# Applied Statistics

IMPRS course

Dr. Jens Schumacher

11.09. - 13.09.2017



## Where to start - where to end?

There are three kinds of lies:
lies, damned lies and statistics

Statistics is the art of honesty



## Overview

## Linear Regression Model

Problem

$$\begin{array}{rcl} \text{predictor} & \Longrightarrow & \text{response} \\ \text{independent variable} & \Longrightarrow & \text{dependent variable} \\ x & \Longrightarrow & y \\ \text{metric} & & \text{metric} \end{array}$$

- prediction of response for given values of the predictor
- asymmetric situation



## Linear Regression Model

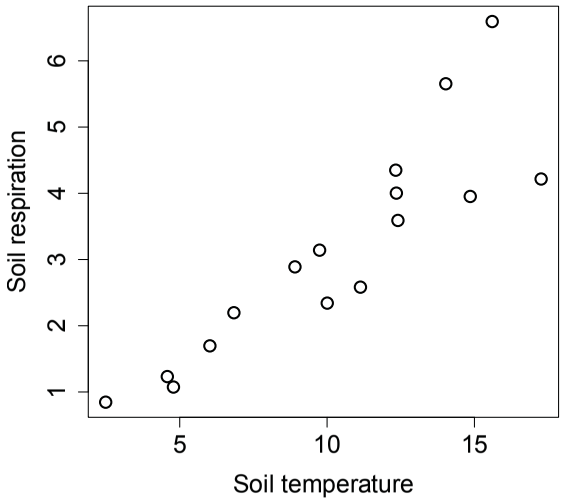
Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

$\beta_0$ – intercept
$\beta_1$ – slope
$\varepsilon_i$ – random error term



## Scatterplot



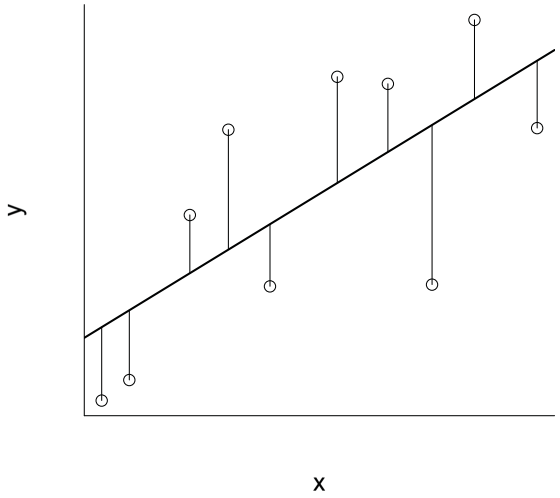



## Calculations - What do we want?

- estimate parameters of linear relation
- How accurate are these estimates?
- predict response for given values of explanatory variable
- How accurate are the predictions?
- How good is the model?

# Calculation

## Method of Least Squares




Applied Statistics
J.Schumacher

# Calculation

## Method of Least Squares

- minimize sum of squared deviations in y direction

$$\text{minimize} \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Applied Statistics
J.Schumacher

# Calculation

## Parameter estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

## Predicted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

## R square

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Applied Statistics
J.Schumacher

# Calculation

## Result

```
Call:
lm(formula = resp ~ temp, data = resp.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.28458 -0.40128  0.05184  0.19982  1.64718

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.25643    0.50740  -0.505    0.621
temp         0.33355    0.04597   7.256 4.19e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7711 on 14 degrees of freedom
Multiple R-squared:  0.7899, Adjusted R-squared:  0.7749
F-statistic: 52.64 on 1 and 14 DF,  p-value: 4.188e-06
```

## Everything about statistical tests

- Formulate a **Null hypothesis** $H_0$
- Define a **test statistic** (measuring deviation from $H_0$)
- Derive **sampling distribution** under $H_0$
- **Compare** observed value of test statistic with sampling distribution
- Reject or accept $H_0$ based on pre-chosen **significance level** $\alpha$ (usually $\alpha = 0.05$)



## Assumptions

**Linearity** relationship between predictor and response can be described by a linear function

**Independence** observations are statistically independent

**Variance homogeneity** random deviations from linearity show equal variability over whole range of observations

**Normality** random deviations follow a normal distribution



## Assumptions

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon_i \quad \text{i.i.d.} \quad N(0, \sigma^2)$$



## Independence

**MOST IMPORTANT ASSUMPTION**

- What does independence mean?
- knowing one random error term gives us **no information at all** about the other random deviations
- observations reflect the **true underlying variability**
- identification of reference population important
  - Are biomass measurements of 5 randomly selected beech trees in a stand statistically independent?
  - **YES** - if this particular stand is my reference population
  - **NO** - if I want to draw conclusions about all Central European beech stands

## Independence

Typical causes of non-independence

- observations over time - **temporal autocorrelation**
- spatial location of observations matters - **spatial autocorrelation**
- repeated observations on the same object/„experimental unit“ - **repeated measurements**
- all observed objects are equal, but some objects are more equal than others - **Clustered sampling**



## Independence

Consequences of non-independence

**Estimation of parameters**

- estimates are still unbiased
- usually only slight changes in parameter estimates
- accuracy of parameter estimates overestimated

**Prediction** prediction accuracy overestimated



## Independence

What shall I do?

- there is no way to correct for non-independence after observations are collected
- ensure independence of sampling units when you design your experiment/observations
- use statistical models which explicitely incorporate statistical dependence of observations

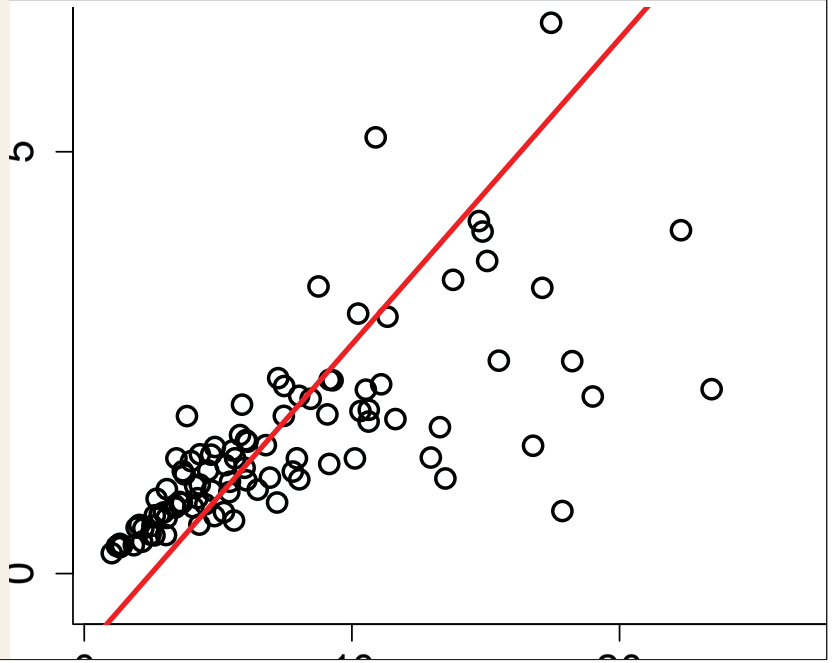


## Variance Homogeneity

What does variance homogeneity mean?

- „size“ of deviations from the model comparable over whole observation range
- typically violated if observation range is large
  - small values of response - small variability
  - large values of response - larger variability

## Variance Homogeneity





## Variance Homogeneity

Consequences of variance heterogeneity

- accuracy of parameter estimates not adequately described
- prediction error ignores differing accuracies



## Variance Homogeneity

Alternatives

- put more emphasis where observations are more accurate
- **weighted least squares**
- variance-stabilizing transformations



## Normality

Normality is **NOT VERY IMPORTANT!!!**

- normality is required for **random deviations**, not for response variable
- problematic if distribution of random deviations **very skewed**
- no distributional assumptions for explanatory variables

# Normality

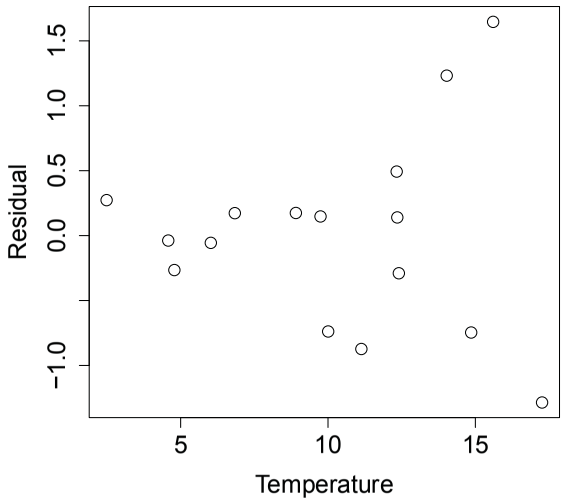
Alternatives

- transformations
- incorporate non-normality into statistical model
- generalized linear models



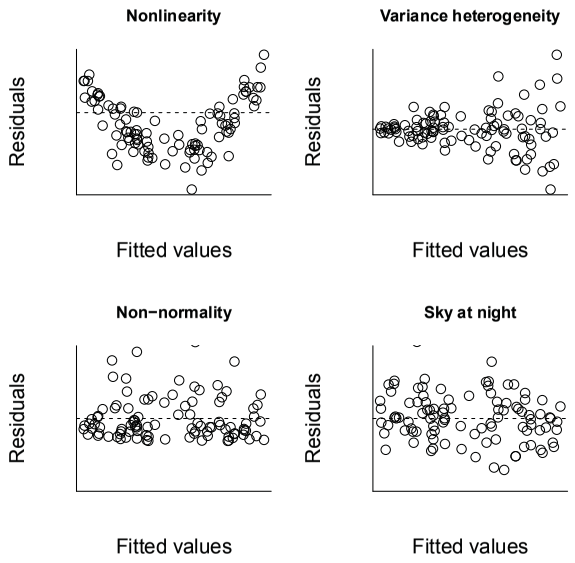
# Checking assumptions

Residual plot



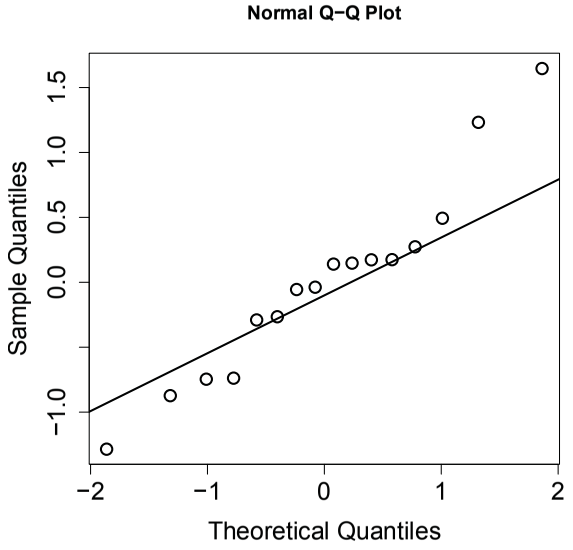



# Checking Assumptions - Violations of assumptions






# Checking assumptions - Normality



# Checking assumptions

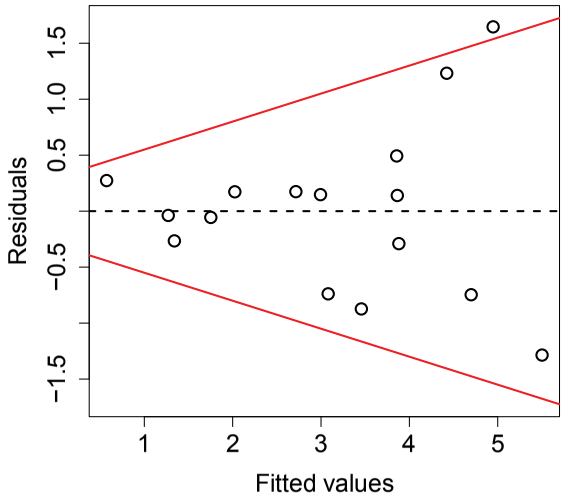
Checking for independence

- temporal dependence: plot residuals in time order
- spatial dependence: plot residuals against spatial coordinates
- clustered sampling: color residuals according to groups



# Checking assumptions

Soil respiration example






# Prediction

- What do we want to predict?
  - the average (expected) value of the response for a given value of the predictor

    $\longrightarrow$ 95%−confidence interval for expected response

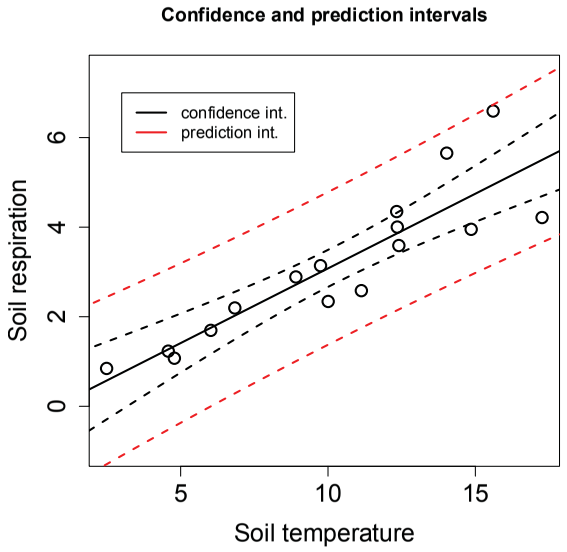
  - a reasonable range where we expect future observations?

    $\longrightarrow$ 95% − prediction interval for future observation

  - distinction especially important if statistical models are incorporated into simulation models



# Prediction



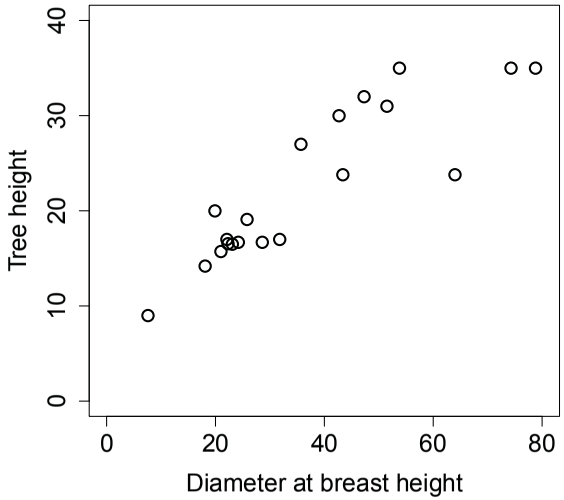
## Prediction

Extrapolation

- confidence bands become wider at both ends of the observed range
- predictions are less accurate
- What happens outside the observed range?
- How do we know that the statistical model is still correct outside the observed range?
- be **extremly cautious** with extrapolation!!!



