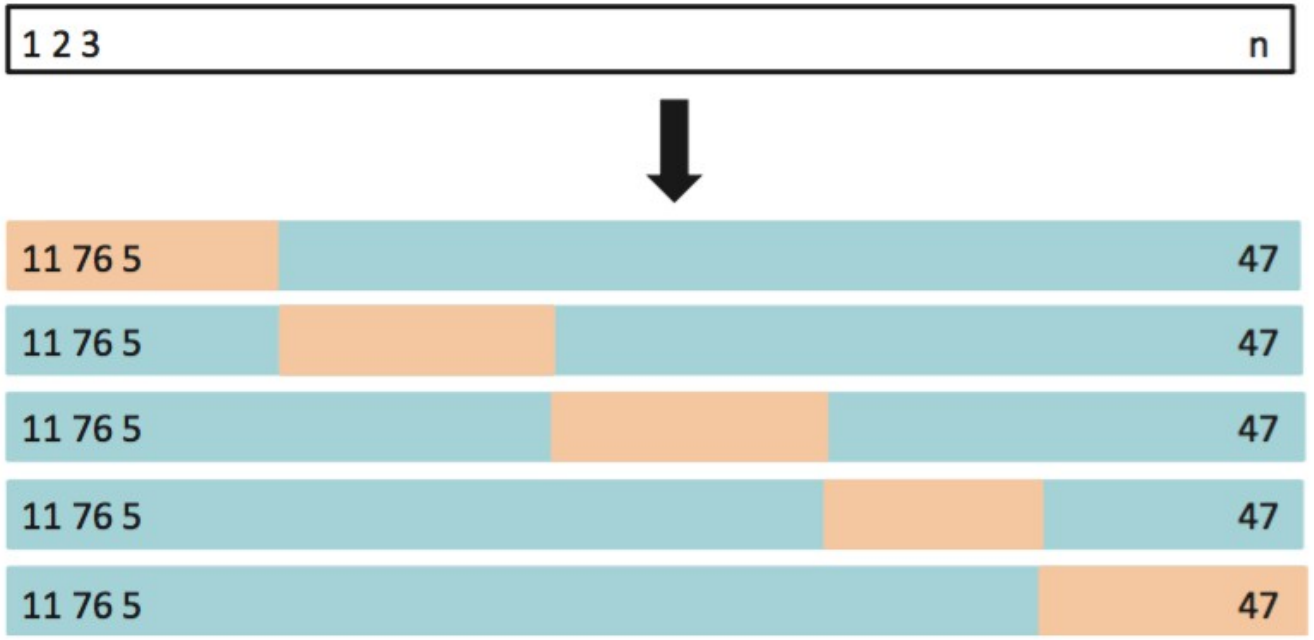


Cross-validation (CV)

- **Aim:** Evaluate predictive accuracy of a fitted model
- Can be checked by predicting (known) responses from independent data sets (that were not used in model fitting) → Rare case
- **Idea:** Split the available data into training and test set and predict (known) observations in test set with a model fitted on the training data
- Algorithm:
  1. Draw *k* random samples without replacement from data
  2. For each *k:*
     1. Fit the model to the other *k*-1 parts
     2. Predict *k* from model and calculate the prediction error
  3. Calculate mean prediction error over the *k* estimates

Cross-validation (CV)

Example: $k = 5$



- Problem of choosing *k*:
  - *k* = n (Leave-one-out CV predicts each observation from all others) → low bias, but high variance
  - *k* = 2 (split data into half) → low variance, but high bias
- *k* typically set to 5 or 10

Taken from James *et al.* 2013: 181

Bias-variance trade-off

Definition in context of model validation:

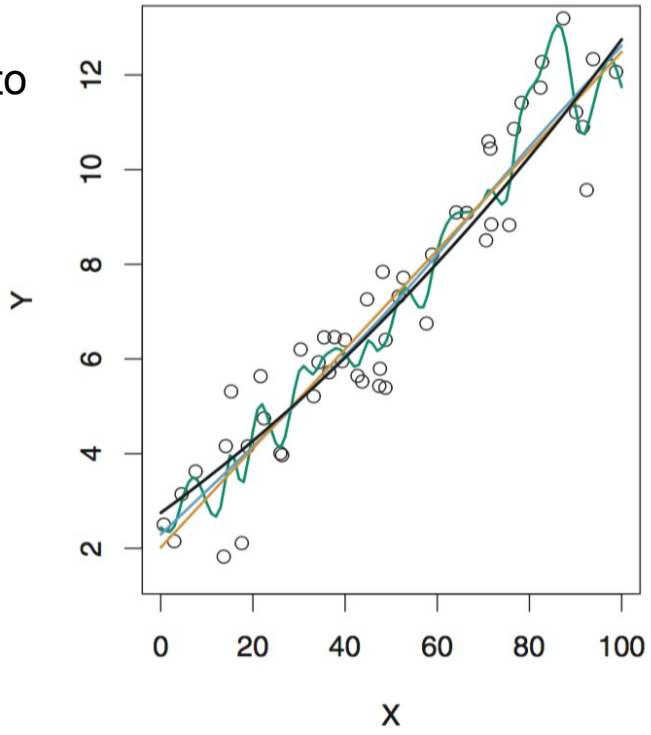
- **Bias:** error when approximating training data
- **Variance:** variability in error when approximating test data

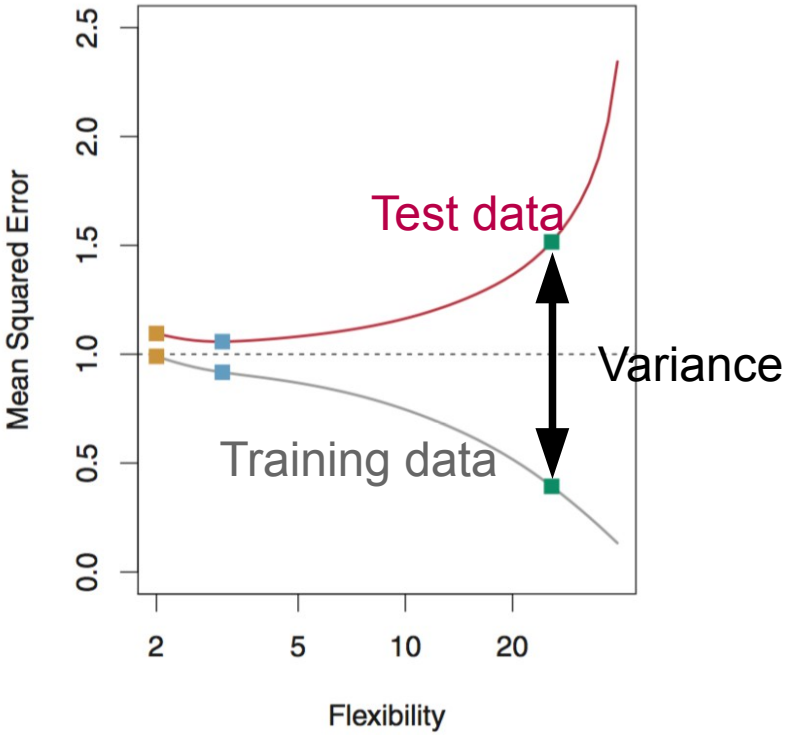
function used to simulate data

highly flexible smoother

little flexible smoother

linear regression



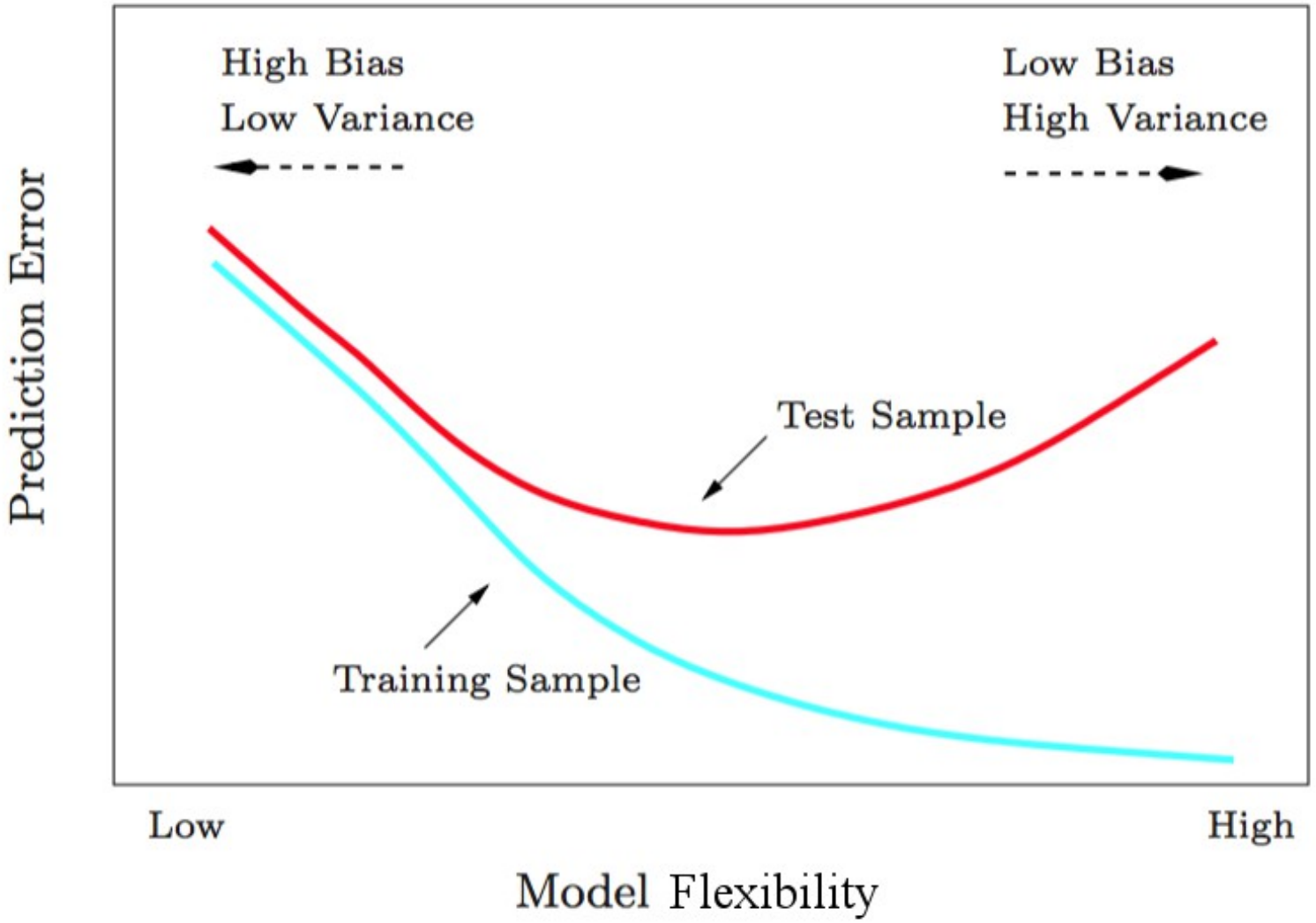


Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will start to increase from some point

Taken from James *et al.* 2013: 33

Bias-variance trade-off

Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will start to increase from some point → Optimise combined error



Taken from Hastie, Tibshirani and Friedman 2011: 38

Linear model with multiple predictors

Research goal: Explanation (identify important explanatory variables)

Example: Which variable(s) do best explain the response of different groups of organisms?

