# Multivariate Statistics in Ecology and Quantitative Genetics **Summary**

Dirk Metzler & Martin Hutzenthaler

http://evol.bio.lmu.de/_statgen

20. July 2012

# Contents

Anova

Linear Models

Generalized Linear Models

Mixed-effects models

Principal Component Analysis (PCA)

Redundancy Analysis (RDA)

Correspondence Analysis (CA)

Canonical Correspondence Analysis (CCA)

QTL Mapping

# Contents

Anova

Linear Models

Generalized Linear Models

Mixed-effects models

Principal Component Analysis (PCA)

Redundancy Analysis (RDA)

Correspondence Analysis (CA)

Canonical Correspondence Analysis (CCA)

QTL Mapping

# Example

Blood-clotting times in rats under 4 different treatments

| gr. | $\overline{x}_{i\cdot}$ | observations | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 61 | 62 | 60 | 63 | 59 | | | | |
| | | $(62-61)^2$ | $(60-61)^2$ | $(63-61)^2$ | $(59-61)^2$ | | | | |
| 2 | 66 | 63 | 67 | 71 | 64 | 65 | 66 | | |
| | | $(63-66)^2$ | $(67-66)^2$ | $(71-66)^2$ | $(64-66)^2$ | $(65-66)^2$ | $(66-66)^2$ | | |
| 3 | 68 | 68 | 66 | 71 | 67 | 68 | 68 | | |
| | | $(68-68)^2$ | $(66-68)^2$ | $(71-68)^2$ | $(67-68)^2$ | $(68-68)^2$ | $(68-68)^2$ | | |
| 4 | 61 | 56 | 62 | 60 | 61 | 63 | 64 | 63 | 59 |
| | | $(56-61)^2$ | $(62-61)^2$ | $(60-61)^2$ | $(61-61)^2$ | $(63-61)^2$ | $(64-61)^2$ | $(63-61)^2$ | $(59-61)^2$ |

global mean $\overline{x}_{\cdot\cdot} = 64$,
group means $\overline{x}_{1\cdot} = 61$, $\overline{x}_{2\cdot} = 66$, $\overline{x}_{3\cdot} = 68$, $\overline{x}_{4\cdot} = 61$.

# Example

Blood-clotting times in rats under 4 different treatments

| gr. | $\overline{x}_{i.}$ | observations | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 61 | 62 | 60 | 63 | 59 | | | | |
| | | $(62-61)^2$ | $(60-61)^2$ | $(63-61)^2$ | $(59-61)^2$ | | | | |
| 2 | 66 | 63 | 67 | 71 | 64 | 65 | 66 | | |
| | | $(63-66)^2$ | $(67-66)^2$ | $(71-66)^2$ | $(64-66)^2$ | $(65-66)^2$ | $(66-66)^2$ | | |
| 3 | 68 | 68 | 66 | 71 | 67 | 68 | 68 | | |
| | | $(68-68)^2$ | $(66-68)^2$ | $(71-68)^2$ | $(67-68)^2$ | $(68-68)^2$ | $(68-68)^2$ | | |
| 4 | 61 | 56 | 62 | 60 | 61 | 63 | 64 | 63 | 59 |
| | | $(56-61)^2$ | $(62-61)^2$ | $(60-61)^2$ | $(61-61)^2$ | $(63-61)^2$ | $(64-61)^2$ | $(63-61)^2$ | $(59-61)^2$ |

global mean $\overline{x}_{..} = 64$,
group means $\overline{x}_{1.} = 61$, $\overline{x}_{2.} = 66$, $\overline{x}_{3.} = 68$, $\overline{x}_{4.} = 61$.

The red Differences (unsquared) are the *residuals*: they are the residual variability which is not explained by the model.

# Example

Blood-clotting times in rats under 4 different treatments

| gr. | $\overline{x}_{i\cdot}$ | observations | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 61 | 62<br>$(62-61)^2$ | 60<br>$(60-61)^2$ | 63<br>$(63-61)^2$ | 59<br>$(59-61)^2$ | | | | |
| 2 | 66 | 63<br>$(63-66)^2$ | 67<br>$(67-66)^2$ | 71<br>$(71-66)^2$ | 64<br>$(64-66)^2$ | 65<br>$(65-66)^2$ | 66<br>$(66-66)^2$ | | |
| 3 | 68 | 68<br>$(68-68)^2$ | 66<br>$(66-68)^2$ | 71<br>$(71-68)^2$ | 67<br>$(67-68)^2$ | 68<br>$(68-68)^2$ | 68<br>$(68-68)^2$ | | |
| 4 | 61 | 56<br>$(56-61)^2$ | 62<br>$(62-61)^2$ | 60<br>$(60-61)^2$ | 61<br>$(61-61)^2$ | 63<br>$(63-61)^2$ | 64<br>$(64-61)^2$ | 63<br>$(63-61)^2$ | 59<br>$(59-61)^2$ |

global mean $\overline{x}_{\cdot\cdot} = 64$,
group means $\overline{x}_{1\cdot} = 61$, $\overline{x}_{2\cdot} = 66$, $\overline{x}_{3\cdot} = 68$, $\overline{x}_{4\cdot} = 61$.