## Regression through the origin






## Linear Regression - Summary

- **Numerics:** We can always „calculate“ the parameter estimates
- **Statistics:** Assumptions are needed if we want to assess the accuracy of estimates
- Assumptions are ususally **violated**

## Model

Multiple linear regression model

Predictors Response

$$\begin{array}{cc} x_1 & \searrow \\ \vdots & \rightarrow \quad y_i \\ x_k & \nearrow \end{array}$$

- response is a linear function of explanatory variables

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

Applied Statistics
J.Schumacher

## Model

Assumptions

$$\varepsilon_i \sim N(0, \sigma^2) \qquad \text{i.i.d.}$$

- same assumptions as for simple linear regression:
  - Linearity
  - Variance homogeneity
  - Normality
  - Independence
- parameter estimation based on Method of least squares is essentially the same
- but there are some new problems and questions

Applied Statistics
J.Schumacher

## Model

New questions and problems

- Problem: graphical display of relationship
- Question: Which explanatory variables are important? $\longrightarrow$ **Model selection**

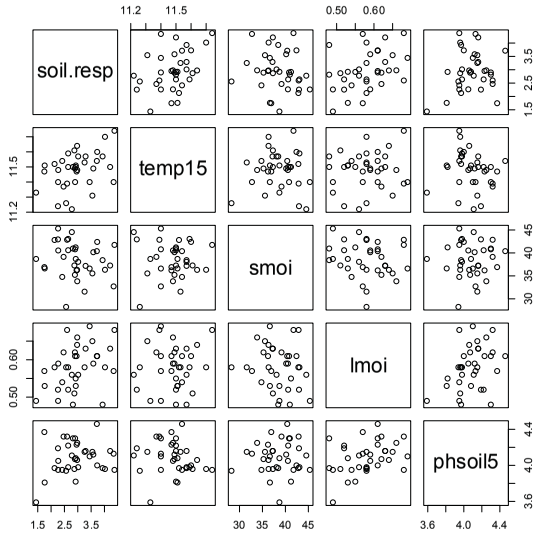
Applied Statistics
J.Schumacher

## Example

Soil respiration Hainich forest

- data from A. Soe
- response variable : soil respiration, measured by manual soil respiration measurement chambers Licor 6400-09
- subset of measured potential explanatory variables:
  - soil temperature, measured at $0cm, 5cm$ and $15cm$ depth (temp0,temp5, temp15)
  - soil moisture (smoi)
  - litter moisture (lmoi)
  - pH value of litter and soil at $5cm$ depth (phlitter, phsoil5)

# Pairwise scatterplots



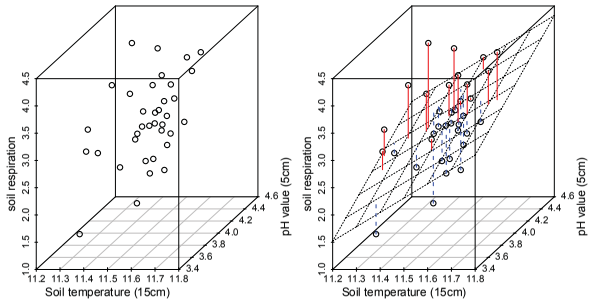

Applied Statistics
J.Schumacher

# Pairwise scatterplots

- provides information about relationships between explanatory variables
- enables idenfication of extreme observations
- only bivariate relationships

Applied Statistics
J.Schumacher

# 3D-Plot

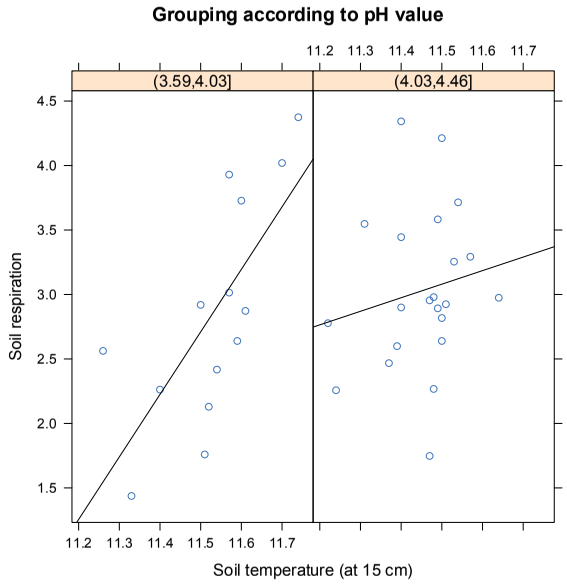



Applied Statistics
J.Schumacher

# 3D-Plot

- small step from 2D to 3D, but what about 4D?
- difficult to identify pattern in 2D projection

# Trellis graphics




Applied Statistics
J.Schumacher

# Trellis graphics

- very flexible
- conditioning also possible for combinations of explanatory variables
- detection of interactions between explanatory variables

Applied Statistics
J.Schumacher

# Calculation

Model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki} + \varepsilon_i$$

Design matrix:

$$\mathbb{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{k1} \\ 1 & x_{12} & \ldots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \ldots & x_{kn} \end{pmatrix}$$

Parameter estimates:

$$\underline{\hat{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\right)^T = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\underline{y}$$

Applied Statistics
J.Schumacher

# Calculation

## Results

```
Call:
lm(formula = soil.resp ~ ., data = hainich.data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0896 -0.4003 -0.1353  0.3715  0.9398

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.88754   11.26280  -3.453 0.001782 **
temp0         0.45785    0.57450   0.797 0.432188
temp5        -4.65635    1.55003  -3.004 0.005561 **
temp15        7.68080    1.89904   4.045 0.000373 ***
smoi         -0.03625    0.02932  -1.236 0.226610
lmoi         -1.71884    2.64404  -0.650 0.520939
phlitter      0.39196    0.34631   1.132 0.267306
phsoil5       0.70166    0.66509   1.055 0.300454
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5782 on 28 degrees of freedom
Multiple R-squared:  0.4899, Adjusted R-squared:  0.3623
F-statistic: 3.841 on 7 and 28 DF,  p-value: 0.004778
```

## Calculation

Interpretation

- (partial) regression coefficients describe the effect of an explanatory variable **if all other explanatory variables are fixed**
- if explanatory variables are highly correlated, isolated change of a single variable may be impossible
- signs of estimated coefficients may be counter-intuitive
- do not overinterpret P values of t-tests, because:
  - Tests are not independent
  - P values may change drastically if variables are removed/added



## Model selection

A word of **WARNING!!!** before

- any statistical procedure can only **support** the model selection process
- do not slavishly trust the result of an automatic statistical model selection
- a model is a model and truth is truth
- there is no true model, we are searching for an adequate model
- accompany the model selection process with your expert knowledge
- or even better: really think about potentially reasonable models before starting any statistical analysis



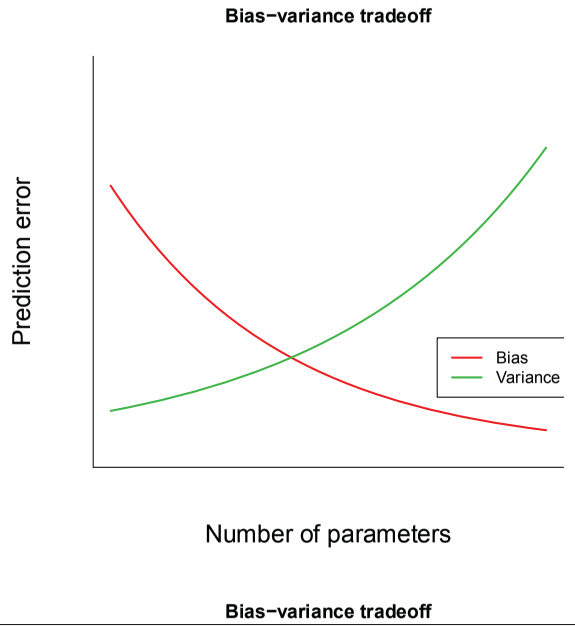
## Model Selection

Preliminaries

- the more predictors the better the model fits the observed data
- the more parameters have to be estimated the larger the estimation error
- bias-variance tradeoff
- adequate model is the best compromise between model fit and model complexity



## Model selection - Bias-variance tradeoff



# Model selection

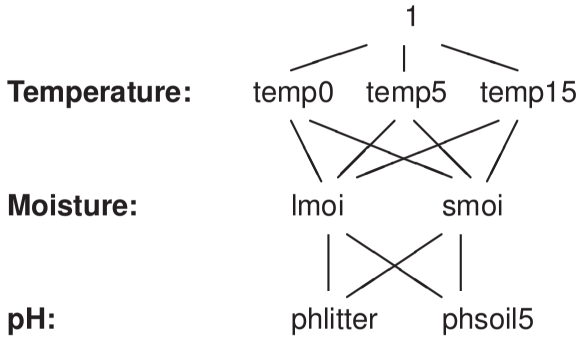
A priori model hierarchy

- soil temperature is certainly important for soil respiration
- question is which depth is best predictor
- moisture may be limiting factor - important in close to extreme situations
- litter moisture or soil moisture?
- not sure whether pH is important - third level in model hierarchy



# Model selection

A priori model hierarchy






# Model selection

Extra sum of squares test - F test

- Follow the model hierarchy from top to bottom
- At each stage perform a statistical test whether inclusion of additional variables **significantly** improves the model
- Appropriate statistical test is **F-test** or **Extra-sum-of-squares test**



# Model selection

Extra sum of squares test - F test

- comparison of two hierarchical models
  - **Model 1:** simpler model with $k$ parameters
  - **Model 2:** larger model with $k + p$ parameters
- **Null hypothesis $H_0$:** simpler model is true
- **Reformulation $H_0$:** all additional coefficients equal 0

$$H_0 : \beta_{k+1} = \ldots = \beta_{k+p} = 0$$

- **Test statistic**: How much better is Model 2?

$$F = \frac{(RSS_1 - RSS_2)/p}{RSS_2/(n-(k+p+1))} = \frac{\text{Model improvement } \textit{per parameter}}{\text{Residual variance of larger model}}$$

# Model selection - Extra sum of squares test

**Level 1:**

```
Single term additions

Model:
soil.resp ~ 1
       Df Sum of Sq    RSS     AIC F value  Pr(>F)
<none>              18.351 -22.258
temp0   1   0.10969 18.241 -20.474  0.2044 0.65403
temp5   1   0.01257 18.338 -20.283  0.0233 0.87955
temp15  1   2.86182 15.489 -26.361  6.2819 0.01714 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

**Level 2:**

```
Single term additions

Model:
soil.resp ~ temp15
       Df Sum of Sq    RSS     AIC F value  Pr(>F)
<none>              15.489 -26.361
smoi    1   0.48137 15.008 -25.498  1.0585 0.31105
lmoi    1   2.34451 13.145 -30.270  5.8859 0.02089 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Applied Statistics
J.Schumacher

# Model selection - Extra sum of squares test

**Level 3:**

```
Single term additions

Model:
soil.resp ~ temp15 + lmoi
         Df Sum of Sq    RSS     AIC F value Pr(>F)
<none>                13.145 -30.270
phlitter  1  0.002259 13.142 -28.276  0.0055 0.9413
phsoil5   1  0.246659 12.898 -28.952  0.6120 0.4398
```

Applied Statistics
J.Schumacher

# Model selection - Hierarchical models

Final model:

```
Call:
lm(formula = soil.resp ~ temp15 + lmoi, data = hainich.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.13926 -0.44079 -0.03058  0.37938  1.60982

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.9374    10.2998  -2.421   0.0211 *
temp15        2.2012     0.9009   2.444   0.0201 *
lmoi          4.5200     1.8631   2.426   0.0209 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 0.6311 on 33 degrees of freedom
Multiple R-squared:  0.2837, Adjusted R-squared:  0.2403
F-statistic: 6.535 on 2 and 33 DF,  p-value: 0.004064
```

Applied Statistics
J.Schumacher

# Model selection - Hierachical models

- F test only valid for hierachical models
- problem of multiple testing
- with arbitrarily chosen significance level $\alpha = 0.05$
- effect of temperature predictors may be evaluated differently after moisture is included in the model

# Model selection

A priori selected candidate models

$soil.resp \sim temp0$
$soil.resp \sim temp5$
$soil.resp \sim temp15$

$soil.resp \sim temp0 + smoi$
$soil.resp \sim temp0 + lmoi$

$soil.resp \sim temp5 + smoi$
$soil.resp \sim temp5 + lmoi$

$soil.resp \sim temp15 + smoi$
$soil.resp \sim temp15 + lmoi$

$soil.resp \sim temp0 + smoi + phlitter$
$soil.resp \sim temp0 + smoi + phsoil5$
$soil.resp \sim temp0 + lmoi + phlitter$
$soil.resp \sim temp0 + lmoi + phsoil5$

$soil.resp \sim temp5 + smoi + phlitter$
$soil.resp \sim temp5 + smoi + phsoil5$
$soil.resp \sim temp5 + lmoi + phlitter$
$soil.resp \sim temp5 + lmoi + phsoil5$

$soil.resp \sim temp15 + smoi + phlitter$
$soil.resp \sim temp15 + smoi + phsoil5$
$soil.resp \sim temp15 + lmoi + phlitter$
$soil.resp \sim temp15 + lmoi + phsoil5$



# Model selection - a priori candidate models

- 21 candidate models
- $R^2$ not suitable to decide between models
- **Akaike's Information Criterion (AIC)** selects a good compromise between model fit to data and model complexity
-

$$AIC = \underbrace{-2 \quad \text{Loglikelihood}}_{\text{Model fit}} + \underbrace{2 \quad \text{Number of parameters}}_{\text{Penalty for model complexity}}$$

- calculate AIC for every candidate model
- best adequate model is the one with the **smallest value of AIC**



# Model selection - AIC

| **Model** | **AIC** | **$\Delta$ AIC** |
|---|---|---|
| $soil.resp \sim temp15 + lmoi$ | 73.89 | 0 |
| $soil.resp \sim temp15 + lmoi + phsoil5$ | 75.21 | 1.32 |
| $soil.resp \sim temp15 + lmoi + phlitter$ | 75.88 | 1.99 |
| $soil.resp \sim temp15$ | 77.80 | 3.91 |
| ... | | |

- absolut values of AIC are (more or less) meaningless
- rule of thumb:
  - models with $\Delta AIC \leq 2$ are reasonably supported by data
  - models with $\Delta AIC \geq 10$ have essentially no support
- identifies the best among candidate models
- model validation based on residuals necessary to show that is is a good model
- it is still a statistical procedure



# Model selection - Best subset selection

- 21 candidate models comprise not all possible models
- for $k$ explanatory variables there are $2^k$ possible models
- evaluating all models may be demanding

# Model selection - Best subset selection

Result

| Model | AIC |
|---|---|
| soil.resp~temp5+temp15+phsoil5 | 66.28 |
| soil.resp~temp5+temp15 | 67.07 |
| soil.resp~temp5+temp15+phlitter | 67.62 |
| soil.resp~temp5+temp15+smoi+phsoil5 | 67.53 |
| soil.resp~temp5+temp15+phlitter+phsoil5 | 67.82 |



# Model selection - best subset selection

- danger of overfitting
- high chance that *surprising* models, not supported by expert knowledge are selected due to peculiarities in the data set
- not robust, new data may select substantially different model