Table 2. Environmental Variables Selected in Linear Model Building with Highest Explanatory Power for the Response Variables Using Explained Variance (r^2) and the Akaike Information Criterion (AIC) as Goodness of Fit Measures

response variable	log mTU_{DM}	T (°C)	conductivity (μS/cm)	turbidity (NTU)	r^2	AIC
$SPEAR_{pesticides}$	x				0.67	−34
SIGNAL	x				0.36	98
bacteria[a]						
flagellates[a]		x	x		0.49	434
ciliates[a]		x		x	0.59	209
amoebas[a]				x	0.78	200

Schäfer et al. 2011

Multiple linear regression model

- Extension of simple linear regression model, we assume true relationship is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$

  $\rightarrow$ Classical definition for case i:

  $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + \varepsilon_i$ with $\varepsilon \sim \text{Normal}(0, \sigma)$

- Using sample date, we estimate β's (b's = regression coefficients) to obtain estimates for y:

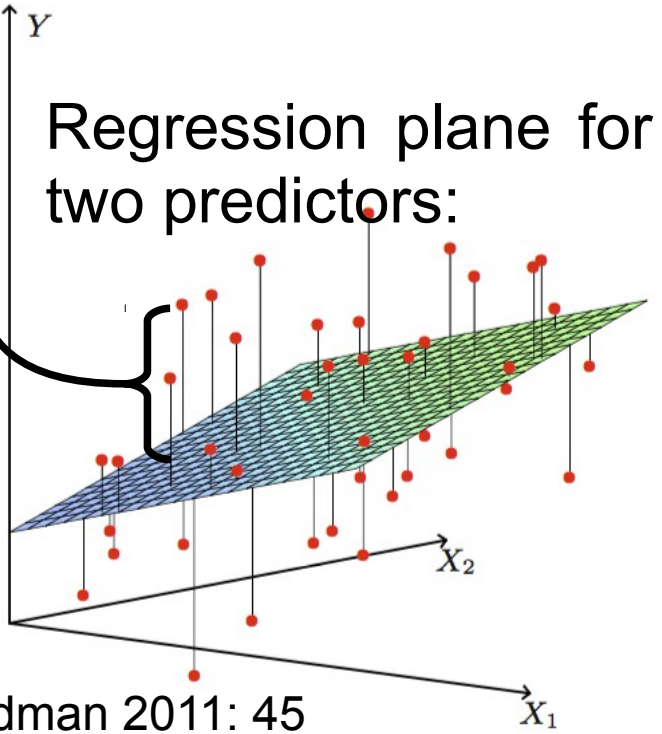
  $\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}$

- Remember: Residual e_i defined as:

  $e_i = y_i - \hat{y}_i$

- Model fitting through minimising the squared sum of residuals (RSS):

$$\text{Find} \arg \min_{b_0, b_1, b_2, \ldots, b_p} \sum_{i=1}^{n} \left(y_i - \left(b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p} \right) \right)^2$$



Taken from Hastie, Tibshirani and Friedman 2011: 45

Case study: Ostracods

Which patterns and factors control the diversity of marine arctic ostracods?

136 ostracod samples from different regions
10 explanatory variables

<u>Aim:</u> Identify most important explanatory variables for diversity of marine ostracods.
→ For explanation search for most parsimonious model

"It is futile to do with more things that which can be done with fewer"

http://www.phdcomics.com/comics/archive.php?comicid=1237

Modelling scheme (mainly for explanation)

Which variables should be included in the multiple regression model?

Full: Model 1: *Diversity* ~ *Var a, Var b, Var c, Var d, Var e*

Reduced: Model 2: *Diversity* ~ *Var a, Var b, Var c*

Model 3: *Diversity* ~ *Var a, Var b, Var c, Var d*

⋮

Model *n*: *Diversity* ~ *Var b, Var d, Var e*

Strategies

- Compare pre-specified models
- Best subset model selection
- Stepwise model selection
- Shrinkage methods

Quantitative model comparison via goodness of fit measures

Best-fit model

- model diagnostics
- model validation

Goodness of fit (GOF) measures

- R^2 or adj. R^2
  - R^2 increases with each additional variable in model (also noise)
  - adj. R^2 should be preferred for model comparison, because it penalises for additional variables
- Information theoretic goodness of fit measures for linear model:

$$\mathrm{AIC} = n \log\left(\frac{RSS}{n}\right) + 2\,p + const.$$

n = sample size

p = parameters in model

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2\,p\,(p+1)}{n-p-1} \qquad \mathrm{BIC} = n \log\left(\frac{RSS}{n}\right) + \ln(n)\,p + const.$$

  - The lower the value, the better the model
- For prediction: Cross-validation with MSPE

Model selection strategies

How to identify the best-fit model?

- Ideally: Comparison of a limited number of *a priori* specified models (based on knowledge)
- Traditionally used: 1) best subset and 2) stepwise model selection

1) Best subset: Compute all 2^p (p = number of parameters) models (w/o interactions) → computationally demanding

2) Stepwise model selection requires start model, computes all models for next step (inclusion or exclusion of variable) and selects best model. Algorithm is repeated until change of included variables would reduce model fit.

- Stepwise selection procedures: backward (variable elimination), forward (variable inclusion), both (combined)

Stepwise model selection

- Can be linked to the assessment of hypotheses:
  - Partial F-test for difference in explained variance between models:

$$\frac{\left(RSS_{\text{reduced model}} - RSS_{\text{full model}}\right) / \left(DoF_{\text{reduced model}} - DoF_{\text{full model}}\right)}{RSS_{\text{full model}} / DoF_{\text{full model}}}$$

  - If models nested and differ only by one predictor, partial F-test is equivalent to t-test for this predictor with H_0: $\beta = 0$
    Remove variable if H_0 not rejected/seems unlikely
  - Multiple inference (e.g. multiple tests on same data or tests on subset of data selected in light of data) leads to inflation of p-values (computed p-values biased low)
    see: Taylor & Tibshirani (2015) PNAS 112: 7629
  - should only be considered for data sets with few variables (< 5) and a high $n:p$ ratio (> 20)
- Can be linked to information-theoretic criteria (AIC, BIC)

Problems of stepwise model selection

Problems include (see Harrell 2015: 68):

- R^2 values biased high $\rightarrow$ Bias variance trade-off
- Standard errors and confidence intervals too low/narrow
- Regression coefficients biased high, require shrinkage
- Collinearity renders variable selection arbitrary
- Allows to not think about the problem

"Let the computer find out" is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting"
(Burnham and Anderson, 2002)

Problems generally apply to the stepwise modelling strategy, irrespective of GOF
(Murtaugh 2014 *Ecology* 95: 611; Harrell 2015: 69)

(Partial) fixes

- Modify stepwise approach or related results:
  - correction of *p*-values for sequential testing (Fithian 2015 *ArXiv e-prints*)
  - employ bootstrapping or cross-validation on all steps of model selection
    (but see Harrell 2015: 70f, Austin 2008 *J Clin Epidem*)
  - apply shrinkage factor(s) *c* to regression coefficients, which is/are estimated via CV:

Global shrinkage factor

$$b_0^s = (1-\hat{c})\bar{y} + \hat{c}\, b_0$$
$$b_j^s = \hat{c}\, b_j;\ \ j = 1,\ldots,p$$

Parameterwise shrinkage factor

$$b_0^s = (1-\hat{c}_0)\bar{y} + \hat{c}_0 b_0$$
$$b_j^s = \hat{c}_j b_j, j = 1,\ldots,p$$

- Use shrinkage method such as the LASSO (Least Absolute Shrinkage and Selection Operator)

Shrinkage method: LASSO

- Ordinary least square regression:

$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j}\right)^2$$

- Linear regression with LASSO:

$$\arg\min_{b_0, b_1} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j}\right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

Other formulation:

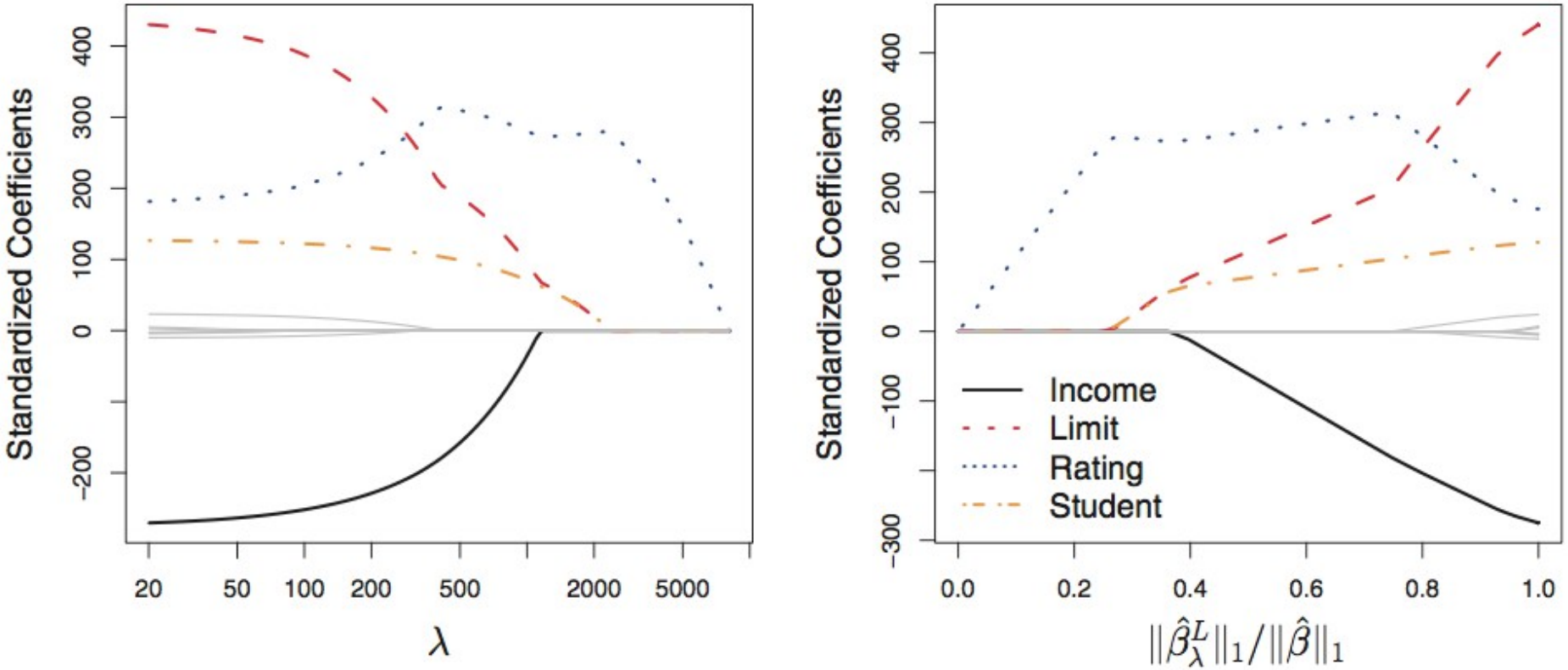
$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j}\right)^2 \text{ subject to } \sum_{j=1}^{p} |b_j| \leq s$$

- Simultaneous selection of variables and estimation of (shrinked) regression coefficients

Shrinkage method: LASSO

$$\underset{b_0, b_1}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

Example plots



- How do we identify the optimal λ? → Cross-validation (CV)

Taken from James *et al.* 2013: 220

LASSO extensions

- When does the LASSO not capture the true model?
  - Case 1: Model with several predictors, most or all relevant. → LASSO likely shrinks small regression coefficients to zero (particular of collinear predictor(s)).
  - Case 2: Model with many predictors, only few relevant. → Optimizing λ regarding prediction (in CV) can lead to selection of noise variables.
  - Alternative: Stability selection.
  - Case 3: High correlation among relevant predictors → LASSO likely selects only one. Alternative: Ridge regression or Elastic net.
  - Case 4: High correlation between relevant and irrelevant predictors. → LASSO may select irrelevant predictor(s). Alternative: Adaptive LASSO.
- Sparsity and absence of collinearity as crucial factors