The red Differences (unsquared) are the *residuals*: they are the residual variability which is not explained by the model.

Sums of squares within groups:

$ss_{\mathrm{within}} = 112$,

# Example

Blood-clotting times in rats under 4 different treatments

| gr. | $\overline{x}_{i\cdot}$ | observations | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 61 | 62<br>$(62-61)^2$ | 60<br>$(60-61)^2$ | 63<br>$(63-61)^2$ | 59<br>$(59-61)^2$ | | | | |
| 2 | 66 | 63<br>$(63-66)^2$ | 67<br>$(67-66)^2$ | 71<br>$(71-66)^2$ | 64<br>$(64-66)^2$ | 65<br>$(65-66)^2$ | 66<br>$(66-66)^2$ | | |
| 3 | 68 | 68<br>$(68-68)^2$ | 66<br>$(66-68)^2$ | 71<br>$(71-68)^2$ | 67<br>$(67-68)^2$ | 68<br>$(68-68)^2$ | 68<br>$(68-68)^2$ | | |
| 4 | 61 | 56<br>$(56-61)^2$ | 62<br>$(62-61)^2$ | 60<br>$(60-61)^2$ | 61<br>$(61-61)^2$ | 63<br>$(63-61)^2$ | 64<br>$(64-61)^2$ | 63<br>$(63-61)^2$ | 59<br>$(59-61)^2$ |

global mean $\overline{x}_{\cdot\cdot} = 64$,
group means $\overline{x}_{1\cdot} = 61$, $\overline{x}_{2\cdot} = 66$, $\overline{x}_{3\cdot} = 68$, $\overline{x}_{4\cdot} = 61$.

The red Differences (unsquared) are the *residuals*: they are the residual variability which is not explained by the model.

Sums of squares within groups:
$ss_{\mathrm{within}} = 112$, 20 degrees of freedom (df)

# Example

Blood-clotting times in rats under 4 different treatments

| gr. | $\bar{x}_{i\cdot}$ | observations | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 61 | 62<br>$(62-61)^2$ | 60<br>$(60-61)^2$ | 63<br>$(63-61)^2$ | 59<br>$(59-61)^2$ | | | | |
| 2 | 66 | 63<br>$(63-66)^2$ | 67<br>$(67-66)^2$ | 71<br>$(71-66)^2$ | 64<br>$(64-66)^2$ | 65<br>$(65-66)^2$ | 66<br>$(66-66)^2$ | | |
| 3 | 68 | 68<br>$(68-68)^2$ | 66<br>$(66-68)^2$ | 71<br>$(71-68)^2$ | 67<br>$(67-68)^2$ | 68<br>$(68-68)^2$ | 68<br>$(68-68)^2$ | | |
| 4 | 61 | 56<br>$(56-61)^2$ | 62<br>$(62-61)^2$ | 60<br>$(60-61)^2$ | 61<br>$(61-61)^2$ | 63<br>$(63-61)^2$ | 64<br>$(64-61)^2$ | 63<br>$(63-61)^2$ | 59<br>$(59-61)^2$ |

global mean $\bar{x}_{\cdot\cdot} = 64$,
group means $\bar{x}_{1\cdot} = 61$, $\bar{x}_{2\cdot} = 66$, $\bar{x}_{3\cdot} = 68$, $\bar{x}_{4\cdot} = 61$.

The red Differences (unsquared) are the *residuals*: they are the residual variability which is not explained by the model.