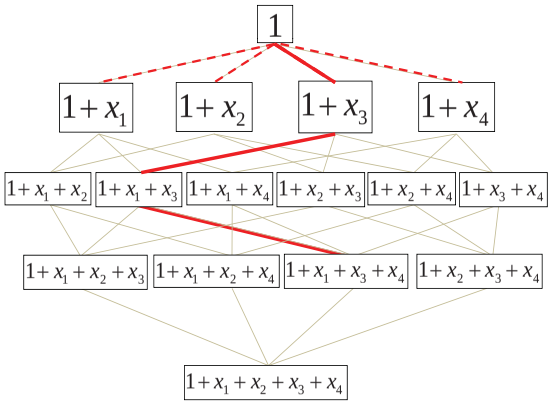


# Model selection - stepwise procedures

- guided tour through world of models
- visit only the most promising models
- three different approaches: forward, backward, stepwise
- individual steps sometimes based on Extra sum of squares test, but this is **problematic**
- better base individual steps on improvement in AIC criterion



# Model selection - stepwise procedures



## Model selection - stepwise procedures

Forward:
- Start with smallest model
- add „best“ predictor variable as long as AIC decreases

Backward:
- Start with largest model containing all predictors
- remove predictor variables as long as AIC decreases

Stepwise:
- after each forward step check whether one of the earlier selected predictors can now be removed

Applied Statistics
J.Schumacher

## Model selection - stepwise procedures

Results:
- all three approaches lead to the same model
- this is an exception, not the rule
- selected best model is the same as for best subset selection
- this is an exception, not the rule

Applied Statistics
J.Schumacher

## Model selection - we are not yet done

- all approaches so far considered only a very limited set of models
- only linear additive models
- nonlinearities?
- transformations?
- and most important: **interactions!**

Applied Statistics
J.Schumacher

## Model selection - Interactions

- **interaction:** quantitative effect of one predictor variable on response variable depends on value of another predictor variable
- **interaction $\neq$ correlation:** clearly distinguish (when thinking AND speaking!) between correlation and interaction
- **marginality principle:** if interaction is included in a model, so are all corresponding main effects
- causes some technical difficulties with the automatic procedures above

## Model selection - Interactions

Forward selection:

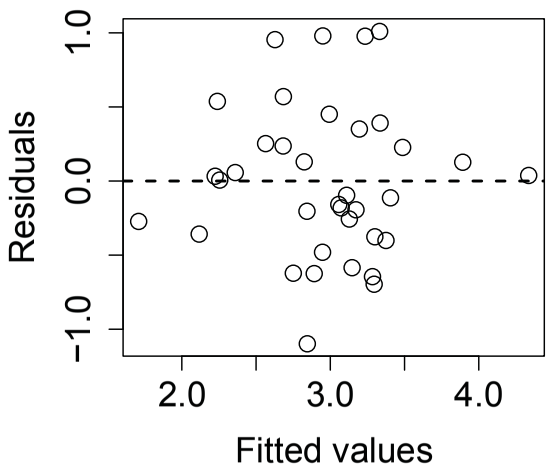
```
soil.resp ~ temp15 + temp5 + phsoil5 + temp15:temp5

Coefficients:
 (Intercept)         temp15          temp5        phsoil5
    1024.105        -86.056        -93.571          1.086

AIC = 65.57
```

Backward selection:

```
selected model contains 10 pairwise interactions

AIC = 60.33
```



## Model selection - Model validation




$R^2 = 0.49$



## Model selection - Summary

- different model selection strategies find different best models
- only slight differences in predictions
- recommendations:
  - think carefully about reasonable models a priori, especially interactions
  - option I: define modeling strategy/hierarchy
  - option II: define set of candidate models
  - do not overinterpret results of automatic selection procedures



## Transformations

- different goals:
  - achieve variance homogeneity
  - reduce influence of few extreme observations
  - linearize relationship

## Transformations

Variance-stabilizing transformations

- logarithmic transformation (very skewed distributions)

$$y_{new} = \log(y_{old} + c)$$

- square-root transformation (count data)

$$y_{new} = \sqrt{y_{old} + c}$$

- power transformation (more flexible)

$$y_{new} = y_{old}^{p}$$

- arcsine transformation (proportions)

$$y_{new} = \arcsin\left(\sqrt{y_{old}}\right)$$



## Linear regression models

Summary

- Data exploration
- Parameter estimation
- Model selection
- Model validation





## Linear Regression Model

Problem

$$\text{independent variable} \Longrightarrow \text{dependent variable}$$
$$\text{grouping variable} \Longrightarrow y$$
$$\text{nominal} \qquad \text{metric}$$

### Model

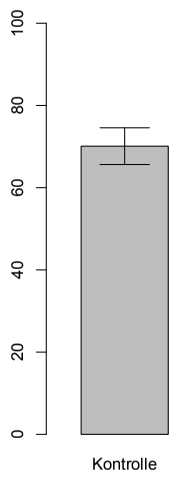
$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \begin{array}{l} i = 1, 2, \ldots, g \\ j = 1, \ldots, n_i \end{array}$$

$\mu$ – general mean
$\alpha_i$ – effect of level $i$
$\varepsilon_{ij}$ – random error term

Applied Statistics

J.Schumacher

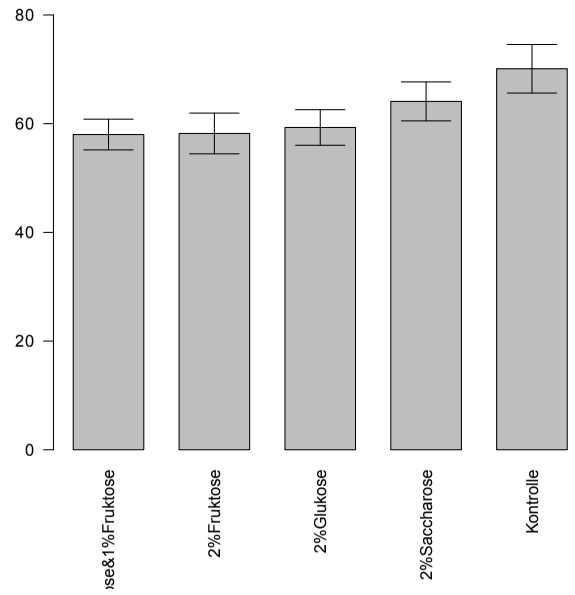
# Bar Chart




Applied Statistics

J.Schumacher

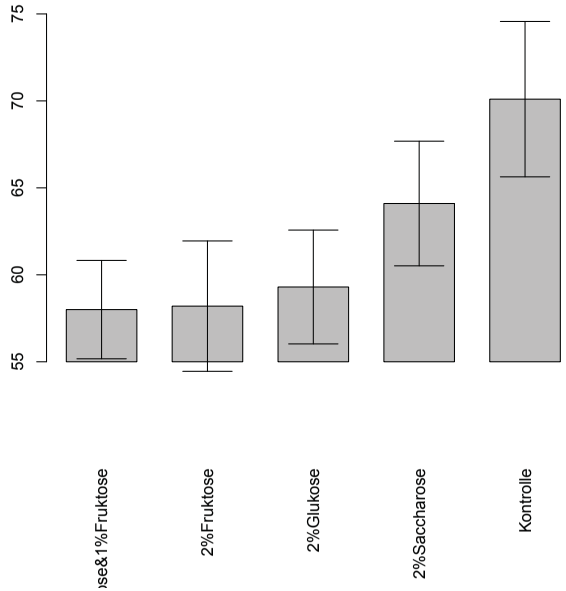
# Bar Chart




Applied Statistics

J.Schumacher

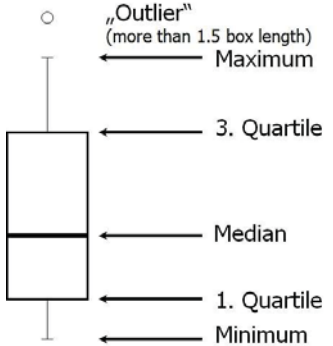
# Bar Chart



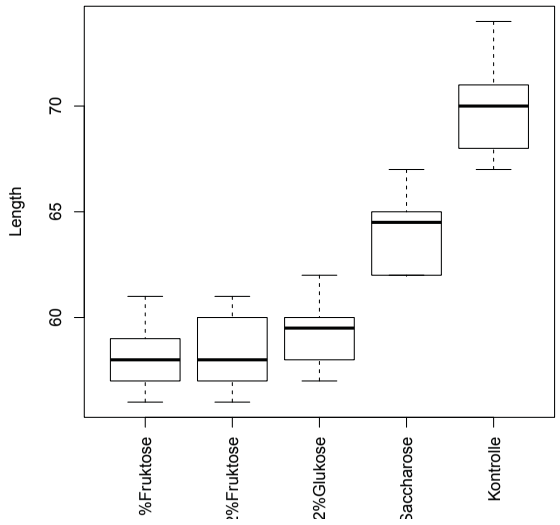

Applied Statistics

J.Schumacher

# Box Plot

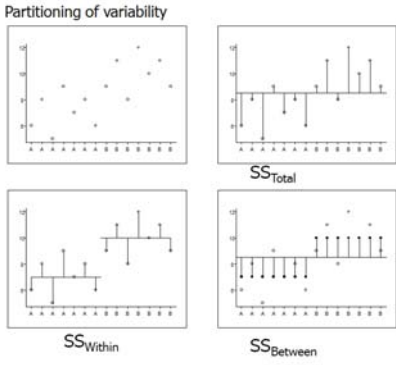


# Bar Chart






# Partitioning of variability






# Partitioning of variability

| Source | SS | DF | MS | Exp. value |
|---|---|---|---|---|
| Between | $\sum_i n_i(\bar{y}_i - \bar{y})^2$ | $g-1$ | $\frac{\sum_i(\bar{y}_i-\bar{y})^2}{g-1}$ | $\sigma_\varepsilon^2 + \sum_i n_i \frac{\alpha_i^2}{g-1}$ |
| Within | $\sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ | $n-g$ | $\frac{\sum_i \sum_j (y_{ij}-\bar{y}_i)^2}{n-g}$ | $\sigma_\varepsilon^2$ |



# Calculations

- Null hypothesis:

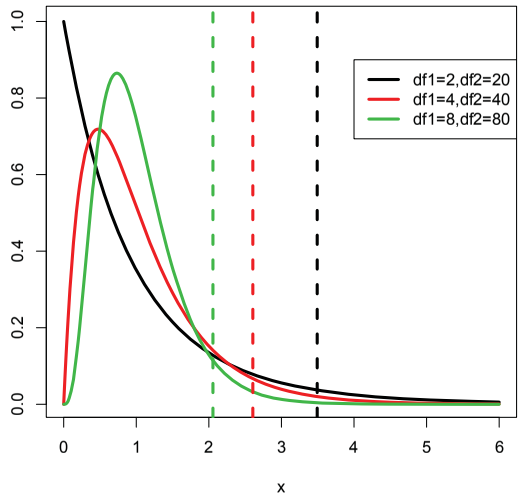
$$H_0 : \alpha_1 = \ldots = \alpha_g = 0$$

- Test Statistic:

$$F = \frac{MS_{between}}{MS_{within}}$$

- Sampling distribution under $H_0$:
  Fisher distribution (after R.A. Fisher)
  Parameter: nominator and denominator d.f.

## Fisher distribution






## Result

```
> summary(peas.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
group        4 1077.3  269.33   82.17 <2e-16
Residuals   45  147.5    3.28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0
```



## Assumptions

**Independence**
observations are statistically independent

**Variance homogeneity**
comparable variability of observations within each group

**Normality** random deviations follow a normal distribution

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \qquad \text{unabhängig}$$



## Regression type output

```
> summary.lm(peas.aov)
Call:
aov(formula = length ~ group, data = peas.dat

Residuals:
   Min     1Q Median     3Q    Max
  -3.1   -1.2   -0.1    0.9    3.9

Coefficients:
                  Estimate Std. Error t value
(Intercept)        58.0000     0.5725 101.307
group2%Fruktose     0.2000     0.8097   0.247
group2%Glukose      1.3000     0.8097   1.606
group2%Saccharose   6.1000     0.8097   7.534
groupKontrolle     12.1000     0.8097  14.944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0
```

## Regression or ANOVA

- treatment (independent) variable is metric
- e.g. different levels of concentration, temperatures
- ANOVA: treat each level separately
- Regression: functional form of relationship
- hierarchical models to test linearity of relationship



## Multiple Comparisons

- multiple pairwise comparisons
- Bonferroni type corrections for multiple testing
- special procedures for ANOVA, e.g. TukeyHSD, Student-Newman-Keuls, Dunnett



## Nonparametric ANOVA

- Nonparametric ANOVA: KRUSKAL-WALLIS Test
- ANOVA based on ranked data
- Assumption: Equal variances!!!!!!
- Wrong Argument:
  Nonnormal data -> Nonparametric Test



## Comparing two groups

- equivalent to two sample t-test
- same P value
- relationship between test statistics:

$$F = t^2$$

## Summary One-way ANOVA

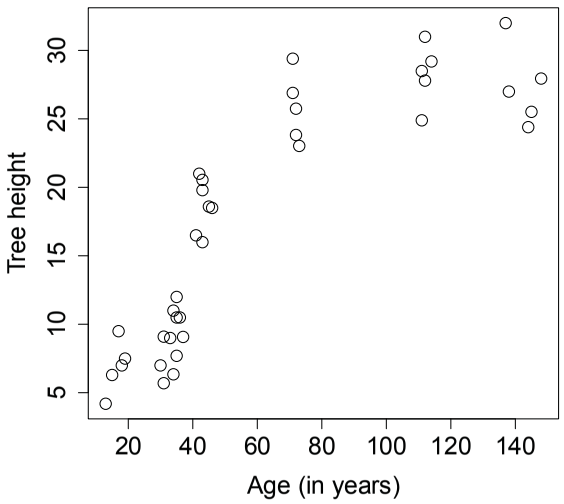
- Assure suitable randomization and replication
- Variance homogeneity more important than normality
- Transformations to achieve homogeneity better than uncritical use of nonparametric ANOVA NA

Applied Statistics
J.Schumacher

Applied Statistics
J.Schumacher

## Example

Height-age relationship for spruce




Applied Statistics
J.Schumacher

## Nonlinear Regression

Polynomial regression

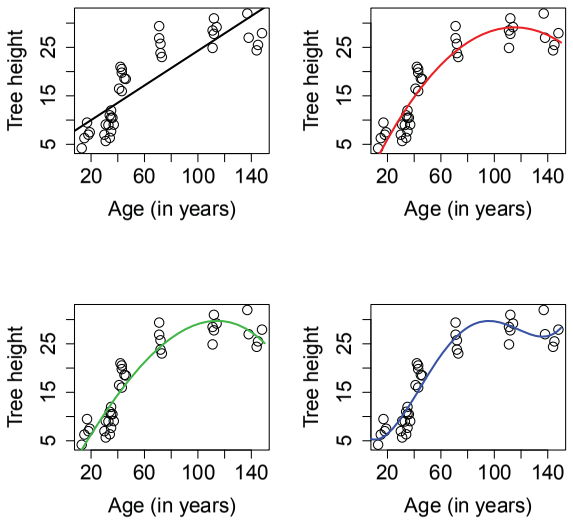
- nonlinear relationship but functional form unknown
- polynomial function as simplest nonlinear function
- any nonlinear function can be approximated by polynomial functions

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$
$$\vdots$$
$$y_i = \beta_0 + \beta_1 x_i + \ldots + \beta_k x_i^k + \varepsilon_i$$

- hierarchy of polynomials of increasing order
- **quasilinear** - linear in parameters $\beta_j$