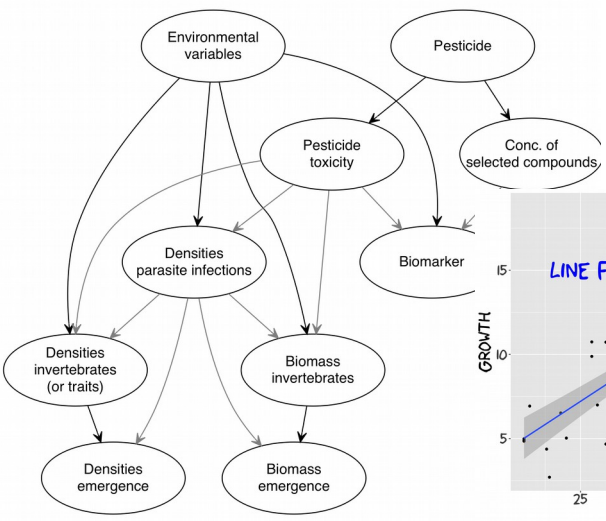
Dealing with small sample sizes

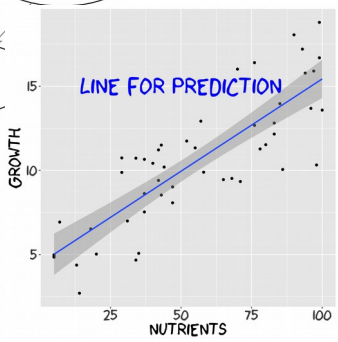
- $\text{sqrt}(n)/p$ > 1; extreme cases (e.g. genetic data): $n < p$
- OLS regression and LASSO unreliable, several modelling approaches not applicable for $n < p$ (e.g. backward elimination)
- Approaches to deal with small sample sizes:
  - Reduce parameters manually: remove variables based on scientific understanding, very low variability or narrow distribution, and missing values
  - Reduce parameters through redundancy techniques: statistical algorithms before modelling that directly reduce number of variables or aid in removal of variables e.g. variable clustering, principal component analysis (PCA)
  - Select alternative model: Elastic net

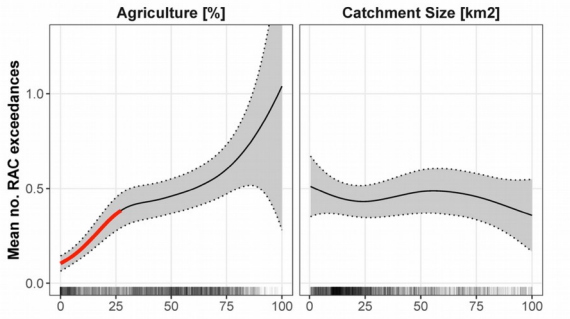
Lecture with openly accessible code, tutorials, slides, video lectures – under construction but covering several issues of modern data analysis

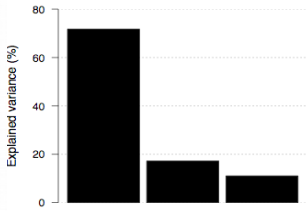
Advanced analysis of ecological data

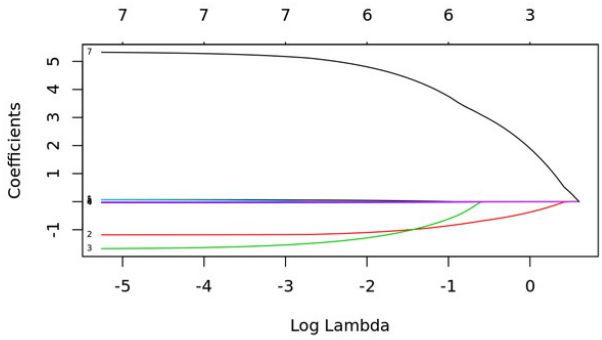
SEFS, Zagreb 2019











Ralf B. Schäfer, Andreas Scharmüller, Moritz Link

UNIVERSITÄT KOBLENZ · LANDAU

These slides and notes are part of the course materials for the workshop “Advanced data analysis” at SEFS 2019. 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

For some of the literature references, see the literature list here: https://github.com/rbslandau/Data_analysis/blob/master/Literature_commented.pdf

The slides from the first part have been taken from the lecture “Tools for complex data analysis”: https://github.com/rbslandau/Data_analysis

Short intro: Ralf Schäfer

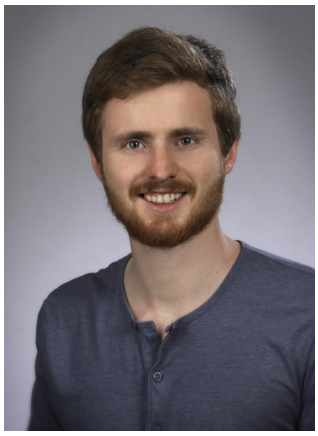
- Prof for Quantitative Landscape Ecology @UKL
- Phd @ UFZ, Leipzig; Postdoc @ RMIT, Australia
- Teaching: Data analysis; GIS; Environmental Modelling; Environmental Philosophy
- Current research projects related to:
  - Community ecology of freshwater invertebrates and microorganisms
  - Response of freshwater ecosystems to different (anthropogenic) stressors (e.g. pollution)
  - Trophic linkages between aquatic & terrestrial systems
- Primarily field studies/experiments and data analyses/ modelling

www.landscapecology.uni-landau.de

@LandscapEcology

Short intro: Andreas Scharmüller

- PhD student Quantitative Landscape Ecology
- Environmental Sciences + Ecotoxicology
- Teaching:
  - Statistics, GIS
- Research:
  - Effects and distribution of pesticides in freshwaters
  - Ecotoxicology
- R programming:
  - Package author: standartox (in preparation)
  - Package contributions: webchem



www.landscapecology.uni-landau.de @andschar

Short intro: Moritz Link

- PhD student, Quantitative Landscape Ecology
- M.Sc. Ecotoxicology @ Uni Koblenz Landau
- Teaching:
  - Course assistence in multivariate statistics
- Research:
  - Data analysis
  - Surface water monitoring
  - Ecosystem services of aquatic fungi

link@uni-landau.de

@LandscapEcology

Course Organisation

10:00-10:15 Short intro & course organisation, Software preparation

10:15-12:00 Model selection for (G)LMs

13:00-14:30 Generalized additive models (GAMs)

15:00-16:30 Structural equation models (SEMs)

16:30-16:45 Course evaluation

Course material:
https://github.com/andreasLD/workshop-sefs11

Course structure: intro – demo – hands on exercises

Part I: Model selection in (G)LMs

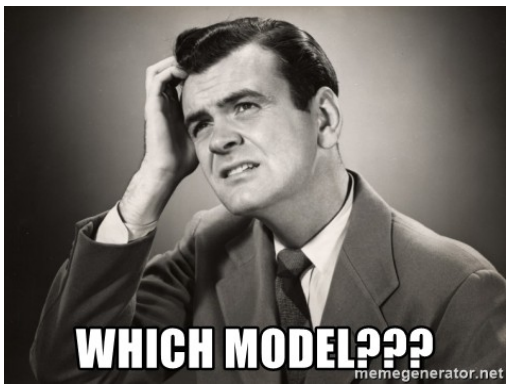
Model 1: *Diversity ~ pH, Temp., NO_3, TU, Cond.*

Model 2: *Diversity ~ TU, Temp., NO_3*

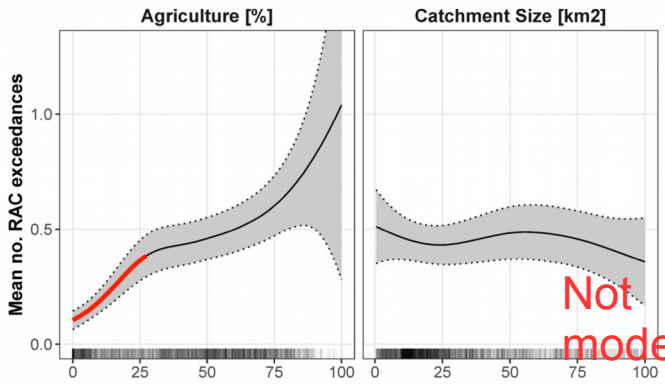
Model 3: *Diversity ~ Temp., pH, TU, NO_3*

⋮

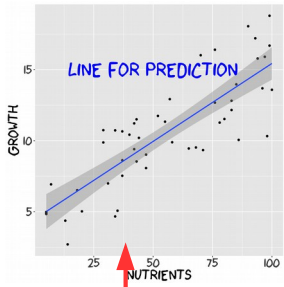
Model n: *Diversity ~ TU, pH*



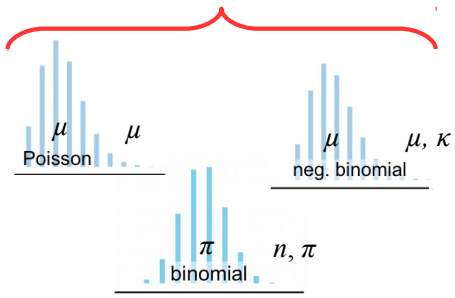
Part II: GAMs



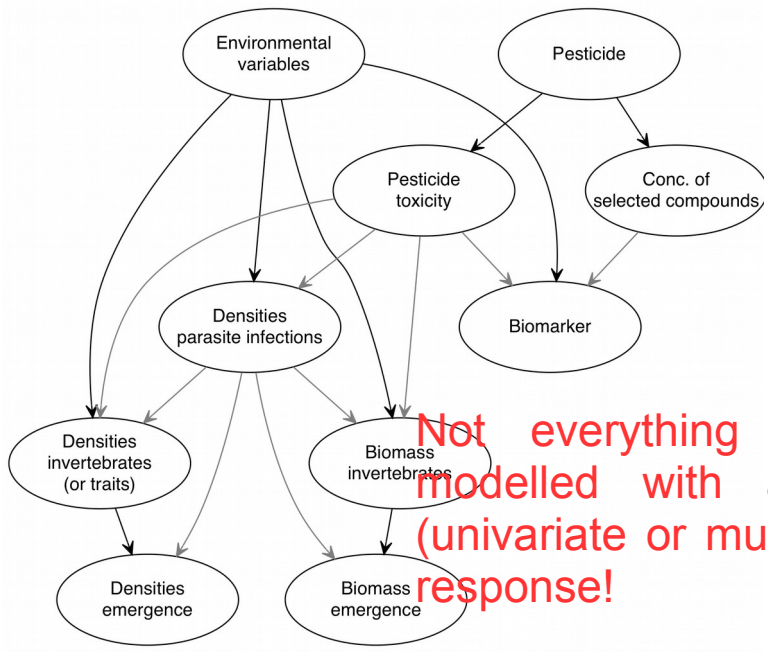
Szöcs et al. 2017 *Environ. Sci. Technol.*



Not everything can be modelled with a linear or generalized model!



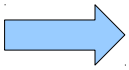
Part III: SEMs



Not everything can be modelled with a single (univariate or multivariate) response!