Sums of squares within groups:
$ss_{\mathrm{within}} = 112$, 20 degrees of freedom (df)
Sums of squares between groups:
$ss_{\mathrm{betw}} = 4 \cdot (61-64)^2 + 6 \cdot (66-64)^2 + 6 \cdot (68-64)^2 + 8 \cdot (61-64)^2 = 228$,

# Example

Blood-clotting times in rats under 4 different treatments

| gr. | $\overline{x}_{i\cdot}$ | observations | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 61 | 62 | 60 | 63 | 59 | | | | |
| | | $(62-61)^2$ | $(60-61)^2$ | $(63-61)^2$ | $(59-61)^2$ | | | | |
| 2 | 66 | 63 | 67 | 71 | 64 | 65 | 66 | | |
| | | $(63-66)^2$ | $(67-66)^2$ | $(71-66)^2$ | $(64-66)^2$ | $(65-66)^2$ | $(66-66)^2$ | | |
| 3 | 68 | 68 | 66 | 71 | 67 | 68 | 68 | | |
| | | $(68-68)^2$ | $(66-68)^2$ | $(71-68)^2$ | $(67-68)^2$ | $(68-68)^2$ | $(68-68)^2$ | | |
| 4 | 61 | 56 | 62 | 60 | 61 | 63 | 64 | 63 | 59 |
| | | $(56-61)^2$ | $(62-61)^2$ | $(60-61)^2$ | $(61-61)^2$ | $(63-61)^2$ | $(64-61)^2$ | $(63-61)^2$ | $(59-61)^2$ |

global mean $\overline{x}_{\cdot\cdot} = 64$,
group means $\overline{x}_{1\cdot} = 61$, $\overline{x}_{2\cdot} = 66$, $\overline{x}_{3\cdot} = 68$, $\overline{x}_{4\cdot} = 61$.

The red Differences (unsquared) are the *residuals*: they are the residual variability which is not explained by the model.

Sums of squares within groups:
$ss_{\mathrm{within}} = 112$, 20 degrees of freedom (df)
Sums of squares between groups:
$ss_{\mathrm{betw}} = 4 \cdot (61-64)^2 + 6 \cdot (66-64)^2 + 6 \cdot (68-64)^2 + 8 \cdot (61-64)^2 = 228$,
3 degrees of freedom (df)

# Example

Blood-clotting times in rats under 4 different treatments

| gr. | $\overline{x}_{i\cdot}$ | observations | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 61 | 62<br>$(62-61)^2$ | 60<br>$(60-61)^2$ | 63<br>$(63-61)^2$ | 59<br>$(59-61)^2$ | | | | |
| 2 | 66 | 63<br>$(63-66)^2$ | 67<br>$(67-66)^2$ | 71<br>$(71-66)^2$ | 64<br>$(64-66)^2$ | 65<br>$(65-66)^2$ | 66<br>$(66-66)^2$ | | |
| 3 | 68 | 68<br>$(68-68)^2$ | 66<br>$(66-68)^2$ | 71<br>$(71-68)^2$ | 67<br>$(67-68)^2$ | 68<br>$(68-68)^2$ | 68<br>$(68-68)^2$ | | |
| 4 | 61 | 56<br>$(56-61)^2$ | 62<br>$(62-61)^2$ | 60<br>$(60-61)^2$ | 61<br>$(61-61)^2$ | 63<br>$(63-61)^2$ | 64<br>$(64-61)^2$ | 63<br>$(63-61)^2$ | 59<br>$(59-61)^2$ |

global mean $\overline{x}_{\cdot\cdot} = 64$,
group means $\overline{x}_{1\cdot} = 61$, $\overline{x}_{2\cdot} = 66$, $\overline{x}_{3\cdot} = 68$, $\overline{x}_{4\cdot} = 61$.

The red Differences (unsquared) are the *residuals*: they are the residual variability which is not explained by the model.

Sums of squares within groups:
$ss_{\mathrm{within}} = 112$, 20 degrees of freedom (df)
Sums of squares between groups:
$ss_{\mathrm{betw}} = 4 \cdot (61-64)^2 + 6 \cdot (66-64)^2 + 6 \cdot (68-64)^2 + 8 \cdot (61-64)^2 = 228$,
3 degrees of freedom (df)

$$F = \frac{ss_{\mathrm{betw}}/3}{ss_{\mathrm{within}}/20} = \frac{76}{5.6} = 13.57$$

## Example: Blood-clotting times in rats under 4 different treatments.

ANOVA table („ANalysis Of VAriance“)

| | df | sum of squares (ss) | mean sum of squares (ss/df) | *F* value |
|---|---|---|---|---|
| groups | 3 | 228 | 76 | 13.57 |
| residuals | 20 | 112 | 5.6 | |

## Example: Blood-clotting times in rats under 4 different treatments.

ANOVA table („ANalysis Of VAriance“)

| | df | sum of squares (ss) | mean sum of squares (ss/df) | $F$ value |
|---|---|---|---|---|
| groups | 3 | 228 | 76 | 13.57 |
| residuals | 20 | 112 | 5.6 | |

Under the Null-Hypothesis $H_0$ “the group means are equal”
(and assuming independent, normally distributed observations)
is $F$ Fisher-distributed with 3 and 20 degrees of freedom, and
$p = \mathrm{Fisher}_{3,20}([13.57, \infty)) \leq 5 \cdot 10^{-5}$.