# Nonlinear regression

## Polynomial regression




Applied Statistics
J.Schumacher

# Nonlinear regression

## Polynomial regression

- Hierarchy of models - Extra sum of squares test

```
Analysis of Variance Table

Model 1: h ~ age
Model 2: h ~ age + I(age^2)
Model 3: h ~ age + I(age^2) + I(age^3)
Model 4: h ~ age + I(age^2) + I(age^3) + I(age^4)
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1     36 806.06
2     35 360.28  1    445.78 48.8897 5.377e-08 ***
3     34 354.86  1      5.42  0.5944   0.44621
4     33 300.89  1     53.97  5.9191   0.02056 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Applied Statistics
J.Schumacher

# Nonlinear regression

## Polynomial regression

- Direct evaluation based on AIC

```
           df      AIC
fichte.lm1  3 229.9131
fichte.lm2  4 201.3130
fichte.lm3  5 202.7370
fichte.lm4  6 198.4679
```

Applied Statistics
J.Schumacher

# Nonlinear regression

## Polynomial regression

- polynomials may be reasonable for limited range
- polynomials show undesired fluctuating behaviour
- extrapolation is extremely dangerous

# Nonlinear regression

Linearization

- nonlinear functional form of relationship is "known"
- Example: allometric relationships

$$y_i = \beta_1 x_i^{\beta_2}$$

- Linearization:

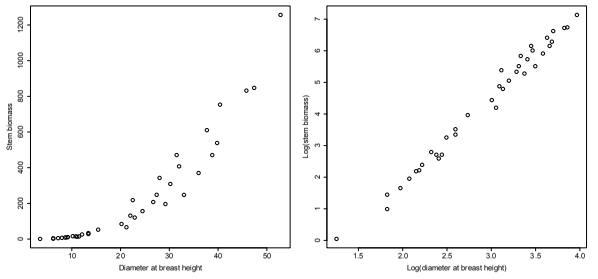
$$\log y_i = \log \beta_1 + \beta_2 \log x_i$$
$$y_i^* = \beta_1^* + \beta_2 x_i^* + \varepsilon_i$$



# Nonlinear regression

Linearization

- Biomass - diameter at breast height for spruce trees






# Nonlinear regression

Linearization

- Advantages:
  - simple, well understood linear models
  - transformation often also homogenizes variances
- Disadvantages:
  - predictions are sought for at original scale
  - nonlinear backtransformation causes bias in predictions



# Nonlinear regression

Truly nonlinear models

- nonlinear relationship which can/shall not be linearized
- functional form of relationship follows from theoretical considerations
- use parametrized class of nonlinear functions which are able to describe main qualitative characteristics of relationship

# Nonlinear regression

Example: growth functions

- logistic growth as a solution of a theoretical (deterministic) growth model

$$y = \frac{\beta_1}{\left(1 + \exp\left(-\frac{x-\beta_2}{\beta_3}\right)\right)}$$

- Michaelis-Menten model of enzyme kinetics

$$y = \frac{\beta_1 x}{\beta_2 + x}$$

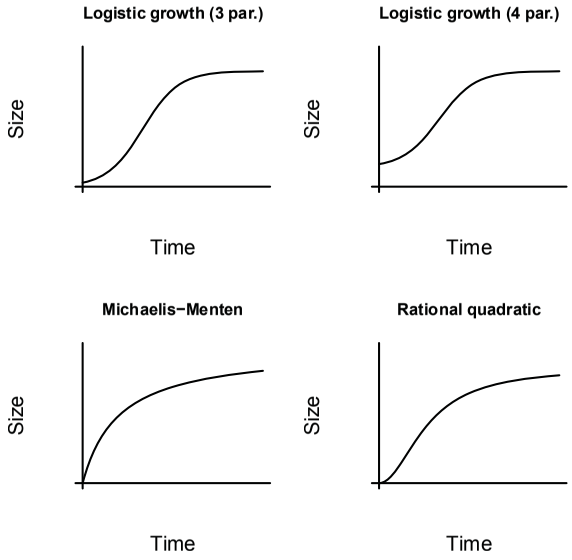
- rational quadratic function

$$y = \frac{\beta_1 * \left(\frac{x}{\beta_2}\right)^2}{\left(1 + \left(\frac{x}{\beta_2}\right)^2\right)}$$



# Nonlinear regression

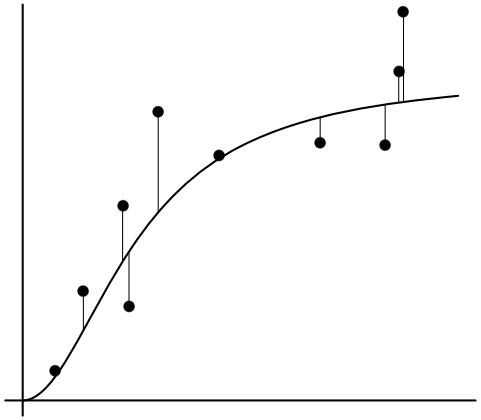
Growth functions






# Nonlinear regression

Nonlinear method of least squares





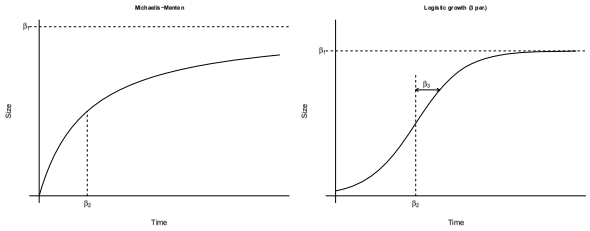
# Nonlinear regression

Starting values

- parameter estimates can not be calculated analytically
- estimates are result of iterative numerical search
- starting values for search must be provided
- possibility of several local minima
- **good** starting values must be provided

# Nonlinear regression

## Starting values



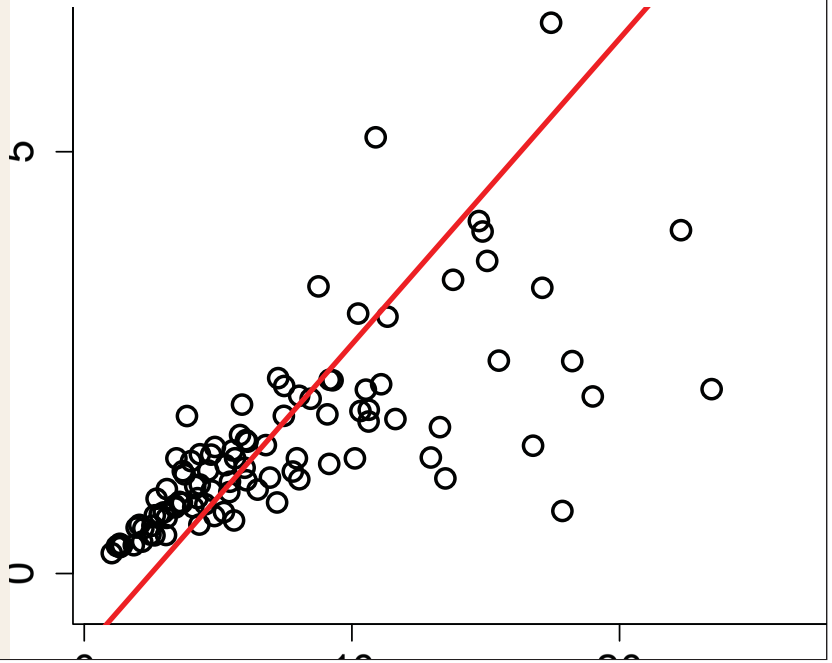

- $\beta_1 =$ asymptotic value for *time* $\to \infty$
- $\beta_2 =$ half saturation

- $\beta_1 =$ asymptotic value for *time* $\to \infty$
- $\beta_2 =$ inflection point
- $\beta_3 =$ distance between $\beta_2$ and point where *response* $= \frac{\beta_1}{1+e^{-1}}$

Applied Statistics
J.Schumacher

Applied Statistics
J.Schumacher

# Variance Homogeneity



Applied Statistics
J.Schumacher

# Variance heterogeneity - Weighted least squares

- method of least squares puts most emphasis on regions, where variability is largest
- **weighted least squares:**

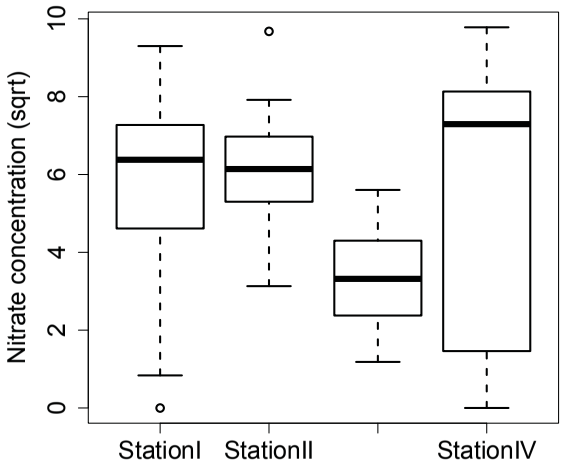
$$\text{minimize} \sum_{i=1}^{n} \mathbf{w_i} \, (y_i - \hat{y}_i)^2$$

- best weights are inversely proportional to variance:

$$\mathbf{w_i} = \frac{1}{\sigma_i^2}$$

- variance is unknown and must be estimated

# Variance heterogeneity - different variances in groups




Applied Statistics
J.Schumacher

# Variance heterogeneity - different variances in groups

- obviously different variability
- transformations not helpful because no tendency
- traditional ANOVA analysis gives:

```
            Df Sum Sq Mean Sq F value Pr(>F)
station      3   66.6  22.189   3.555 0.0199 *
Residuals   56  349.5   6.241
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Applied Statistics
J.Schumacher

# Variance heterogeneity - different variances in groups

- estimate variances within groups and use these to calculate weights
- compare equal variances model with different variances model:
- 

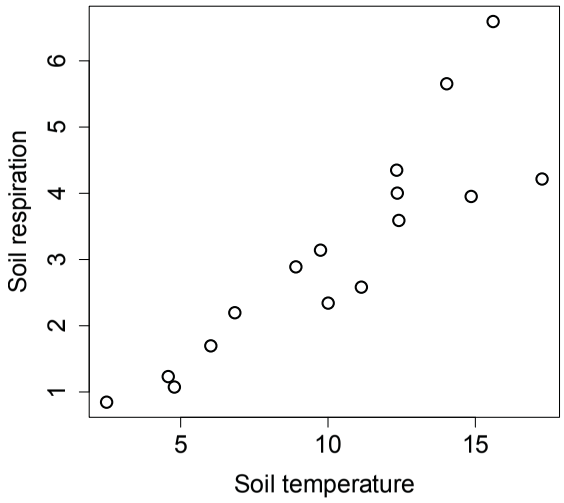
```
         Model df    AIC   logLik L.Ratio p-value
equal variances  5 282.29  -136.14
diff. variances  8 270.58  -127.29  17.70   5e-04
```

- modified model gives:

```
            numDF  F-value p-value
station         3  10.2366  <.0001
```

Applied Statistics
J.Schumacher

# Scatterplot



# Variance heterogeneity - Variance models

- no groups - no possibility to estimate variance within groups
- create groups by binning and estimate variance within groups
- groups to small: small number of observations, unreliable estimates
- groups to large: systematic change of response within group inflates variance estimate



# Variance heterogeneity - Variance models

- variance changes continuously
- often related to explanatory variable
- or related to values of the response
- model variance as a function of explanatory variable

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_i^2)$$
$$\sigma_i^2 = \sigma^2 x_i^{2\delta}$$

- simultaneously estimate $\beta_0, \beta_1$ and $\delta$



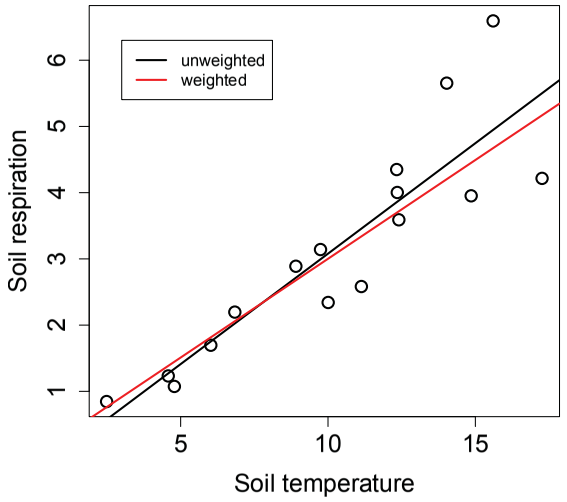
# Variance heterogeneity

```
 Parameter estimates:
   power
1.276296

Coefficients:
                 Value  Std.Error   t-value p-value
(Intercept) 0.01881966 0.13720137  0.137168  0.8929
temp        0.29827654 0.02581033 11.556477  0.0000
```



# Variance heterogeneity

Comparison of results



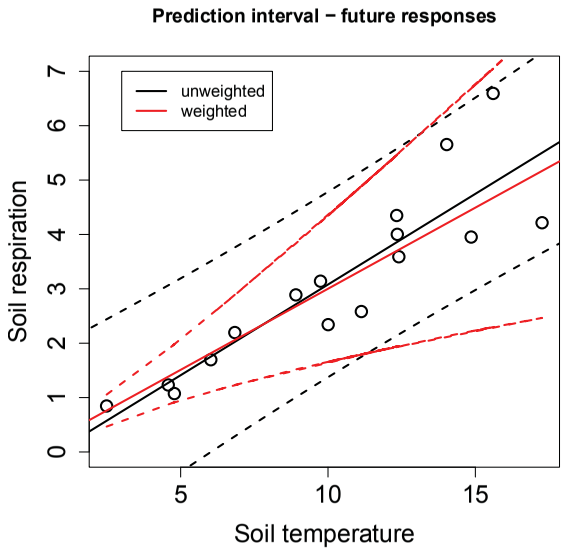
# Variance heterogeneity - Comparison of results

## Accuracy of estimates

```
Unweighted least squares:

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.25643    0.50740  -0.505    0.621
temp         0.33355    0.04597   7.256 4.19e-06 ***

Weighted least squares

                Value  Std.Error   t-value p-value
(Intercept) 0.01881966 0.13720137  0.137168  0.8929
temp        0.29827654 0.02581033 11.556477  0.0000
```

Applied Statistics
J.Schumacher

# Variance heterogeneity - Comparison of results



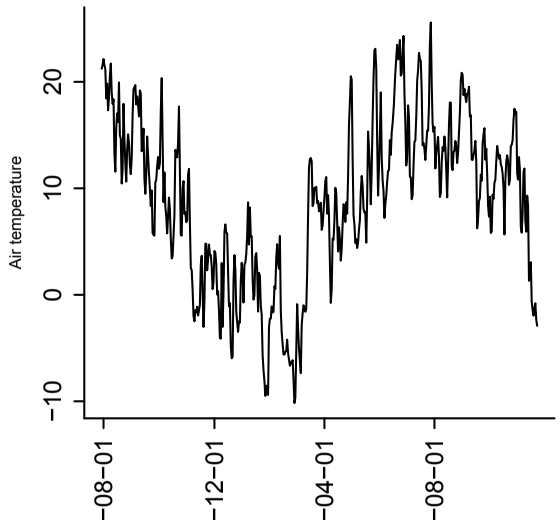

Applied Statistics
J.Schumacher

Applied Statistics
J.Schumacher

# Example streamflow

- streamflow is measured at a spring in Conventwald in 15 min resolution
- time span is August 2004 to November 2005
- additionally meterological parameters are measured
- we use daily data for our analysis