Your background

- 95% beginner or intermediate R users and data analysts
- ~70% have used (G)LMs, 20% lack knowledge
- 50% don't know GAMs, 40% have a little bit of prior knowledge
- > 80% don't know SEMs

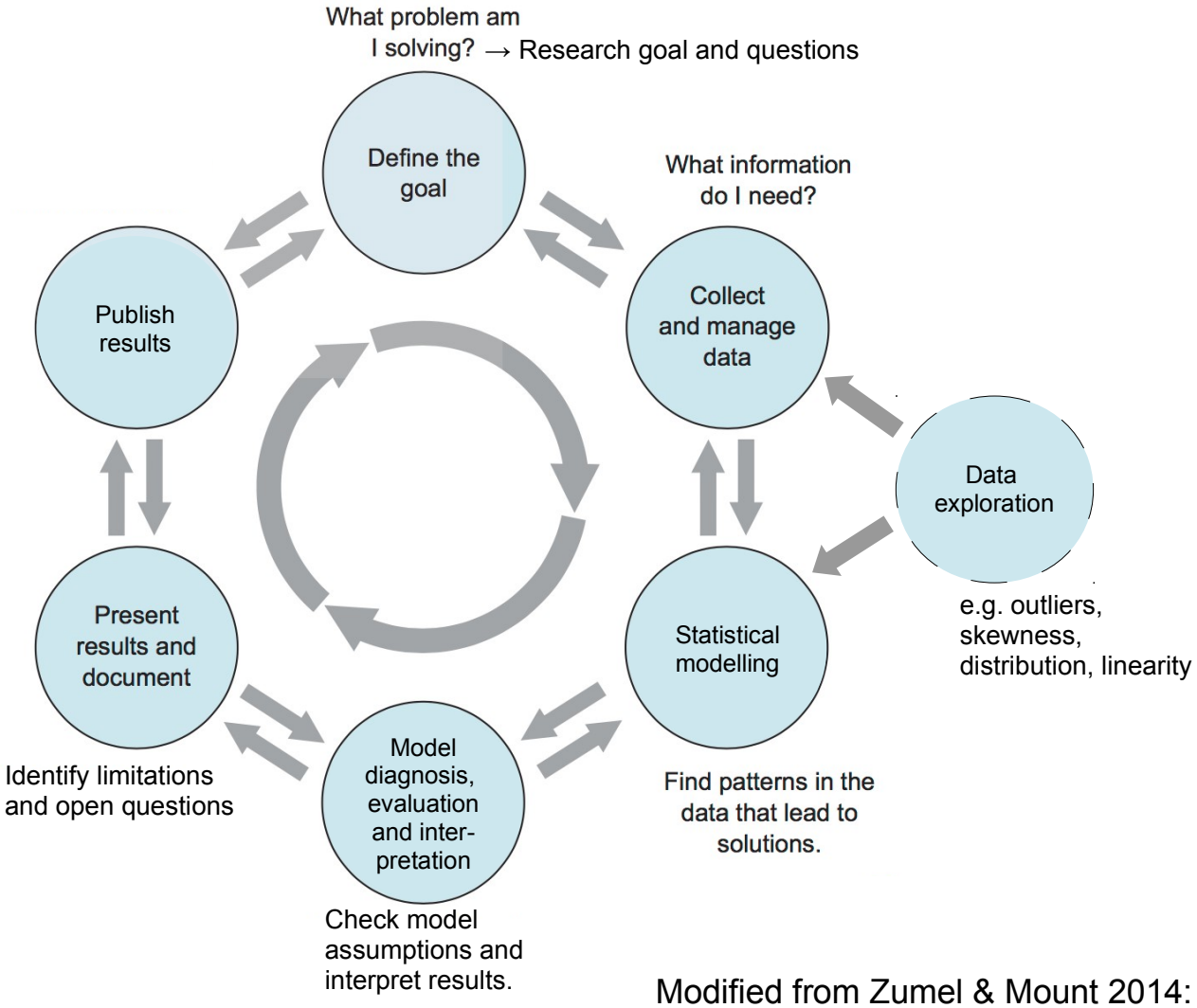


Very brief intro into (G)LMs, more extensive intro into GAMs and SEMs

Block I

Model selection for Linear and Generalised Linear models

Data analysis cycle



11

Modified from Zumel & Mount 2014: 6

Zumel N. & Mount J. (2014) Practical data science with R. Manning Publications Co, Shelter Island, NY.

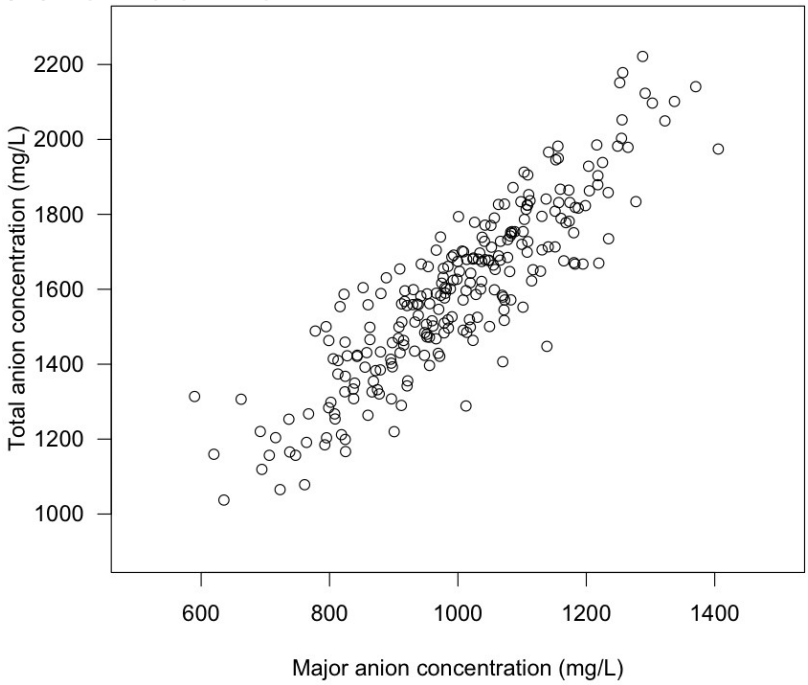
Data exploration visualised with dashed line as it will depend on the research context if and when data exploration is conducted. However, most frequently data exploration (e.g. descriptive statistics such as data summaries) is employed before statistical modelling to aid in model selection and to identify errors or outliers. Moreover, the goal of some studies is exploration and eventually no statistical modelling is done. In this case, data exploration would directly lead to the presentation of results. Conversely, in case that a clear research hypothesis has been established before data collection and the data is known to be free from outliers or errors, data exploration may be unnecessary before statistical modelling. Still, the tools related to data exploration will be needed to check model assumptions.

Case study: Water concentrations

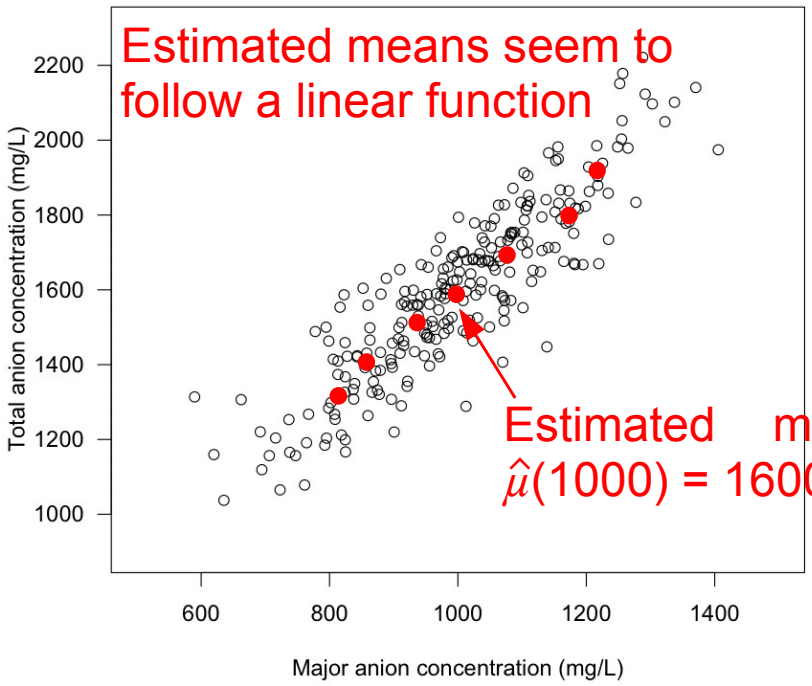
Research question: Can we predict the total anion concentration in water from the concentration of major anions (Cl^-, SO_4^{2-}, PO_4^{3-})?

Study: Samples of total anion and major anion concentrations from 250 streams.

https://upload.wikimedia.org/wikipedia/commons/1/11/Water_resources%2C_taking_a_water_sample.jpg



Predicting Y from X with a linear function



We assume a linear function of the true population $\mu(X)$:

$\mu(X) = \beta_0 + \beta_1 X$ from which follows that: $Y = \beta_0 + \beta_1 X + \varepsilon$

β_0 and β_1 : regression coefficients

13

(Simple) Linear regression model

Assuming that the true relationship is a linear function of the form $Y = \beta_0 + \beta_1 X + \varepsilon$, we can use sample data to obtain estimates of β_0 and β_1, denoted as $\hat{\beta}_0 = b_0$ and $\hat{\beta}_1 = b_1$, and subsequently predict $\hat{Y}$:

$\hat{Y} = b_0 + b_1 X$ for realisations of X we can rewrite this to:

$\hat{y}_i = b_0 + b_1 x_i$

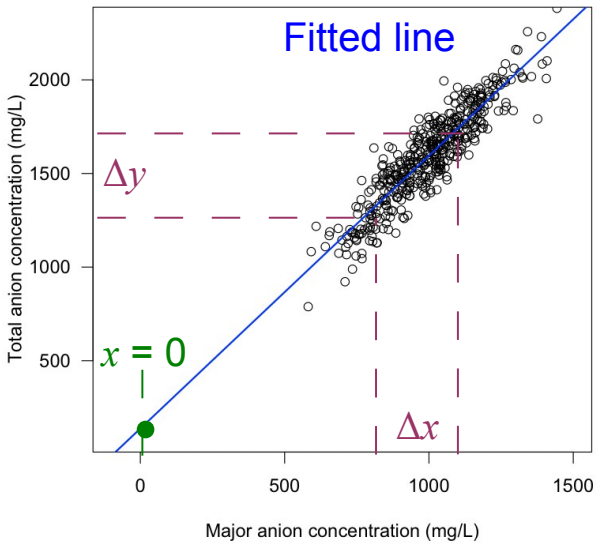
What are β_0 and β_1?

$\beta_0 = E(Y|X = 0)$ "intercept"

$\beta_1 = \frac{dy}{dx}$ "slope"

$b_0 = 138$

$b_1 = 1.5$



The model that contains only one explanatory variable/predictor is called simple linear regression model.

The intercept of a linear regression model relates to the expected value for $X = 0$. In a figure, it is the intersection of the regression line with the Y axis at $X = 0$.

The slope is represented in the figure as the change in y per change in x. In the figure, we use Δx for the difference between two given numbers for x. By contrast, dx refers to an infinitesimally small change in x.

What is the optimal regression line?