## Example: Blood-clotting times in rats under 4 different treatments.

ANOVA table („ANalysis Of VAriance“)

| | df | sum of squares (ss) | mean sum of squares (ss/df) | *F* value |
|---|---|---|---|---|
| groups | 3 | 228 | 76 | 13.57 |
| residuals | 20 | 112 | 5.6 | |

Under the Null-Hypothesis $H_0$ “the group means are equal”
(and assuming independent, normally distributed observations)
is $F$ Fisher-distributed with 3 and 20 degrees of freedom, and
$p = \mathrm{Fisher}_{3,20}([13.57, \infty)) \leq 5 \cdot 10^{-5}$.

Thus, we can reject $H_0$.

# $F$-Test

$n = n_1 + n_2 + \cdots + n_I$ obersvations in $I$ groups,
$X_{ij} = j$-th observation in $i$-th group, $j = 1, \ldots, n_i$.

## $F$-Test

$n = n_1 + n_2 + \cdots + n_I$ obersvations in $I$ groups,
$X_{ij} = j$-th observation in $i$-th group, $j = 1, \ldots, n_i$.

Model assumption: $X_{ij} = \mu_i + \varepsilon_{ij}$,
with independent, normally distributed $\varepsilon_{ij}$, $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

## *F*-Test

$n = n_1 + n_2 + \cdots + n_I$ obersvations in $I$ groups,
$X_{ij} = j$-th observation in $i$-th group, $j = 1, \ldots, n_i$.

Model assumption: $X_{ij} = \mu_i + \varepsilon_{ij}$,
with independent, normally distributed $\varepsilon_{ij}$, $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

($\mu_i$ is the “true” mean within group $i$.)

## *F*-Test

$n = n_1 + n_2 + \cdots + n_I$ obersvations in $I$ groups,
$X_{ij} =$ $j$-th observation in $i$-th group, $j = 1, \ldots, n_i$.

Model assumption: $X_{ij} = \mu_i + \varepsilon_{ij}$,
with independent, normally distributed $\varepsilon_{ij}$, $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

($\mu_i$ is the “true” mean within group $i$.)

$\overline{X}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} X_{ij}$ (empirical) “global mean”
$\overline{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ (empirical) mean of group $i$

## $F$-Test

$n = n_1 + n_2 + \cdots + n_I$ obersvations in $I$ groups,
$X_{ij} = j$-th observation in $i$-th group, $j = 1, \ldots, n_i$.

Model assumption: $X_{ij} = \mu_i + \varepsilon_{ij}$,
with independent, normally distributed $\varepsilon_{ij}$, $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

($\mu_i$ is the “true” mean within group $i$.)

$\overline{X}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} X_{ij}$ (empirical) “global mean”
$\overline{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ (empirical) mean of group $i$

$$SS_{\mathrm{within}} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2$$

sum of squares within the groups,
$n - I$ degrees of freedom

$$SS_{\mathrm{betw}} = \sum_{i=1}^{I} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$

sum of squares between the groups,
$I - 1$ degrees of freedom

## *F*-Test

$n = n_1 + n_2 + \cdots + n_I$ obersvations in $I$ groups,
$X_{ij} = j$-th observation in $i$-th group, $j = 1, \ldots, n_i$.

Model assumption: $X_{ij} = \mu_i + \varepsilon_{ij}$,
with independent, normally distributed $\varepsilon_{ij}$, $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

($\mu_i$ is the “true” mean within group $i$.)

$\overline{X}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} X_{ij}$ (empirical) “global mean”
$\overline{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ (empirical) mean of group $i$

$$SS_{\mathrm{within}} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2$$

sum of squares within the groups,
$n - I$ degrees of freedom

$$SS_{\mathrm{betw}} = \sum_{i=1}^{I} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$

sum of squares between the groups,
$I - 1$ degrees of freedom

$$F = \frac{SS_{\mathrm{betw}}/(I-1)}{SS_{\mathrm{within}}/(n-I)}$$

## $F$-Test

$X_{ij} = j$-th observation $i$-th group, $j = 1, \ldots, n_i$,

Model assumption: $X_{ij} = \mu_i + \varepsilon_{ij}$. $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

$$SS_{\text{within}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2$$

sum of squares within groups,
$n - I$ degrees of feedom

$$SS_{\text{betw}} = \sum_{i=1}^{I} n_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$

sum of squares between groups,
$I - 1$ degrees of freedom

$$F = \frac{SS_{\text{betw}}/(I-1)}{SS_{\text{within}}/(n-I)}$$

Under the hypothesis $H_0 : \mu_1 = \cdots = \mu_I$ (“all $\mu_i$ are equal”)
$F$ is Fisher-distributed with $I - 1$ and $n - I$ degrees of freedom
(no matter what the true joint value of $\mu_i$ is).