# Example streamflow



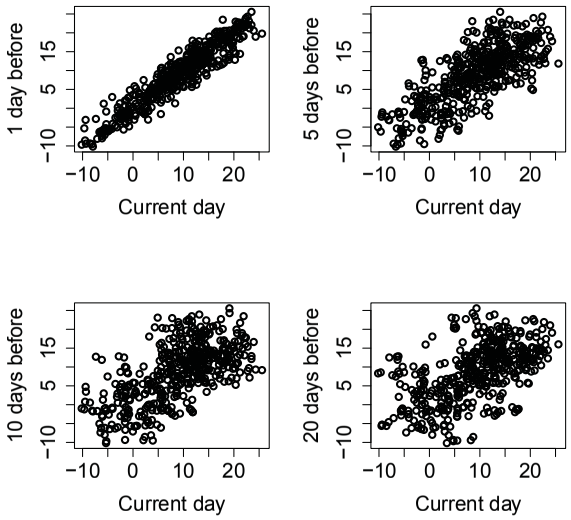

Applied Statistics
J.Schumacher

# Example streamflow

- measurements close in time tend to be similar
- tomorrow we will have the same weather as today is a forecast which is hard to beat
- measurements are stochastically dependent
- we try to quantify this dependence to incorporate it into statistical models

Applied Statistics
J.Schumacher

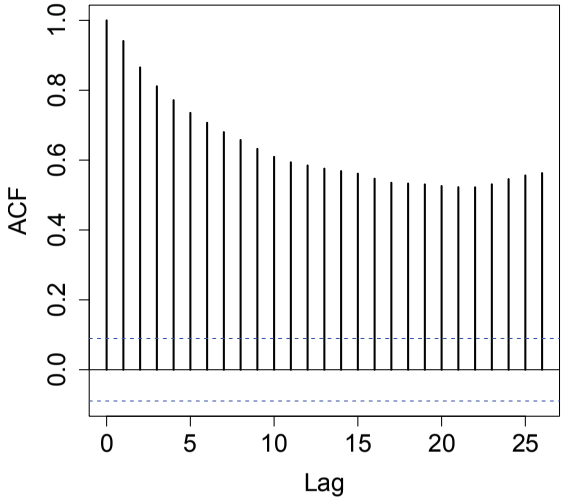
# Correlation for different time lags




Applied Statistics
J.Schumacher

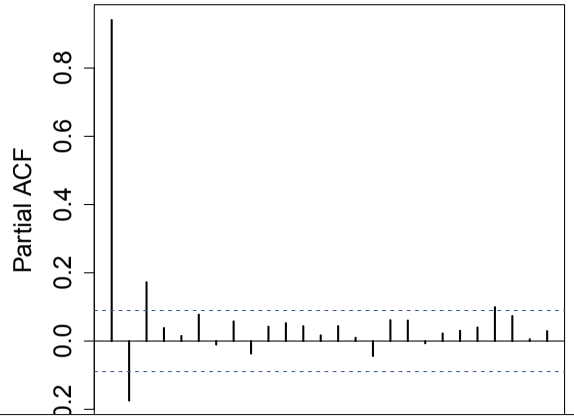
# Autocorrelation function

- correlation coefficient as a function of time lag



## Partial autocorrelation function

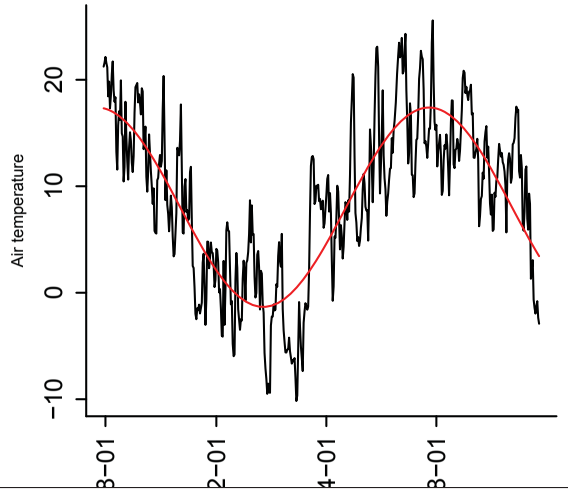
- day 1 is correlated with day 2 which is correlated with day 3 ...
- remove the indirect effect from the ACF -> partial ACF






## Deseasoning

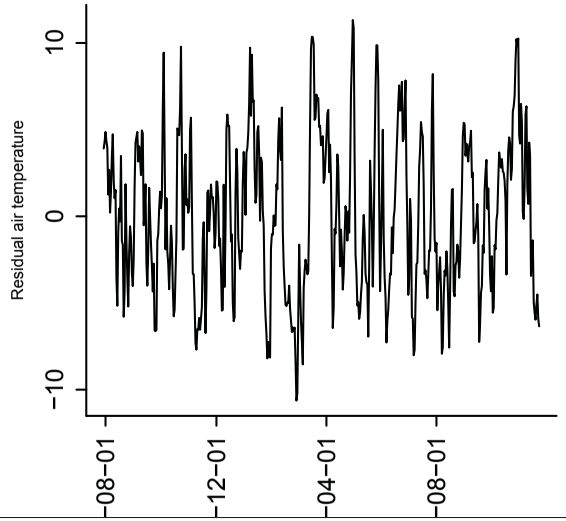
- clear seasonal pattern - not all fluctuations are random!





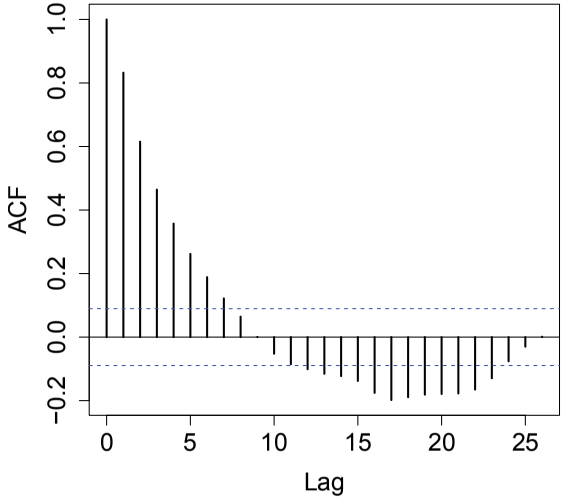

## Deseasoning

- analyse deviations from seasonal trend

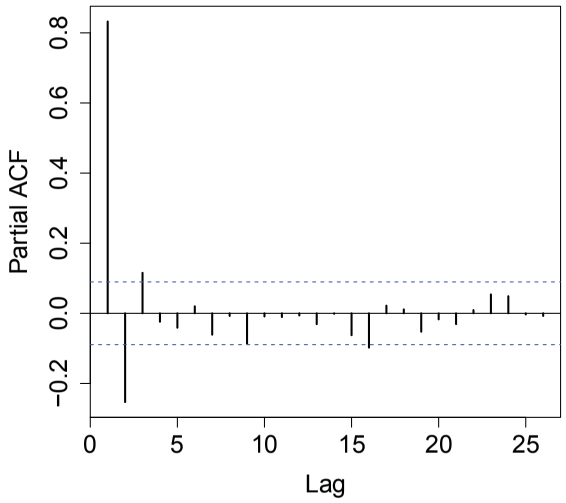





## Deseasoning - Autocorrelation function



## Deseasoning - Partial autocorrelation function






## Time series models

$$y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + \varepsilon_t$$

- errors $\varepsilon_t$ are correlated
- degree of correlation diminishes with time difference

$$cov(\varepsilon_{t_1}, \varepsilon_{t_2}) = h(|t_1 - t_2|)$$



## Autoregressive models

- **AR(1)**: autoregressive process of order 1

$$\varepsilon_t = \underbrace{\varphi}_{\text{autoregression coefficient}} \varepsilon_{t-1} + \underbrace{\eta_t}_{\text{innovation}}$$

- current deviation is a linear function of previous deviation
- plus an **independent** innovation
- $|\varphi|$ must be $< 1$, otherwise deviations explode
- generalization **AR(p)**: autoregressive process of order p

$$\varepsilon_t = \varphi_1 \varepsilon_{t-1} + \ldots + \varphi_p \varepsilon_{t-p} + \eta_t$$



## Autoregressive process - Correlations

$$\begin{aligned} cor(\varepsilon_t, \varepsilon_{t-1}) &= cor(\varphi\varepsilon_{t-1} + \eta_t, \varepsilon_{t-1}) \\ &= \varphi cor(\varepsilon_{t-1}, \varepsilon_{t-1}) \\ &= \varphi \end{aligned}$$

$$\begin{aligned} cor(\varepsilon_t, \varepsilon_{t-2}) &= cor(\varphi\varepsilon_{t-1} + \eta_t, \varepsilon_{t-2}) \\ &= cor(\varphi(\varphi\varepsilon_{t-2} + \eta_{t-1}) + \eta_t, \varepsilon_{t-2}) \\ &= \varphi^2 cor(\varepsilon_{t-2}, \varepsilon_{t-2}) \\ &= \varphi^2 \\ &\vdots \end{aligned}$$

$$cor(\quad) = \varphi^d$$

## Autoregressive process - Correlations

Correlation matrix:

$$cor(\varepsilon_1, \ldots, \varepsilon_n) = \begin{pmatrix} 1 & \varphi & \varphi^2 & \ldots & \varphi^{n-1} \\ \varphi & 1 & \varphi & & \varphi^{n-2} \\ \varphi^2 & \varphi & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \varphi \\ \varphi^{n-1} & \ldots & & \varphi & 1 \end{pmatrix}$$

- structure allows calculation of correlation also for non-equidistant time series



## Moving average models

- **MA(1)**: moving average process of order 1

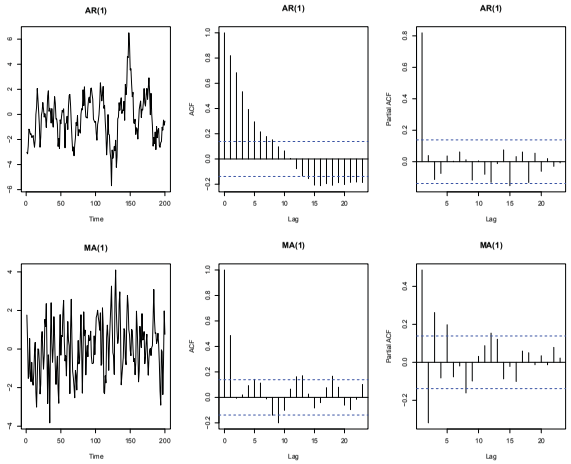
$$\varepsilon_t = \underbrace{\theta}_{\text{moving average coefficient}} \eta_{t-1} + \eta_t$$

- current deviation is a linear function of previous **innovation**
- generalization **MA(p)**: moving average process of order p

$$\varepsilon_t = \theta_1 \eta_{t-1} + \ldots + \theta_p \eta_{t-p} + \eta_t$$



## Time series models






## Time series models

- combination of AR() and MA() models possible $\longrightarrow$ **ARMA(p,q) models**
- model selection based on AIC
- AR models seem to be more often applicable
- for shorter time series AR(1) usually sufficient description of dependence

## Example streamflow - Model fitting

| Model | df | AIC |
|---|---|---|
| temp.day.ar1 | 3 | 2222.299 |
| temp.day.ar2 | 4 | 2192.112 |
| temp.day.ar3 | 5 | 2187.246 |
| temp.day.ma1 | 3 | 2413.953 |
| temp.day.ma2 | 4 | 2264.153 |
| temp.day.ma3 | 5 | 2231.573 |



## Example streamflow - Model fitting

Iterative parameter estimation

- fit seasonal trend parameters using method of least squares
- fit a time series model to the residuals
- improve the estimates for the seasonal trend by incorporating dependence structure in least squares estimation (generalized least squares)
- fit a time series model to the new residuals
- ...
- this is usually done simultaneously by software programms



## Example streamflow - Model fitting

Result:

```
 Parameter estimate(s):
      Phi1       Phi2       Phi3
 1.0834644 -0.3811414  0.1241373

Coefficients:
                Value Std.Error  t-value p-value
(Intercept)  8.013442 0.6157113 13.01493       0
Amplitude    9.499088 0.8963411 10.59763       0

AIC  2186.845
```

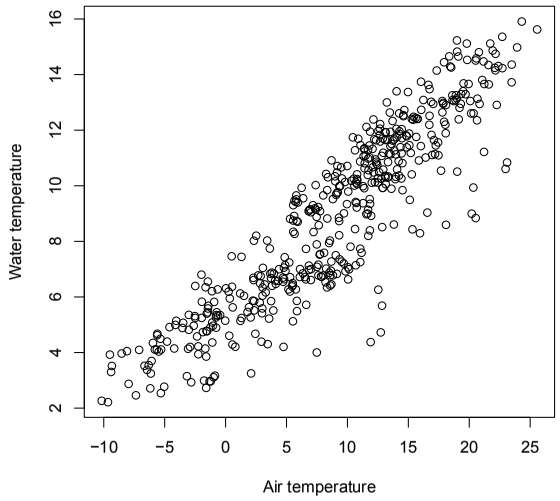


## Regression in Time

- we want to quantify the relationship between air temperatur and water temperature in the spring
- temporal autocorrelation has to be taken into consideration
- influences accuracy of estimates
- possibly influences model selection
- regression model with temporal dependence

$$y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + \varepsilon_t$$
$$cov(\varepsilon_{t_1}, \varepsilon_{t_2}) = h(|t_1 - t_2|)$$

# Regression in Time




Applied Statistics
J.Schumacher

# Regression in Time

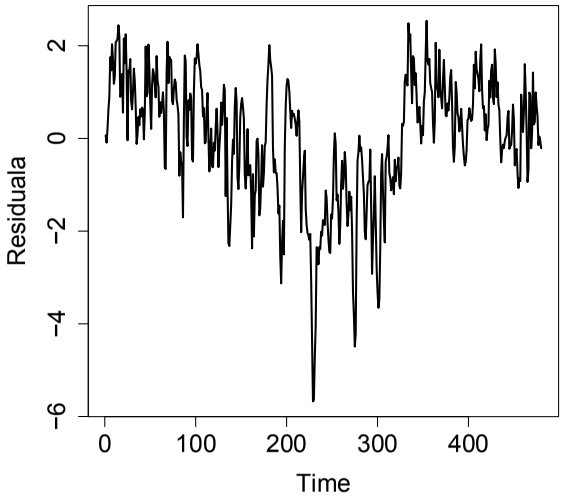
## Naive analysis

```
Call:
lm(formula = convent.daily.ts[, "temp.water"] ~ convent
    "temp.air"])

Residuals:
    Min      1Q  Median      3Q     Max
-5.6748 -0.7571  0.2190  0.9905  2.5385

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.525171   0.095005   58.16   <2e-16 ***
temp.air    0.380257   0.008069   47.13   <2e-16 ***
---

Residual standard error: 1.366 on 478 degrees of freed
Multiple R-squared: 0.8229,     Adjusted R-squared: 0.
F-statistic:  2221 on 1 and 478 DF,  p-value: < 2.2e-1
```