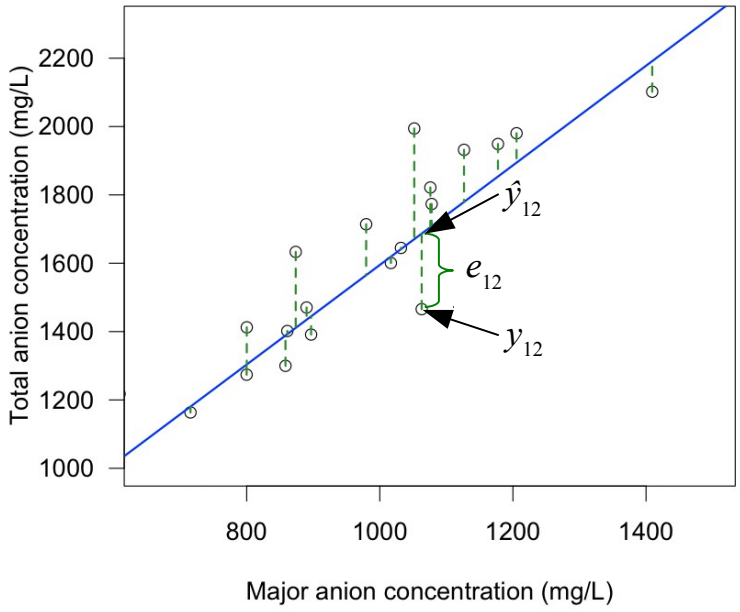
We defined for prediction: $\hat{Y} = f(X)$ and $Y = f(X) + \varepsilon$

$\Rightarrow Y = \hat{Y} + \varepsilon \quad \Leftrightarrow \quad \varepsilon = Y - \hat{Y}$

For sample data ($i = 1,2,3, \ldots, n$) and the regression model, we defined: $\hat{y}_i = b_0 + b_1 x_i$

We define the residual e_i as: $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$

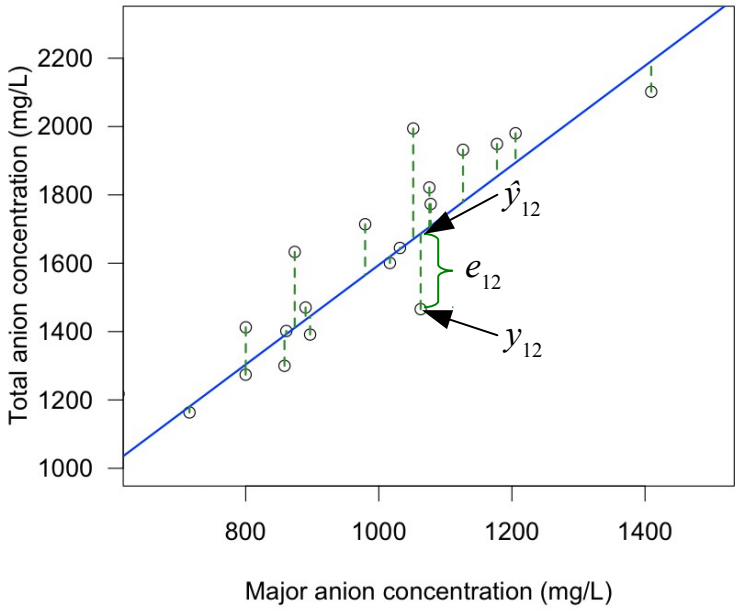
Example for observation i = 12



15

What is the optimal regression line?

Example for observation i = 12



Optimal line minimises the Residual Sum of Squares (RSS):

$$\text{RSS} = e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2$$

$$= (y_1 - (b_0 + b_1 x_1))^2 + (y_2 - (b_0 + b_1 x_2))^2 + \ldots + (y_n - (b_0 + b_1 x_n))^2$$

$$= \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \Rightarrow \text{Find} \arg\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

16

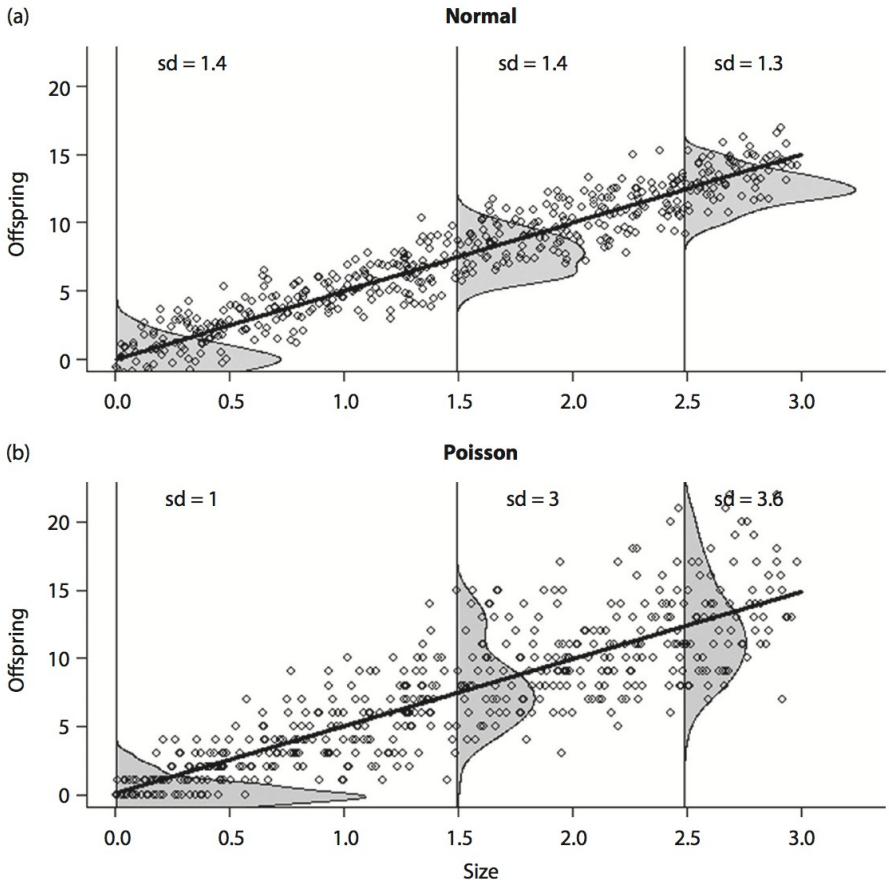
We have defined ε as the statistical error. However, we can only determine the statistical error, if we know the true linear relationship, i.e. the true values for β_0 and β_1. Given that we obtain our predictions from the estimates b_0 and b_1, we cannot determine the true error but only an estimate of the error, called residual e. Consequently, we call the sum of the squared estimates of the errors the residual sum of squares (RSS). Other terms used are sum of squared residuals (SSR) and sum of squared residual errors (SSE). Sometimes the misleading term sum of squared errors (SSE) is used, but this should be avoided, because we only deal with estimates of errors.

Note also that in some text books the residuals are simply called error and it is not distinguished between the population error and the sample estimate of the error (residual).

Due to the focus on the minimisation of the sum of squares, the model is also called *ordinary least squares* (OLS). The term "ordinary" was added to distinguish the model from the many others, later developed, relying on least squares approaches (e.g. weighted least squares, generalised linear models).

Extending the linear model: Motivation

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



17

Buckley 2015: 134

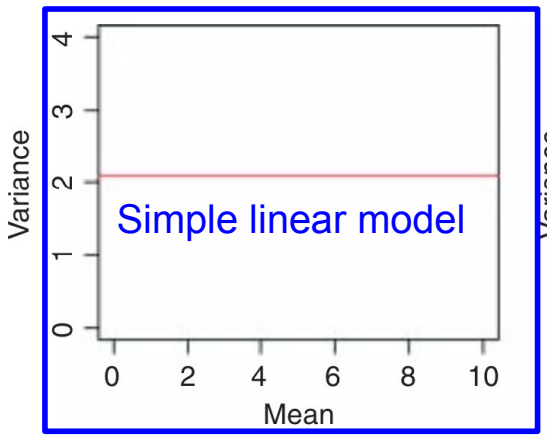
The top figure shows normally distributed residuals for the relationship between the number of offspring and body size. Although count data may approach a normal distribution in case of large data sets, the displayed residuals, which have been simulated, are not plausible, as they are associated with 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.

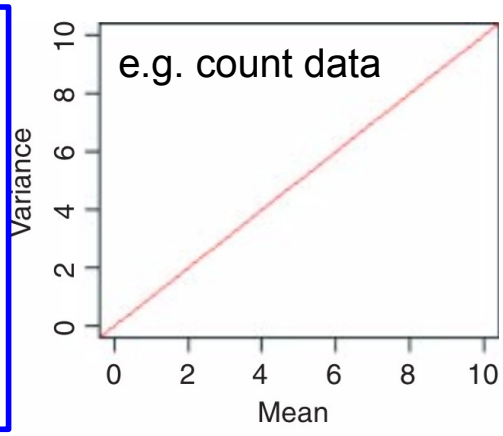
The figure below shows residuals that have been simulated using a Poisson distribution. Count data are typically Poisson distributed and we can see that the variance is not constant but increases with the mean, as for the logistic regression.

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

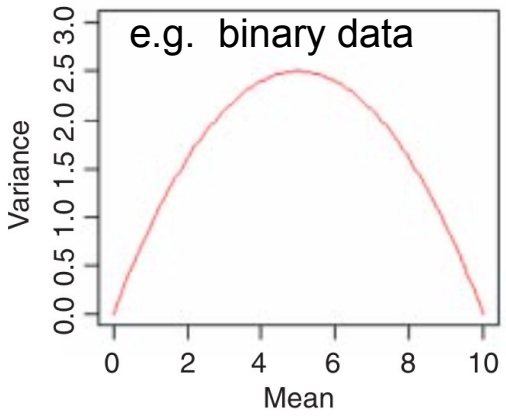
Modelling the mean-variance relationship

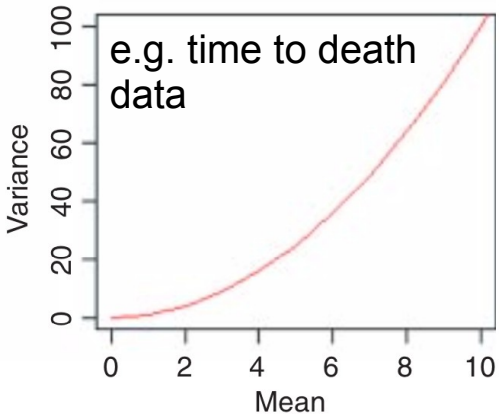
Idea: Express variance as a function of the mean!





taken from Crawley 2012: 557





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

19

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. 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.

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

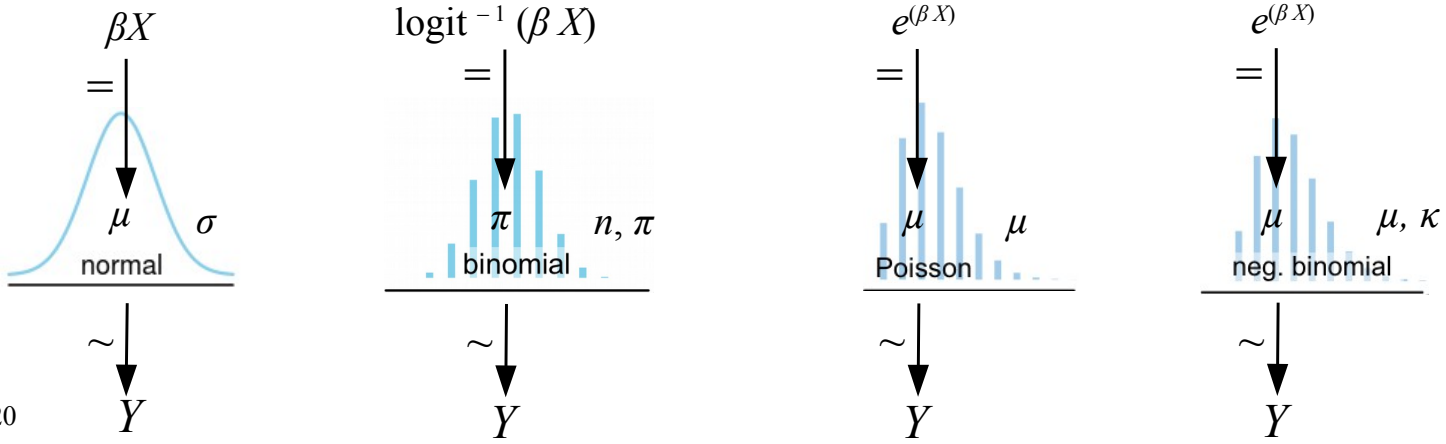
$Y \sim \text{Bin}(n, \pi)$
$E(Y) = \pi$
$\text{Var}(Y) = \frac{\pi(n-\pi)}{n}$
$\text{logit}(\pi) = \beta X$
$\varepsilon = y - \pi$

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 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 before. 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).

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.

Cross-validation (CV)

- **Aim:** Evaluate predictive accuracy of a fitted model
- Can be checked by predicting (known) responses from independent data sets (that were not used in model fitting) → Rare case
- **Idea:** Split the available data into training and test set and predict (known) observations in test set with a model fitted on the training data
- Algorithm:
  1. Draw *k* random samples without replacement from data
  2. For each *k:*
     1. Fit the model to the other *k*-1 parts
     2. Predict *k* from model and calculate the prediction error
  3. Calculate mean prediction error over the *k* estimates

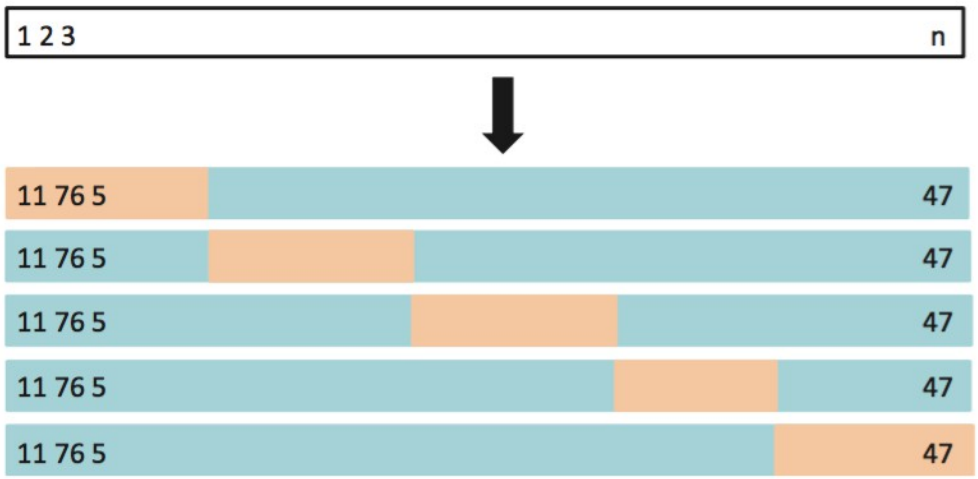
21

Predictive accuracy measures the accuracy of predictions for new data.

CV is typically used in validation, but can also be used as goodness-of-fit measure to guide parameter estimation (see shrinkage methods later).

Cross-validation (CV)

Example: $k = 5$



- Problem of choosing *k*:
  - *k* = n (Leave-one-out CV predicts each observation from all others) → low bias, but high variance
  - *k* = 2 (split data into half) → low variance, but high bias
- *k* typically set to 5 or 10

Taken from James *et al.* 2013: 181

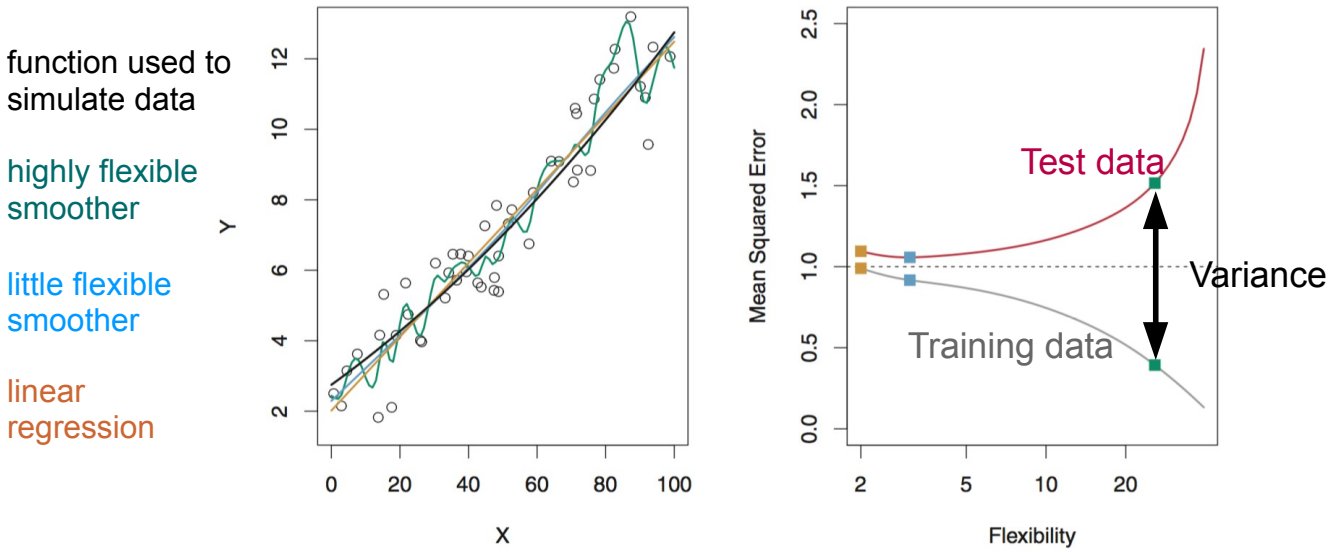
The bias-variance trade-off will be discussed in a more general context hereafter. Here, we discuss it with respect to the prediction accuracy, i.e. using cross validation to estimate the prediction accuracy. There is a trade-off between bias (error when estimating the 'true' prediction accuracy of the sample data) and variance (variability of the error when estimating new data). If we use a major fraction of the data (extreme case: $k = n$, where we use n-1 observations) in model fitting, the error of estimating the prediction accuracy of the full data is probably very low (low bias). However, the variability of the error when predicting a few (or only one for $k = n$) observations from different training sets is most likely high, which translates to a high variance. Conversely, if we use only half of the data ($k = 2$) in model fitting, we decrease the variance. In other words, the error when predicting the test set is most likely similar for the two training sets. But this comes at the cost of bias. In the case of $k = 2$, we are estimating the predictive accuracy from only a fraction of the data, whereas in practice all observations will be used in prediction. The prediction accuracy estimated from the fraction of the data is likely to differ (i.e. lower or higher) from that of the complete data set, i.e. exhibit bias. Thus, the bias increases when the relative size of the training set in CV decreases.

k is typically set to 5 or 10, i.e. the data is partitioned in 5 or 10 groups during CV as a compromise between bias and variance. Leave-one-out CV is considered less reliable than 5- or 10-fold CV (see Harrell 2015: 172).

Bias-variance trade-off

Definition in context of model validation:

- **Bias:** error when approximating training data
- **Variance:** variability in error when approximating test data



Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will start to increase from some point

Taken from James *et al.* 2013: 33

23

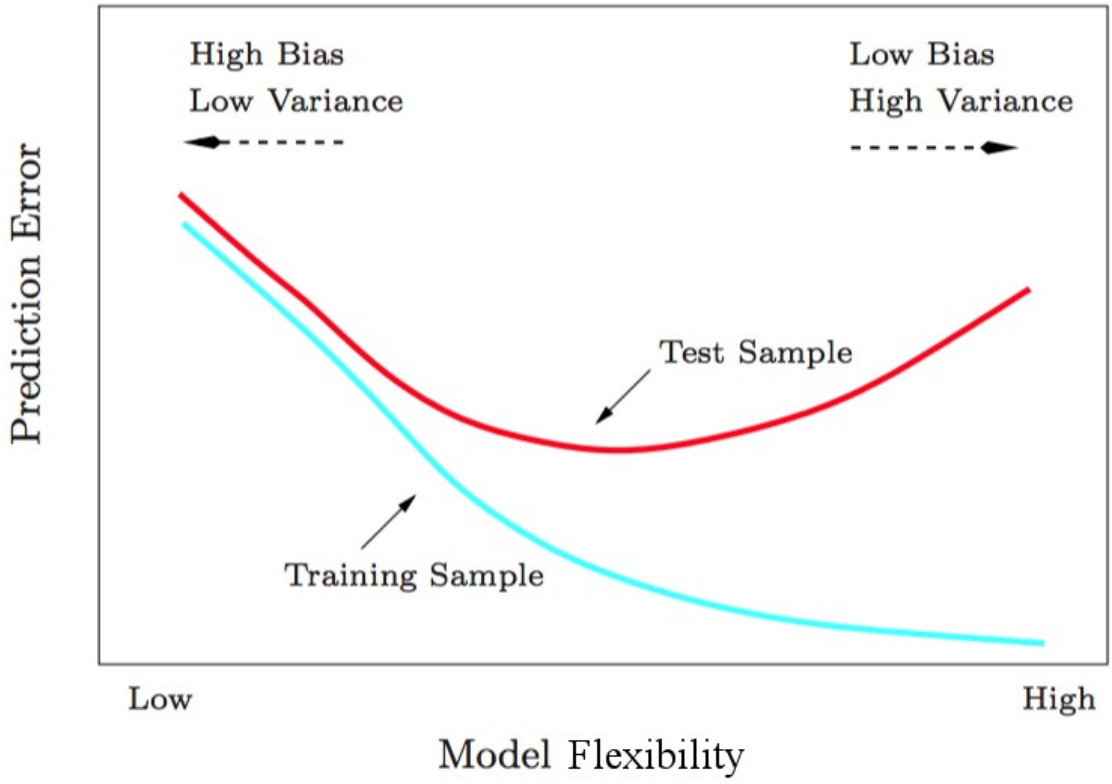
The left figure displays the fit of different models to data originating from the function plotted in black.

The models rank regarding bias: linear regression > little flexible smoother > highly flexible smoother.

Regarding variance (see right figure), the ranking is: highly flexible smoother > little flexible smoother > linear regression.

Bias-variance trade-off

Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will start to increase from some point → Optimise combined error



Taken from Hastie, Tibshirani and Friedman 2011: 38

For a mathematical derivation of the bias-variance trade-off see Matloff(2017): 48f.

24

Linear model with multiple predictors

Research goal: Explanation (identify important explanatory variables)

Example: Which variable(s) do best explain the response of different groups of organisms?

Table 2. Environmental Variables Selected in Linear Model Building with Highest Explanatory Power for the Response Variables Using Explained Variance (r^2) and the Akaike Information Criterion (AIC) as Goodness of Fit Measures

response variable	log mTU_{DM}	T (°C)	conductivity (μS/cm)	turbidity (NTU)	r^2	AIC
$SPEAR_{pesticides}$	x				0.67	−34
SIGNAL	x				0.36	98
bacteria[a]						
flagellates[a]		x	x		0.49	434
ciliates[a]		x		x	0.59	209
amoebas[a]				x	0.78	200

25

Schäfer et al. 2011

As stated before, the goal of explanation becomes much more relevant with multiple explanatory variables that, in many cases, offer competing explanations for the response. Hence, the goal is to identify the explanatory variable or variables with the highest explanatory power.