Applied Statistics
J.Schumacher

# Regression in Time

## Naive analysis - Residuals




Applied Statistics
J.Schumacher

# Regression in Time

## Incorporating temporal dependence as an AR(2)-process

```
 Parameter estimate(s):
     Phi1       Phi2
 1.2073034 -0.2166271

Coefficients:
              Value Std.Error   t-value p-value
(Intercept) 7.606263 1.4197970  5.357289       0
temp.air    0.144851 0.0061376 23.600432       0

Standardized residuals:
        Min          Q1         Med          Q3
-1.77703564 -0.70730464  0.08623659  0.68027050  1.748

Residual standard error: 2.787154
Degrees of freedom: 480 total; 478 residual
```

# Regression in Time

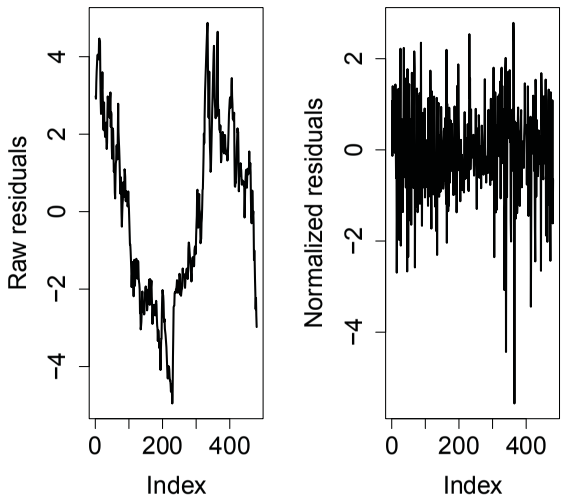
Model comparison

| | df | AIC |
|---|---|---|
| independent | 3 | 1677.0961 |
| AR(1) | 4 | 352.6384 |
| AR(2) | 5 | 333.8741 |



# Regression in Time

Residual plot





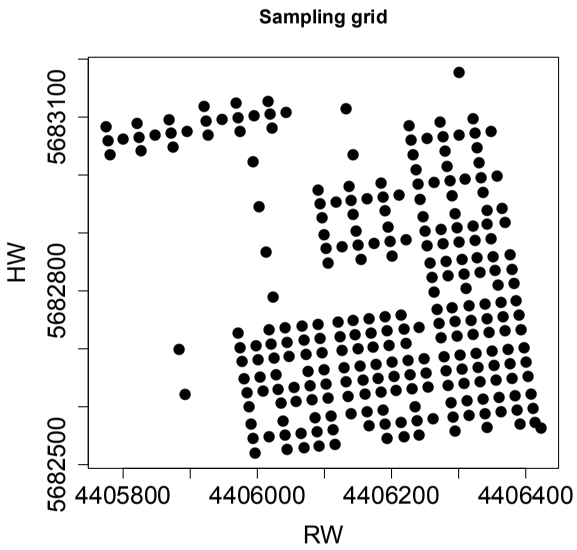



# Example

Soil carbon concentration at Mehrstedt

- initial soil sampling for BIOTREE experiment in 2003/2004
- extensively managed grassland
- regular sampling grid ($24m \times 24m$) over a $17ha$ site
- 284 soil cores
- organic carbon concentration at $5cm$ depth

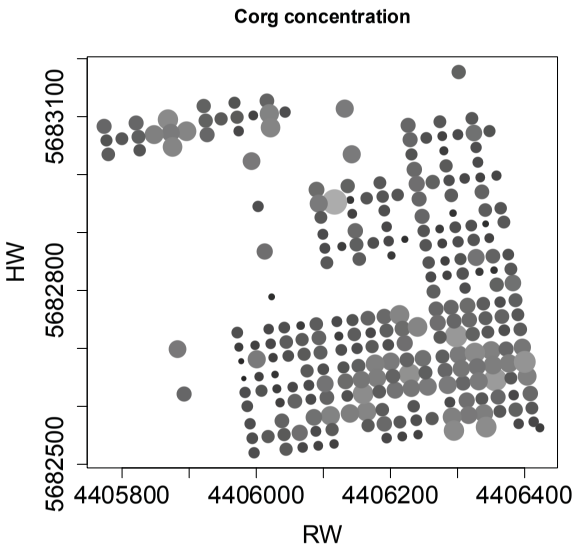
# Mehrstedt





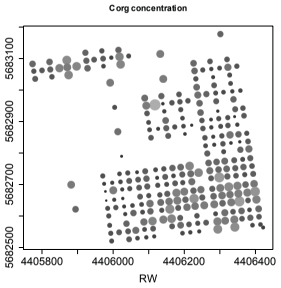

# Mehrstedt

Organic carbon concentration






# Mehrstedt




- nearby samples tend to be similar
- they are stochastically dependent
- consequences for accuracy of estimates
- **large scale** and **small scale** variation
- spatial analysis tries to separate large-scale and small-scale variation



# Mehrstedt

Quantification of spatial dependence

$$\begin{aligned} Var(Y_i - Y_j) &= Var(Y_i) + Var(Y_j) - \mathbf{2cov}(\mathbf{Y_i}, \mathbf{Y_j}) \\ &= 2\sigma^2 - 2cov(Y_i, Y_j) \end{aligned}$$

Stationarity

$$= 2\gamma(h)$$

Isotropy

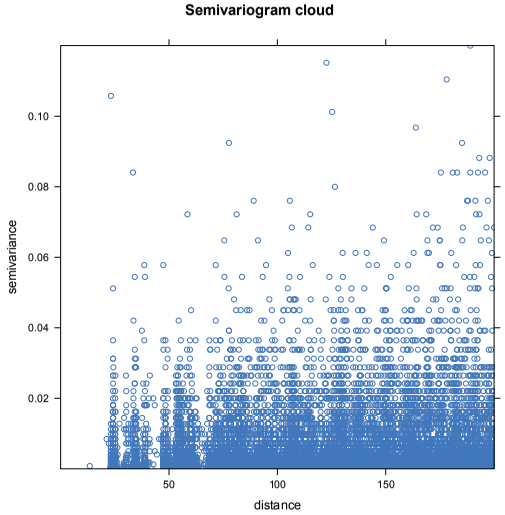
$h$ - distance between $Y_i$ and $Y_j$

**Semivariogram:**

- $\gamma(h)$ as a function of the distance between points is called the **Semivariogram** or simply **Variogram**.
- Factor 2 ensures that we can *see* the underlying variance $\sigma^2$ if we estimate $\gamma(h)$.

# Estimation of the semivariogram

Variogram Cloud: plot the squared differences of measured values as a function of distance




Applied Statistics
J.Schumacher

# Estimation of the semivariogram

Empirical semivariogram

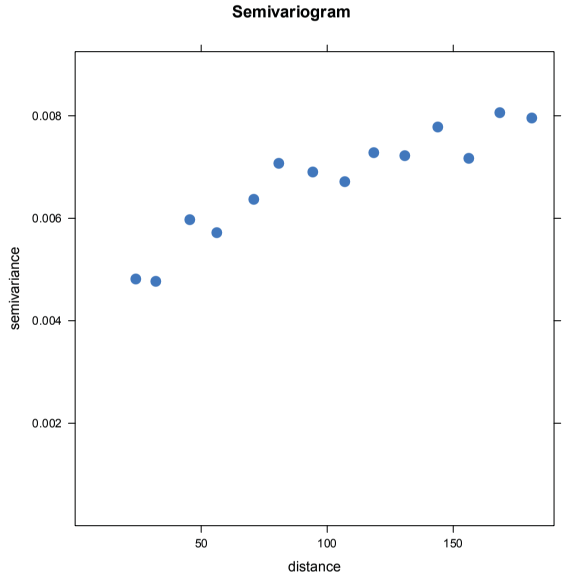
- divide distance range into bins
- calculate mean within bins

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} (y_i - y_j)^2$$

Applied Statistics
J.Schumacher

# Estimation of the semivariogram

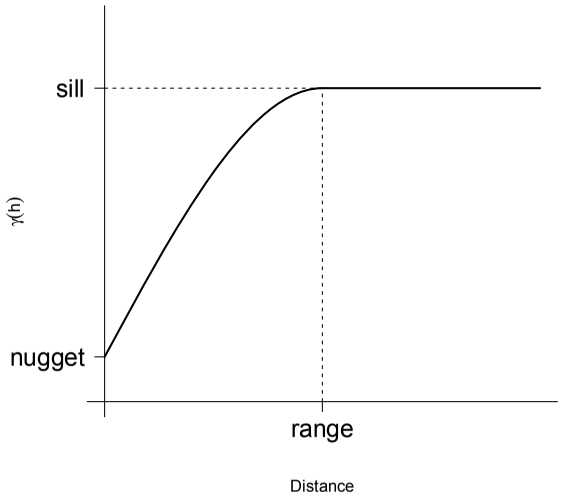
Empirical semivariogram




Applied Statistics
J.Schumacher

# Estimation of the semivariogram

Typical Semivariogram



# Semivariogram

Characteristics of a semivariogram

**nugget effect**
micro-scale variation, measurement error, variability if sampling would be repeated at exactly the same location

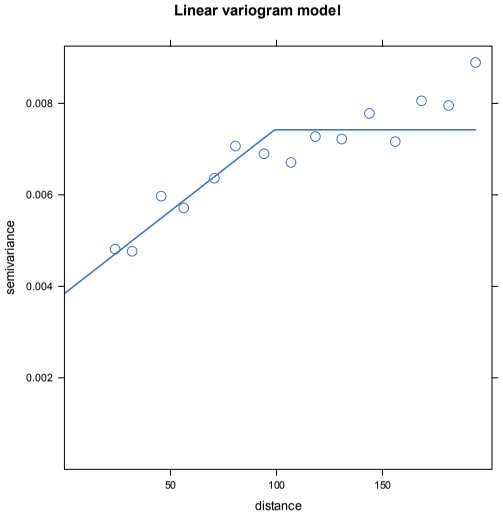
**sill** asymptotic value of $\gamma(h)$, underlying variability, variability between independent measurements

**range**
the distance at which data can be considered independent



# Variogram models - Linear variogram

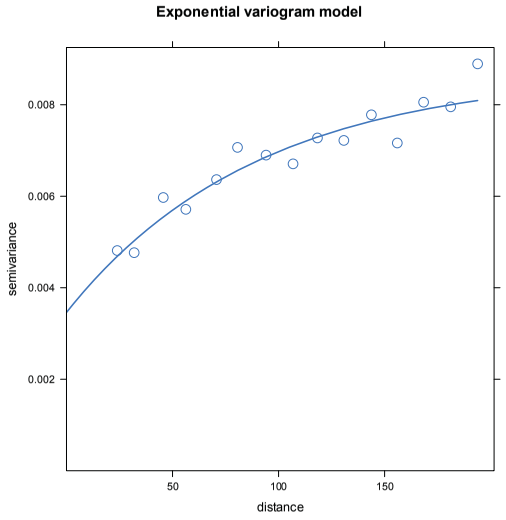
$$\gamma(h) = \begin{cases} a+(\sigma^2-a)*\left(1-\left(1-\frac{h}{r}\right)\right) & 0<h\leq r \\ \sigma^2 & \text{otherwise} \end{cases}$$






# Variogram models - Exponential variogram

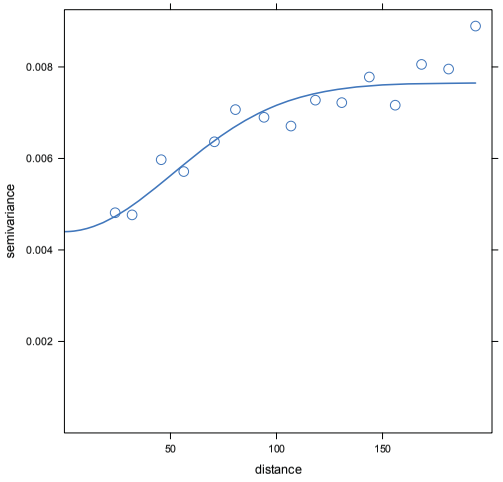
$$\gamma(h) = a + (\sigma^2 - a) * \left(1 - \exp\{-\frac{3h}{r}\}\right)$$





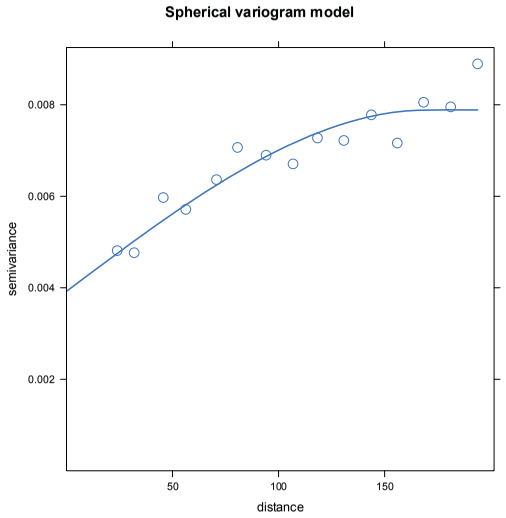

# Variogram models - Gaussian variogram

$$\gamma(h) = a + (\sigma^2 - a) * \left(1 - \exp\{-\frac{3h^2}{r^2}\}\right)$$



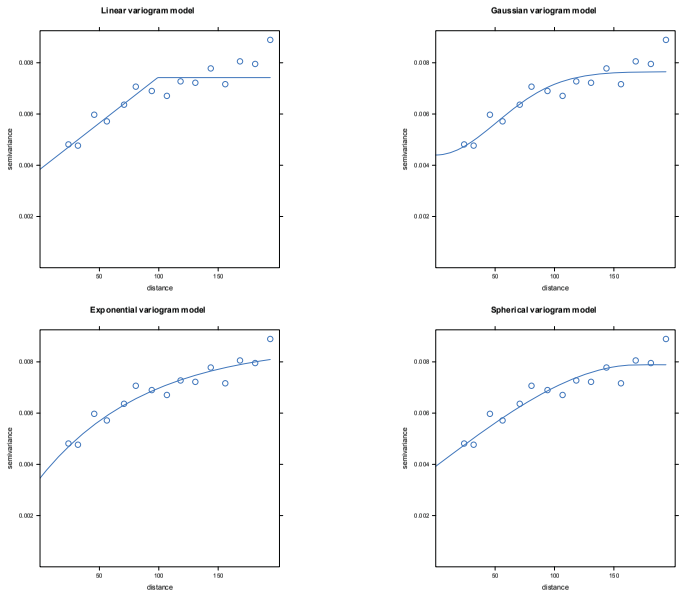
# Variogram models - Spherical variogram

$$\gamma(h) = \begin{cases} a+(\sigma^2-a)*\left(1-1.5\frac{h}{r}+0.5\left(\frac{h}{r}\right)^3\right) & 0<h\leq r \\ \sigma^2 & \text{otherwise} \end{cases}$$





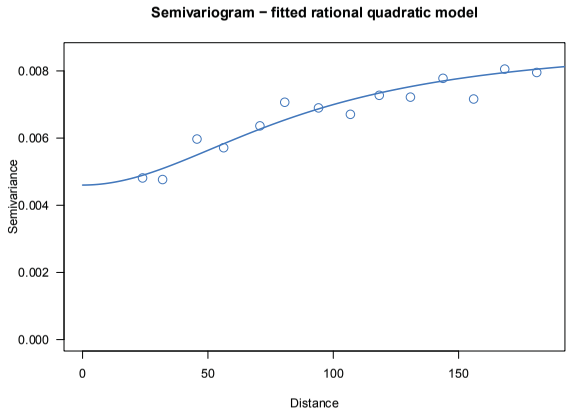

# Fitted variogram models






# Variogram models - Rational quadratic

$$\gamma(h) = a + (\sigma^2 - a) * \frac{(h/r)^2}{1 + (h/r)^2}$$

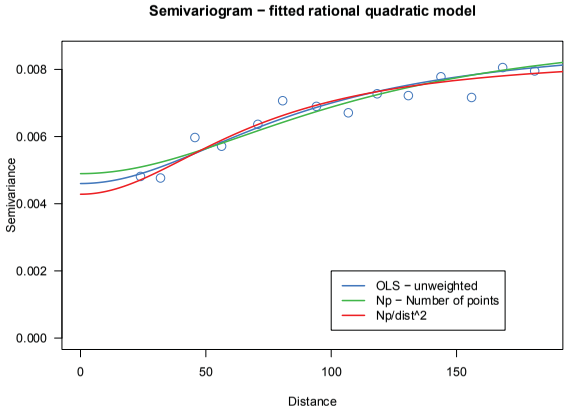





# Fitting variogram models

How are semivariogram models fitted?

- nonlinear regression problem
- different numbers of point pairs in differents bins
- larger variability at larger distances
- $\Longrightarrow$ **weighted nonlinear least squares**



# Semivariogram

Selecting a best variogram model

- it is **very important** to account for spatial dependence
- it is less important, **how** you account for spatial dependence
- differences between models often appear at small distances where no observations available
- visual inspection of model fit



# Consequences of spatial dependence

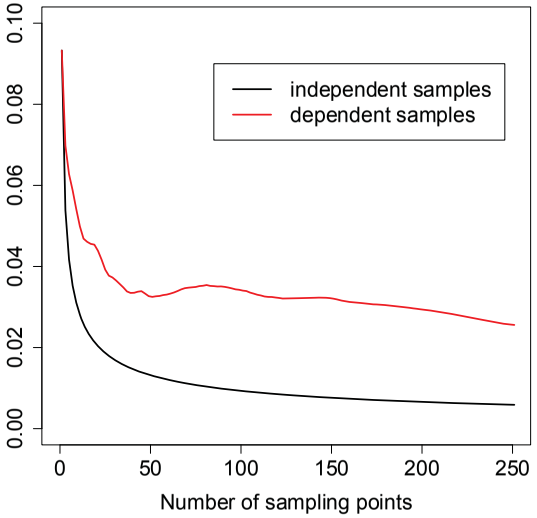
How accurate is the estimate of mean carbon concentration?

- Independent samples: standard error = $\frac{\sigma^2}{\sqrt{\text{sample size}}}$
- Autocorrelated samples:
  - randomly selected sampling points on the grid
  - use fitted exponential semivariogram model for modeling spatial dependence
- **effective sample size** = How many independent samples achieve the same accuracy?



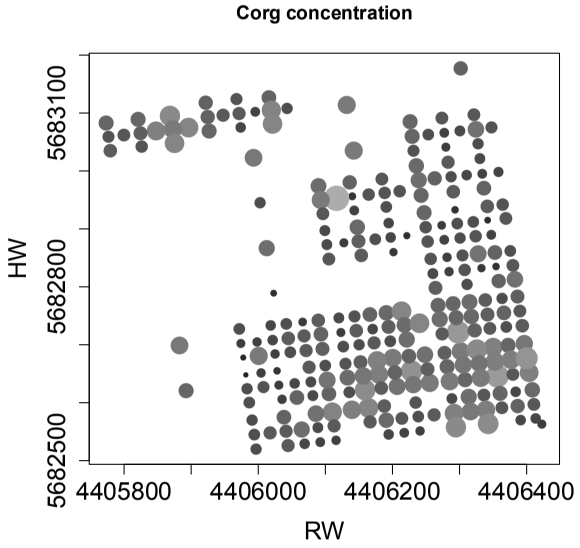
# Consequences of spatial dependence

Accuracy of estimates



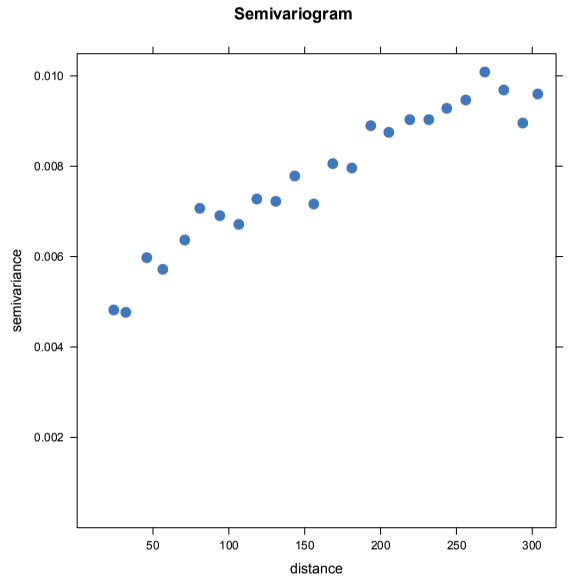



# Fitting the large scale trend



# Fitting the large-scale trend

## Identifying non-stationarity




Applied Statistics
J.Schumacher

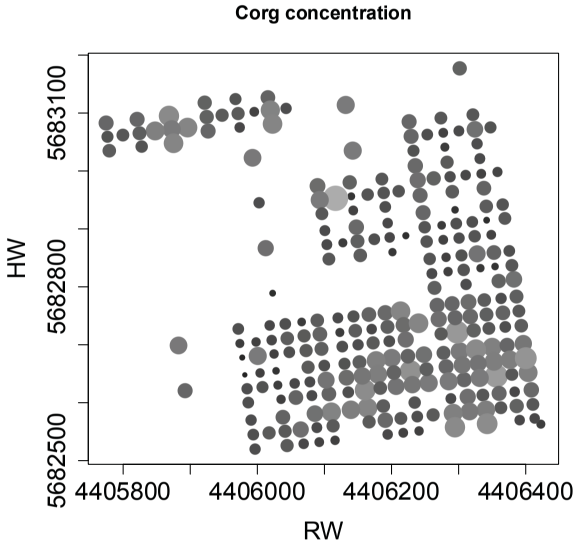
# Fitting the large-scale trend

Identifying non-stationarity

- no saturation
- the more distant the points, the more different their measured values
- indication of systematic trend over longer distances

Applied Statistics
J.Schumacher

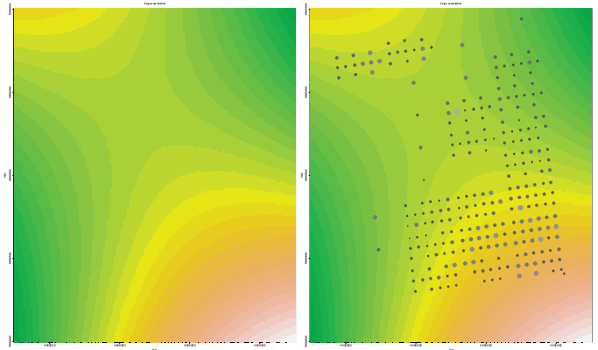
# Fitting the large scale trend




Applied Statistics
J.Schumacher

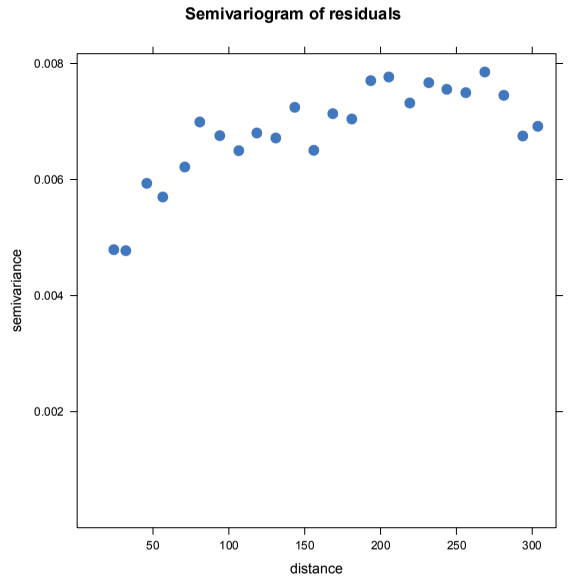
# Fitting the large scale trend

Quadratic surface



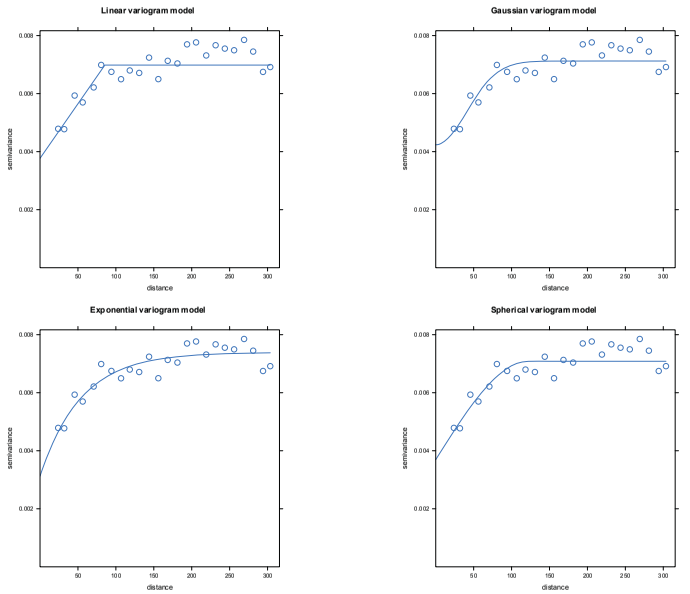
# Fitting the large scale trend

## Remaining small-scale variation






# Fitting the large scale trend






# Spatial prediction - Kriging

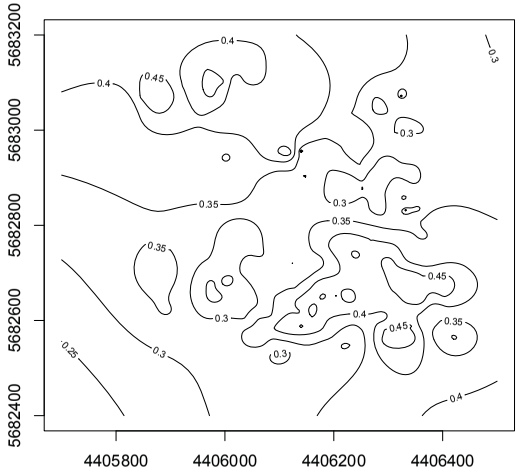
Combining large and small scales

- quadratic regression model predicts the general large scale pattern
- predictions can be improved by taking into account the small scale dependence
- $\longrightarrow$ **Kriging**



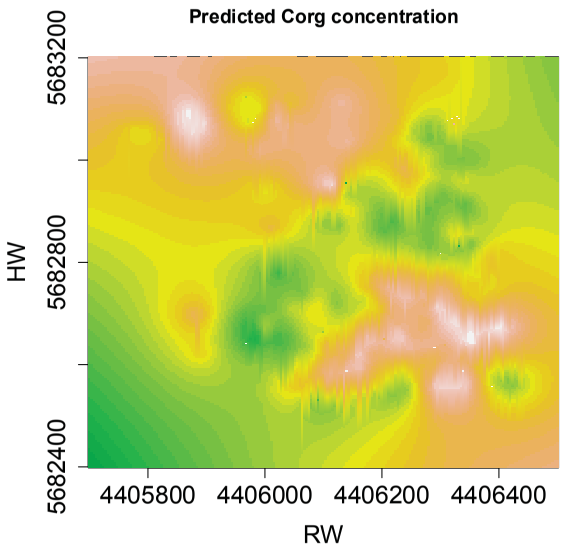
# Spatial prediction - Kriging

## Combining large and small scales



# Spatial prediction - Kriging

Combining large and small scales

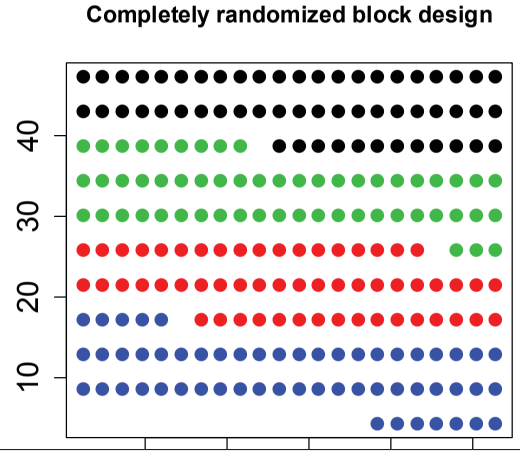





# Incorporating spatial dependence into statistical analysis

Wheat yield example

- Wheat yield trial with 56 different varieties
- completely randomized block design with 4 blocks






# Wheat yield example

Traditional ANOVA analysis

```
             Df Sum Sq Mean Sq F value     Pr(>F)
Block         3 1809.1  603.03 12.1621 3.127e-07 ***
variety      55 2387.5   43.41  0.8755     0.7119
Residuals   165 8181.1   49.58
```

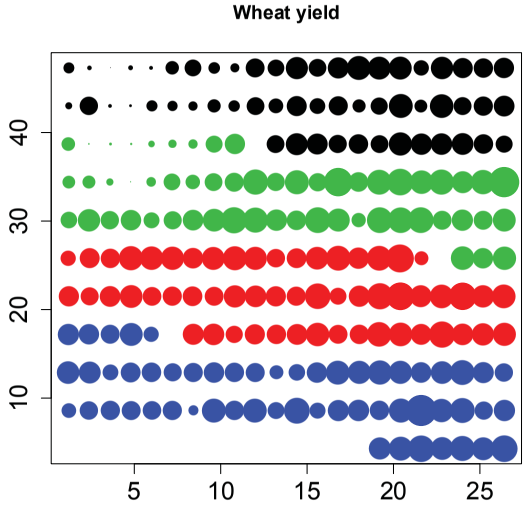
- statistically significant block effect
- **NO** statistically significant yield differences between varieties

```
   NE83432     NE87499     NE84557 ROUGHRIDER        CODY
   19.7250     20.4125     20.5250     21.1875     21.2125
...
  NE83498 CENTURK78    REDLAND    NE86501    NE87619    NE86503
  30.1250    30.3000    30.5000    30.9375    31.2625    32.6500
```



# Wheat yield example

A closer look



## Wheat yield example

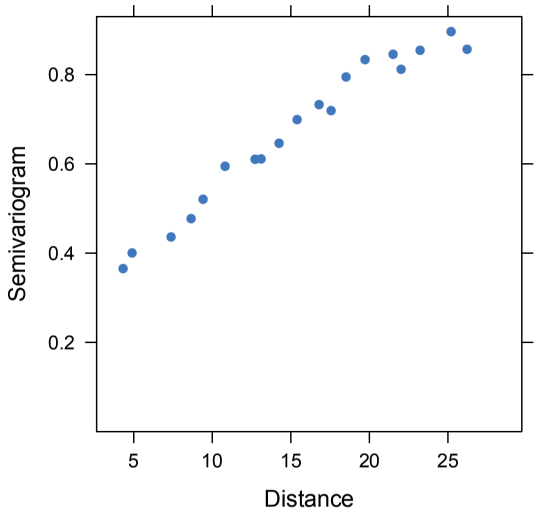
A closer look

- blocks are relatively large
- blocks are not really homogeneous
- there is small scale variation without any apparent trend