In the example, we measured several stressors that may explain an ecological index or the total abundance of an organism group, and our research goal was to identify the variable(s) with the highest explanatory power for the index. The study was conducted in 24 streams in South-East Australia and was to designed to detect potential relationships between pesticides and invertebrates as well as microorganisms. The SPEAR index has been developed to indicate pesticide effects in invertebrate communities, whereas the SIGNAL index has been developed to indicate general ecological degradation. The log mTU is a proxy for the pesticide toxicity in a sampling site, calculated for pesticides in 24 South-East Australian streams.

Schäfer R.B., Pettigrove V., Rose G., Allinson G., Wightwick A., von der Ohe P.C., et al. (2011) Effects of pesticides monitored with three sampling methods in 24 sites on macroinvertebrates and microorganisms. Environmental Science & Technology 45, 1665–1672.

Multiple linear regression model

- Extension of simple linear regression model, we assume true relationship is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$

  → Classical definition for case i:

  $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + \varepsilon_i$ with $\varepsilon \sim \text{Normal}(0, \sigma)$

- Using sample date, we estimate β's (b's = regression coefficients) to obtain estimates for y:

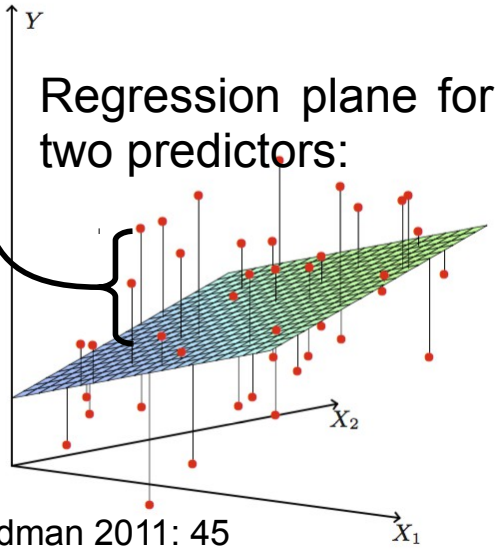
  $\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}$

- Remember: Residual e_i defined as:

  $e_i = y_i - \hat{y}_i$

- Model fitting through minimising the squared sum of residuals (RSS):

$$\text{Find} \arg \min_{b_0, b_1, b_2, \ldots, b_p} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}))^2$$



Taken from Hastie, Tibshirani and Friedman 2011: 45

26

Case study: Ostracods

Which patterns and factors control the diversity of marine arctic ostracods?

136 ostracod samples from different regions
10 explanatory variables

<u>Aim:</u> Identify most important explanatory variables for diversity of marine ostracods.
→ For explanation search for most parsimonious model

"It is futile to do with more things that which can be done with fewer"

http://www.phdcomics.com/comics/archive.php?comicid=1237

27

The ostracod picture has been taken from: https://www4.uwm.edu/fieldstation/naturalhistory/bugoftheweek/images/ostracod12-10.jpg

We assume that the relationship between explanatory variables and the species richness is largely linear (or quadratic) in the case study – see for details:
Yasuhara M., Hunt G., van Dijken G., Arrigo K.R., Cronin T.M. & Wollenburg J.E. (2012) Patterns and controlling factors of species diversity in the Arctic Ocean. *Journal of Biogeography* 39, 2081–2088.

Occam's razor may be translated to our situation as: Given a similar predictive or explanatory power, models with fewer variables are generally better than those with more variables.

Harrell (2015: 70, 95ff) argues that the full model (including all possible predictors) typically provides meaningful p-values, confidence intervals and parameter estimates and has the highest predictive power. Thus, model parsimony is primarily relevant when we aim to identify the most important variables. Notwithstanding, when building models for prediction, we also prefer the model with fewer variables to one with more variables for a similar predictive power. See also Matloff (2017): 339ff.

Modelling scheme (mainly for explanation)

Which variables should be included in the multiple regression model?

Full: Model 1: *Diversity ~ Var a, Var b, Var c, Var d, Var e*

Reduced: Model 2: *Diversity ~ Var a, Var b, Var c*

Model 3: *Diversity ~ Var a, Var b, Var c, Var d*

⋮

Model *n*: *Diversity ~ Var b, Var d, Var e*

Strategies

- Compare pre-specified models
- Best subset model selection
- Stepwise model selection
- Shrinkage methods

Quantitative model comparison via goodness of fit measures

Best-fit model

- model diagnostics
- model validation

28

For prediction, we often can use the full model and do not need to select a modelling strategy. If we aim to determine an effect size or assess a specific hypothesis, we should have a pre-specified model and other strategies are largely irrelevant.

Why do we not assess the importance of variables from multiple simple linear regressions? This is because in simple linear regressions, important variables may be ignored that exert a high explanatory power in the presence of other variables in the model. For further explanation see:

Sun G.-W., Shook T.L. & Kay G.L. (1996) Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of Clinical Epidemiology 49, 907 – 916.

Goodness of fit (GOF) measures

- R^2 or adj. R^2
  - R^2 increases with each additional variable in model (also noise)
  - adj. R^2 should be preferred for model comparison, because it penalises for additional variables
- Information theoretic goodness of fit measures for linear model:

$$\mathrm{AIC} = n \log\left(\frac{RSS}{n}\right) + 2p + const.$$

n = sample size

p = parameters in model

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2p(p+1)}{n-p-1} \qquad \mathrm{BIC} = n \log\left(\frac{RSS}{n}\right) + \ln(n)\, p + const.$$

  - The lower the value, the better the model
- For prediction: Cross-validation with MSPE

Here, the information criteria are expressed for models subject to least square fitting. Generally, the AIC is given as: AIC = -2 log(L) + 2p, where L is the likelihood function for the parameters in the model. Similarly, the BIC is given as: BIC = -2 log(L) + p log(n). These implementations based on the concept of log likelihood are used for the Generalized Linear Model (GLM).

Note that the R function AIC() calculates a simplified version of AIC without the constant term. This constant term is -n log(n). Hence, when models for the same data are compared, this omission is justified because the constant is the same for all models and the addition of a constant does not affect the ranking of the absolute values of the models. The BIC gives higher penalty to more complex models than the AIC for $n \geq 8$ and may thus aid in selection of more parsimonious (sparser) models. The AIC tends towards over-fitting especially for smaller data sets (e.g. $n < 50$). The corrected AIC ($\mathrm{AIC_C}$) is the recommended alternative. In fact, the corrected AIC could always be used as it converges with the AIC for larger sample sizes.

The adjusted (adj.) R^2 should be preferred over the normal R^2 as it takes the number of explanatory variables p into account.

$$\mathrm{adj.}\, R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Model selection strategies

How to identify the best-fit model?

- Ideally: Comparison of a limited number of *a priori* specified models (based on knowledge)
- Traditionally used: 1) best subset and 2) stepwise model selection

1) Best subset: Compute all 2^p (p = number of parameters) models (w/o interactions) → computationally demanding

2) Stepwise model selection requires start model, computes all models for next step (inclusion or exclusion of variable) and selects best model. Algorithm is repeated until change of included variables would reduce model fit.

- Stepwise selection procedures: backward (variable elimination), forward (variable inclusion), both (combined)

30

If our aim is to estimate effect sizes or to test hypotheses, we should pre-specify a model (or a few models).

Computing all possible models represents an exhaustive search for the best regression model. This procedure is most useful when no prior knowledge on the ranking of the scientific relevance of variables is available. If the number of possible models is large, different models may have a similar GOF so that the selection of only one model is problematic. Model averaging over all models up to a certain threshold of a GOF measure can be applied in this case. A review by Grueber et al. (2011) discusses several issues associated with model selection and averaging: http://onlinelibrary.wiley.com/doi/10.1111/j.1420-9101.2010.02210.x/abstract. Cade (2015) cautions regarding the use of model averaging: http://www.esajournals.org/doi/pdf/10.1890/14-1639.1.

Best subset selection quickly becomes computationally demanding, especially if model validation would be applied to the whole process. For example, for ten-fold CV and 10 explanatory variables, 1024 * 10 models need to be fitted. The number of possible models becomes even higher if interaction terms are included. Moreover, compared to a more constrained search, best subset selection is likely to yield to a model with higher variance in prediction (see Hastie, Tibshirani and Friedman 2017: 59).

The criticism on stepwise model selection (see next slides) also applies to best subset selection.

Stepwise model selection

- Can be linked to the assessment of hypotheses:
  - Partial F-test for difference in explained variance between models:

$$\frac{\left(RSS_{\text{reduced model}} - RSS_{\text{full model}}\right) / \left(DoF_{\text{reduced model}} - DoF_{\text{full model}}\right)}{RSS_{\text{full model}} / DoF_{\text{full model}}}$$

  - If models nested and differ only by one predictor, partial F-test is equivalent to t-test for this predictor with $H_0: \beta = 0$
    Remove variable if H_0 not rejected/seems unlikely
  - Multiple inference (e.g. multiple tests on same data or tests on subset of data selected in light of data) leads to inflation of p-values (computed p-values biased low)
    see: Taylor & Tibshirani (2015) PNAS 112: 7629
  - should only be considered for data sets with few variables (< 5) and a high $n:p$ ratio (> 20)
- Can be linked to information-theoretic criteria (AIC, BIC)

31

n = sample size, p = parameters in model

Murtaugh (2014; Ecology 95: 611-617) pointed out that the p-value and the information-theoretic criteria (AIC, BIC) are intimately linked. Thus, they face similar problems in model comparison and selection.

The paper by Taylor & Tibshirani (2015) can be freely accessed:

http://www.pnas.org/content/112/25/7629.abstract

Problems of stepwise model selection

Problems include (see Harrell 2015: 68):

- R^2 values biased high → Bias variance trade-off
- Standard errors and confidence intervals too low/narrow
- Regression coefficients biased high, require shrinkage
- Collinearity renders variable selection arbitrary
- Allows to not think about the problem

"Let the computer find out" is a poor strategy and usually reflects the fact that the researcher did not bother to think clearly about the problem of interest and its scientific setting"
(Burnham and Anderson, 2002)

Problems generally apply to the stepwise modelling strategy, irrespective of GOF
(Murtaugh 2014 *Ecology* 95: 611; Harrell 2015: 69)

32

If stepwise model selection is used, Harrell (2015: p. 70) suggests to use backward selection, because this would perform better in the presence of collinear variables and starts with the full model, which is the only model providing accurate p-values, standard errors etc. However, backward selection cannot be used, if $n < p$ (this case is discussed later).

The issue of collinearity will be discussed in detail later.

(Partial) fixes

- Modify stepwise approach or related results:
  - correction of *p*-values for sequential testing (Fithian 2015 *ArXiv e-prints*)
  - employ bootstrapping or cross-validation on all steps of model selection
    (but see Harrell 2015: 70f, Austin 2008 *J Clin Epidem*)
  - apply shrinkage factor(s) *c* to regression coefficients, which is/are estimated via CV:

Global shrinkage factor

$$b_0^s = (1-\hat{c})\bar{y} + \hat{c}\, b_0$$
$$b_j^s = \hat{c}\, b_j;\ j = 1,\ldots,p$$

Parameterwise shrinkage factor

$$b_0^s = (1-\hat{c}_0)\bar{y} + \hat{c}_0 b_0$$
$$b_j^s = \hat{c}_j b_j, j = 1,\ldots,p$$

- Use shrinkage method such as the LASSO (Least Absolute Shrinkage and Selection Operator)

Austin (2008) found no improved performance of bootstrapping model selection compared to backward stepwise selection. Harrell (2015: 70f) discusses several drawbacks of the bootstrap approach.
Cross-Validation is similarly likely to underestimate the true variance.