## Wheat yield example

Semivariogram of residuals






## Wheat yield example

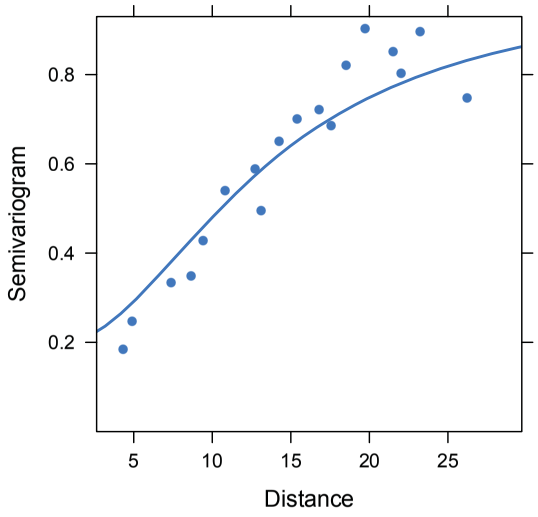
Incorporating spatial dependence into statistical analysis

- Null hypotheses: no yield differences between varieties
- Alternative hypotheses: yields are different
- Rational quadratic semivariogram model



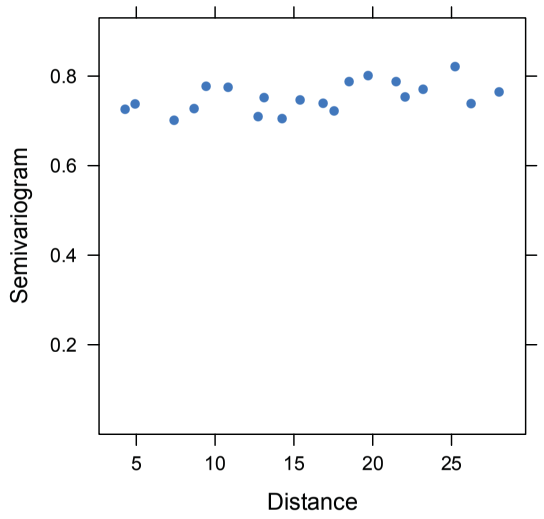
## Wheat yield example

Rational quadratic semivariogram



# Model validation

Variogram of standardized residuals






# Wheat yield example

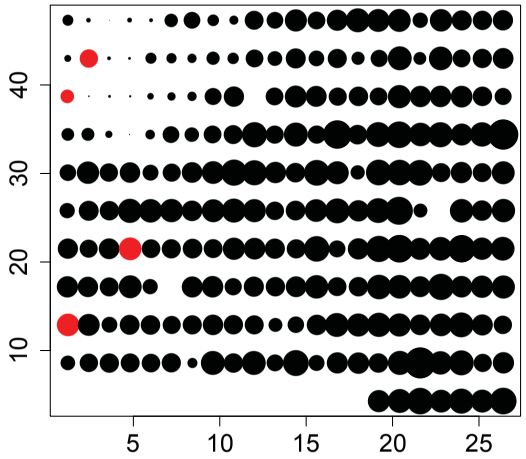
Incorporating spatial dependence into statistical analysis

```
             Model df      AIC      logLik  L.Ratio p-value
Constant         1  4 1358.815  -675.4074
Variety          2 59 1363.794  -622.8971 105.0207   1e-04
...
varietyNE87613 varietyNE85556 varietyNE87612
      27.73219       27.91699       28.32995

 varietyNE87619  varietyNE83498 varietyBUCKSKIN
       28.66144        29.08406        35.03727
```



# Wheat yield example

The winner is ... **BUCKSKIN**






# Spatial dependence - Summary

- Data obtained close together are likely to be correlated
- correlated data reduce the accuracy of estimated means
- the degree of spatial autocorrelation can be quantified by empirical semivariograms
- semivariogram models can be fitted to empirical semivariograms
- incorporating spatial correlation leads to reliable estimates of accuracy
- incorporating spatial correlation can help to reveal hidden information

## What are linear mixed models?

- linear mixed models are linear models where fixed and random effects are mixed

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

- linear mixed models are linear models with more than one random effect



## Fixed and Random Effects

Experimental factors

Fixed factors

- all levels of interest are included in the analysis
- all levels included in the analysis are of direct interest
- e.g. experimental factors (fertilizer, land-use, light)
- exactly the same levels would be used in a repeated experiment

Random factors

- level can be considered as randomly selected
- e.g. randomly selected experimental units (plots, genotypes, individuals)
- different levels would be used in a repeated experiment
- levels are not informative (arbitrary id number)



## Fixed and random effects

Continuous explanatory variables

Fixed effects

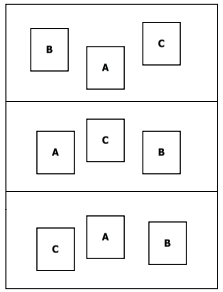
- interest is in quantification of the effects induced by changes in explanatory variable

Random effects

- individual-specific random modification of fixed effect

# Design approach to LMM

Block ANOVA design




Model:

$$y_{ij} = \mu + \underbrace{\alpha_j}_{\text{treatment effect}} + \underbrace{b_i}_{\text{block effect}} + \varepsilon_{ij}$$

- differences between blocks contribute to variability
- but we are not really interested in the effect of particular block
- nuisance parameter
- random effect



# Design approach to LMM

Block ANOVA model

$$y_{ij} = \mu + \alpha_j + b_i + \varepsilon_{ij}$$
$$b_i \sim N(0, \sigma_b^2) \qquad b_i \text{ unabhängig}$$
$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \qquad \varepsilon_{ij} \text{ unabhängig, auch von } b_i$$

Stochastic dependence structure:

- observations from different blocks:

$$\begin{aligned} cov(y_{ij}, y_{kl}) &= cov(\mu + \alpha_j + b_i + \varepsilon_{ij}, \mu + \alpha_l + b_k + \varepsilon_{kl}) \\ &= cov(b_i + \varepsilon_{ij}, b_k + \varepsilon_{kl}) \\ &= cov(b_i, b_k) + cov(b_i, \varepsilon_{kl}) \\ &\qquad + cov(\varepsilon_{ij}, b_k) + cov(\varepsilon_{ij}, \varepsilon_{kl}) \\ &= 0 \end{aligned}$$

- observations from different blocks are stochastically independent



# Design approach to LMM

Block ANOVA model

$$y_{ij} = \mu + \alpha_j + b_i + \varepsilon_{ij}$$
$$b_i \sim N(0, \sigma_b^2) \qquad b_i \text{ unabhängig}$$
$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \qquad \varepsilon_{ij} \text{ unabhängig, auch von } b_i$$

Stochastic dependence structure:

- observations from the same block:

$$\begin{aligned} cov(y_{ij}, y_{il}) &= cov(\mu + \alpha_j + b_i + \varepsilon_{ij}, \mu + \alpha_l + b_i + \varepsilon_{il}) \\ &= cov(b_i + \varepsilon_{ij}, b_i + \varepsilon_{il}) \\ &= cov(b_i, b_i) + cov(b_i, \varepsilon_{il}) \\ &\qquad + cov(\varepsilon_{ij}, b_i) + cov(\varepsilon_{ij}, \varepsilon_{il}) \\ &= \sigma_b^2 \end{aligned}$$

$$Var(y_{ij}) = cov(y_{ij}, y_{ij}) = \sigma_b^2 + \sigma_\varepsilon^2$$



# Design approach to LMM

- observations from the same blocks are stochastically dependent
- intraclass correlation coefficient:

$$cor(y_{ij}, y_{il}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\varepsilon^2}$$

# Design approach to LMM

Block ANOVA - Covariance matrix:

$$\begin{pmatrix} \sigma_b^2+\sigma_\varepsilon^2 & \sigma_b^2 & \cdots & \sigma_b^2 & 0 & \cdots & 0 \\ \sigma_b^2 & \sigma_b^2+\sigma_\varepsilon^2 & \ddots & \sigma_b^2 & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \vdots & & & \\ \sigma_b^2 & \cdots & \sigma_b^2 & \sigma_b^2+\sigma_\varepsilon^2 & 0 & \cdots & 0 \\ & & & & \ddots & & \\ 0 & \cdots & 0 & & \sigma_b^2+\sigma_\varepsilon^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \vdots & & \vdots & & \sigma_b^2 & \sigma_b^2+\sigma_\varepsilon^2 & \ddots & \sigma_b^2 \\ & & & & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & & \sigma_b^2 & \cdots & \sigma_b^2 & \sigma_b^2+\sigma_\varepsilon^2 \end{pmatrix}$$

block diagonal matrix



# Design approch to LMM

Split-plot design I - Biodiversity experiment

- experimental plots differ in diversity: species richness: 1,2,4,8,16
- „phytometer“ are planted in each plot:
  - one individual of *Trifolium repens*
  - one individual of *Festuca pratensis*
- interest is in
  - diversity effects
  - legume/grass differences



# Design approch to LMM

Split-plot design I - Biodiversity experiment

- the two individuals planted in the same plot share all plot-specific peculiarities
- $\longrightarrow$ random plot effect
- Model:

$$y_{ijk} = \mu + \underbrace{\alpha_i}_{\text{diversity}} + \underbrace{\beta_j}_{\text{Legume/grass}} + \underbrace{b_k}_{\text{Plot}} + \varepsilon_{ijk}$$

- formal model description syntax:

  $y \sim$ diversity + species + (1|plot)

  $y \sim$ diversity + species, random = $\sim$ 1|plot



# Design approch to LMM

Split-plot design II - Biodiversity experiment

- experimental plots differ in diversity: species richness: 1,2,4,8,16
- „phytometer“ are planted in each plot:
  - one individual of *Trifolium repens*
  - one individual of *Festuca pratensis*
- **BUT:**
  - five leafs per plant are analysed

## Design approch to LMM

Split-plot design II

- the two individuals planted in the same plot share all plot-specific peculiarities
- the five leafs share a common plant effect
- results would be different if we harvested one leaf from 5 different plants
- the plant individuals belong to exactly one plot
- $\longrightarrow$ nested random effects
- Model:
- formal model description syntax:

$y \sim$ diversity + species + (1|plot)+(1|plot:plant.id)

$y \sim$ diversity + species, random = $\sim$ 1|plot/plant.id



## Design approch to LMM

Split-plot design III

- experimental plots differ in diversity: species richness: 1,2,4,8,16
- „phytometer“ are planted in each plot:
  - one individual each of *10 grass species*
  - one individual of *10 legume species*
- species are **randomly** selected to represent variability within the functional group
- species are not nested within plots because the same species occur on all plots
- $\longrightarrow$ **crossed random effects**
- formal model description syntax:

$y \sim$ diversity + species + (1|plot)+(1|species)

$y \sim$ diversity + species, random = ???



## Example: irrigation trial

- four irrigation systems
- can only be applied to large fields
- 8 fields
- 2 varieties of wheat are sown on 2 x 3 smaller plots per field
- What are fixed and random factors?



## Regression approach to LMM

Linear regression model

$$y_{ij} = \beta_0 + \beta_1 x_{1,ij} + \ldots + \beta_k x_{k,ij} + \varepsilon_{ij}$$

- second index indicates, that there is a group structure
- this grouping is induced by a random factor
- regression coefficients may show small group-specific random deviations

## Regression approach to LMM

Linear mixed model - Random intercept model

- random differences between groups occur only in the intercept

$$y_{ij} = \beta_0 + \gamma_i + \beta_1 x_{1,ij} + \ldots + \beta_k x_{k,ij} + \varepsilon_{ij}$$

- block design and split-plot design I are examples of random intercept models



## Regression approach to LMM

Linear mixed model - Random intercept and slope model

- some of the other regression coefficients (or even all) show random group-specific modifications

$$y_{ij} = (\beta_0 + \gamma_{0,i}) + (\beta_1 + \gamma_{1,i})\, x_{1,ij} + \ldots + \beta_k x_{k,ij} + \varepsilon_{ij}$$

- fixed and random effects are usually separated

$$y_{ij} = \beta_0 + \beta_1 x_{1,ij} + \ldots + \beta_k x_{k,ij} + \gamma_{0,i} + \gamma_{1,i} + \varepsilon_{ij}$$

- formal model description syntax:

$$y \sim 1 + x_1 + \ldots + x_k + (1 + x_1 | group)$$
$$y \sim 1 + x_1 + \ldots + x_k, \text{ random = } \sim 1 + x_1 | group)$$



## Regression approach to LMM

Stochastic dependence between random effects

$$cov \begin{pmatrix} \gamma_{0,i} \\ \gamma_{1,i} \end{pmatrix} = \Sigma_\gamma = \begin{pmatrix} \sigma^2_{\gamma_0} & \rho\sigma_{\gamma_0}\sigma_{\gamma_1} \\ \rho\sigma_{\gamma_0}\sigma_{\gamma_1} & \sigma^2_{\gamma_1} \end{pmatrix}$$

- fast increase in number of parameters



## Parameter estimation

Maximum likelihood principle

- if all model parameters were known, one could calculate the probability of observing exactly the data observed
- since model parameters are not known, it is reasonable to take those parameter values as estimates which maximize the probability of observing exactly our data
- two versions of likelihood-based estimation:
  - maximum likelihood estimation (ML) in the strict sense
  - restricted maximum likelihood estimation (REML)

## Parameter estimation

ML vs. REML estimation

- ML (maximum likelihood) tends to underestimate variance parameters of the model
- REML (restricted maximum likelihood) usually leads to improved estimates (less biased) of variance parameters
- model selection concerning fixed effects works only with ML
- REML especially advisable if estimates of variance components are of direct interest
- REML can be interpreted as maximum likelihood restricted to variance parameters



## Parameter estimation

ML vs. REML example

$$y_i \sim N(\mu, \sigma^2), \qquad i = 1, \ldots n$$

Variance estimates:

- ML:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- REML:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \text{unbiased estimate}$$



## A modeling strategy

- fixed and random effects have to be treated separately
- we are usually more interested in fixed effects
- random effects are nuisance parameters, have to be incorporated to describe dependence structure
- keep random effects model simple - quickly rising number of parameters
- interference between model selection for fixed and random effects



## A modeling strategy

Step I: Selection of a random effects structure

- start with a (too) large fixed effect model
- start with a minimal random effects model reflecting the structure of the data
- fit models with different random effect structures using REML
- compare hierarchical models with likelihood ratio tests or use AIC to find an optimal random effects structure

## A modeling strategy

Step II: Selection of a fixed effects structure

- comparison of models with different fixed effects must be based on ML estimation of parameters
- compare hierarchical models with likelihood ratio tests

Step III: Refit the model with REML

- if estimates for random effects are to be reported, the final model should be refit using REML

Applied Statistics
J.Schumacher

## Variance Homogeneity