The application of shrinkage factors after model selection is called post-selection shrinkage.

A simulation study found that backward stepwise elimination performed equally well as the LASSO in the identification of true predictors, particularly in conjunction with parameter-wise shrinkage (Houwelingen & Sauerbrei 2013). However, no approach performed best in all scenarios. Interestingly, backward stepwise elimination yielded often to more parsimonious (sparser) models than the LASSO (see next slides).

References

Austin P.C. (2008) Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: a simulation study. Journal of Clinical Epidemiology 61, 1009 – 1017.e1.
Fithian W., Taylor J., Tibshirani R. & Tibshirani R. (2015) Selective Sequential Model Selection. ArXiv e-prints, 1–36. http://adsabs.harvard.edu/abs/2015arXiv151202565F (an updated version can be found here: http://www.stat.cmu.edu/~ryantibs/papers/seqinf.pdf)
Houwelingen H.C. van & Sauerbrei W. (2013) Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited. Open Journal of Statistics 03, 79–102.

Shrinkage method: LASSO

- Ordinary least square regression:

$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j}\right)^2$$

- Linear regression with LASSO:

$$\arg\min_{b_0, b_1} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j}\right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

Other formulation:

$$\arg\min_{b_0, b_1} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j}\right)^2 \text{ subject to } \sum_{j=1}^{p} |b_j| \leq s$$

- Simultaneous selection of variables and estimation of (shrinked) regression coefficients

34

Using the LASSO is motivated by two problems of OLS regression:

1. Regression coefficients are biased high, leading to high variance in prediction. Shrinking of regression coefficients increases the bias (remember the bias-variance tradeoff), but reduces the variance.
2. For a large number of predictors, interpretation of the OLS result becomes tricky. Focusing on a smaller subset improves interpretation.

Through introduction of the penalty term in LASSO, the regression coefficients are shrunk. With increasing penalty (i.e. larger λ and smaller *s*, respectively) some regression coefficients are shrunk to 0, which means that the LASSO performs variable selection.

The penalty term is called ℓ_1-norm in mathematical terms, where norm is a function that assigns a value to a vector (in our case to the vector of regression coefficients) and the ℓ notation is a reference to a specific mathematical space.

A comparison of regression using LASSO and several other techniques can be found in Hastie, Tibshirani & Friedman (2017: 61ff). The LASSO is typically among the methods with the lowest prediction error. However, the optimal λ for prediction does not guarantee the selection of the most parsimonious model for explanation (Zou 2006).

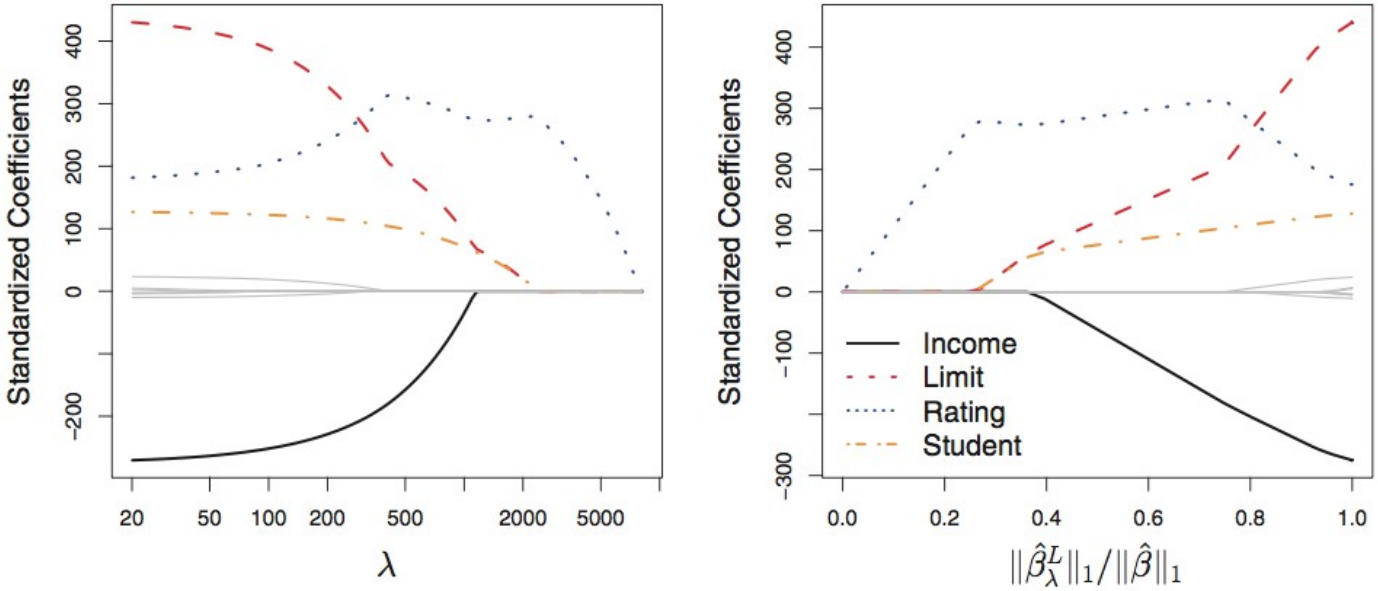
For application in R see James et al. (2013) and for further developments of shrinkage (or more precise: sparse) methods see Hastie, Tibshirani & Wainwright (2015). All these books are freely downloadable, the URLs are provided with the literature list.

Zou H. (2006). The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association 101, 1418–1429. https://doi.org/10.1198/016214506000000735

Shrinkage method: LASSO

$$\arg\min_{b_0, b_1} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

Example plots



- How do we identify the optimal λ? → Cross-validation (CV)

Taken from James *et al.* 2013: 220

For the LASSO analysis, variables are typically standardized to zero mean and standard deviation of one. The advantage is that all variables have the same units after standardization, i.e. are represented on the same scale and their importance can be ranked based on the size of the coefficient. Moreover, the response Y is centred (through subtraction of the mean), which leads to a zero intercept, i.e. b_0 can be removed.

The left plot shows the standardised regression coefficients along increasing λ on the x axis. For very low values of λ, the regression coefficients are the same as for OLS regression. As λ becomes higher, all regression coefficients are shrunk towards zero and for a very high λ, we eventually obtain the null model.

The right plot displays the ratio of the absolute sum of the standardized regression coefficients for the LASSO (i.e. ℓ^1-norm) and the absolute sum of the standardized regression coefficients from OLS. How to choose an optimal λ using the prediction error in cross validation is shown in the R tutorial.

LASSO extensions

- When does the LASSO not capture the true model?
  - Case 1: Model with several predictors, most or all relevant. → LASSO likely shrinks small regression coefficients to zero (particular of collinear predictor(s)).
  - Case 2: Model with many predictors, only few relevant. → Optimizing λ regarding prediction (in CV) can lead to selection of noise variables.
  - Alternative: Stability selection.
  - Case 3: High correlation among relevant predictors → LASSO likely selects only one. Alternative: Ridge regression or Elastic net.
  - Case 4: High correlation between relevant and irrelevant predictors. → LASSO may select irrelevant predictor(s). Alternative: Adaptive LASSO.
- Sparsity and absence of collinearity as crucial factors

The LASSO is typically used, if sparsity can be assumed, i.e. that of a large number of predictors only a small subset is relevant. In other words, the true model in the statistical population only contains a few predictors. I outline a few cases where the LASSO may not capture the true model. This may not be that relevant if the research goal is prediction, but it is if the research goal is explanation.

An alternative to the conventional LASSO approach to determine the optimal λ via the prediction error in CV, is to employ a resampling procedure such as bootstrapping, where for each λ several bootstrap samples are used to identify the relevant variables. The related method is called stability selection and identifies the most relevant variables based on their probability of selection in resampling (see Meinhausen & Bühlmann 2010 for details).

Moreover, the true model may not be captured if the relevant predictors exhibit a high intercorrelation (i.e. collinearity) or a high correlation with irrelevant variables. However, for real data where the true regression coefficients are unknown, it is difficult to evaluate whether our data falls into one of these cases outlined.

Alternatives for the case of collinearity among the predictors are ridge regression and the elastic net (Zou & Hastie 2005). Ridge regression represents a shrinkage method similar to the LASSO, but uses a quadratic penalty term called ℓ_2-norm (see James, Witten, Hastie and Tibshirani 2013: 61ff):

$$\arg\min_{b_0, b_1} \left\{ \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} b_j^2 \right\}$$

Compared to the LASSO, ridge regression can better deal with collinear variables, but does not perform variable selection (i.e. regression coefficients are not shrunk to zero). Both ridge regression and the LASSO represent special cases of the elastic net, which combines both ℓ_1 *and* ℓ_2 penalties. Ridge regression should be used if the (relevant) predictors are collinear, and all variables should remain in the model. The elastic net should be used if (relevant) predictors are collinear and variables should be shrunk to zero. Another advantage (besides accounting for collinearity) of the elastic net over the LASSO is that it can deal with low sample sizes (further details later).

The adaptive LASSO relies on individual penalties for each predictor through adaptive weights, which are typically determined via ridge regression (in case of collinearity).

Meinshausen N. & Bühlmann P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 417–473. Freely accessible at: https://pdfs.semanticscholar.org/9476/3a504ed7d835051d3e52f288c9d9e4d80e03.pdf

Zou H. & Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 301–320. Freely accessible at: https://web.stanford.edu/~hastie/Papers/B67.2%20%282005%29%20301-320%20Zou%20&%20Hastie.pdf

Zou H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101, 1418–1429. Freely accessible at: http://users.stat.umn.edu/~zouxx019/Papers/adalasso.pdf

Dealing with small sample sizes

- $\text{sqrt}(n)/p$ > 1; extreme cases (e.g. genetic data): $n < p$
- OLS regression and LASSO unreliable, several modelling approaches not applicable for $n < p$ (e.g. backward elimination)
- Approaches to deal with small sample sizes:
  - Reduce parameters manually: remove variables based on scientific understanding, very low variability or narrow distribution, and missing values
  - Reduce parameters through redundancy techniques: statistical algorithms before modelling that directly reduce number of variables or aid in removal of variables e.g. variable clustering, principal component analysis (PCA)
  - Select alternative model: Elastic net

37

For multiple regression analysis with OLS, the number of parameters p in the model should be smaller than $\text{sqrt}(n)$, see Matloff (2017) p. 441. Harrel (2015) pp. 72-74 suggests a more restrictive rule: the number of parameters p in the model should be smaller or equal than $n/15$. Consult both references for further details.

Running a PCA before regression analysis is also called principal component regression (PCR). Briefly, PCA constructs new, non-correlated (orthogonal) gradients from a data set and in case of collinearity this can help to reduce the number of variables. See Harrell (2015) pp. 79 for further details and techniques. The approach is particularly powerful, if the predictors are strongly correlated, for example, in the case of bioclimatic or water quality variables. For an application of PCR see Bhowmik & Schäfer (2015).

The elastic net combines both ℓ^1 *and* ℓ^2 penalties of the LASSO and ridge regression. It can deal with $n < p$ situations and with collinear variables. For details see Zou & Hastie (2005) and Hastie, Tibshirani & Friedman (2017: 661ff).

Another alternative model, but with a completely different modelling approach, is the Random Forest model.

Bhowmik & Schäfer (2015) Large Scale Relationship between Aquatic Insect Traits and Climate. PLoS ONE 10, e0130025. Freely accessible at: http://dx.doi.org/10.1371%2Fjournal.pone.0130025

Zou H. & Hastie T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

Lecture with openly accessible code, tutorials, slides, video lectures – under construction but covering several issues of modern data analysis

38