$F$-Test: We reject $H_0$ on the level of significance $\alpha$ if $F \geq q_\alpha$, whereas $q_\alpha$ is the $(1 - \alpha)$-quantile of the Fisher-distribution with $I - 1$ and $n - I$ degrees of freedom.

```
> a <- aov(meas~flab)
> a
Call:
   aov(formula = meas ~ flab)

Terms:
                     flab Residuals
Sum of Squares  0.1247371 0.2314000
Deg. of Freedom         6        63

Residual standard error: 0.06060541
Estimated effects may be unbalanced
> summary(a)
            Df  Sum Sq  Mean Sq F value    Pr(>F)
flab         6 0.12474 0.020789  5.6601 9.453e-05 ***
Residuals   63 0.23140 0.003673
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

only for balanced designs:

```
> TukeyHSD(a)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = meas ~ flab)

$flab
      diff          lwr          upr     p adj
2-1 -0.065 -0.147546752  0.017546752 0.2165897
3-1 -0.059 -0.141546752  0.023546752 0.3226101
4-1 -0.142 -0.224546752 -0.059453248 0.0000396
5-1 -0.105 -0.187546752 -0.022453248 0.0045796
6-1 -0.107 -0.189546752 -0.024453248 0.0036211
7-1 -0.064 -0.146546752  0.018546752 0.2323813
3-2  0.006 -0.076546752  0.088546752 0.9999894
[...]
```

```
> kruskal.test(meas~flab)

Kruskal-Wallis rank sum test

data:  meas by flab
Kruskal-Wallis chi-squared = 29.606, df = 6, p-value = 4.
```

Let $i$ be the index for the row of a data table. The data are subdivided into groups and $G_i$ is the group row $i$ (or patient $i$) belongs to; e.g. $G_i$ can be the treatment of patient $i$. Let $Y_i$ be the response variable, e.g. the blood pressure of patient $i$. We can apply an anova to check whether $Y$ depends on $G$, and the model behind it is:

$$Y_i = b_{G_i} + \varepsilon_i$$

where the $\varepsilon_i$ are assumed to be independent and normally distributed with expectation 0, and all $\varepsilon_i$ have the same variance $\sigma^2$. During the ANOVA we estimate the influence $b_{G_i}$ of the group on $Y_i$ by the group mean $\widehat{b_g}$. Thus, the residuals $r_i := Y_i - \widehat{b_{G_i}} \approx Y_i - b_{G_i} = \varepsilon_i$ should be approximately normally distributed.

More than one factor can play a role. For example we may take into account that the blood pressure $Y_i$ of a patient may depend on the sex $S_i$ of the patient. In this case the model behind the anova takes the form

$$Y_i = b_{G_i} + c_{S_i} + \varepsilon_i.$$

$b_{G_i}$ depends only on the treatment group and $c_{S_i}$ only on the sex of the female. If we also want allow in *interaction* between the treatment and the sex, we need another variable $d_{G_i,S_i}$ that may depend on both:

$$Y_i = b_{G_i} + c_{S_i} + d_{G_i,S_i} + \varepsilon_i.$$

This makes possible, for example, that a certain treatment has a stronger effect for males than for females.

A *balanced design* means, that the sample size are the same for each combination of factors. E.g. 10 males and 10 females in each treatment group. Some ANOVA-based method will only work for balanced designs. Therefore, it is preferable to use a balanced design when planning an experiment. If the data, however, are observations from nature, the “design” is usually unbalanced and this has to be taken into account in the analysis.

One of the methods for which you need a balanced design is Tukey’s HSD (honest significat differences). From an anova it computes confidence intervals for the pairwise differences between the group means with mulptiple-testing correction (cf. R-script).

Another thing to be careful with is the interpretation of ANOVA tables. The R command anova, applied to a single model gives a so-called “Type I Anova”, where each line take only the variables in the lines above into account:

```
> anova(model4)
Analysis of Variance Table

Response: log(ccrt)
           Df  Sum Sq Mean Sq F value    Pr(>F)
line        1  1.2224 1.22238 13.1486 0.0003812 ***
day        11  2.8471 0.25883  2.7841 0.0023769 **
person      1  0.0850 0.08504  0.9147 0.3402393
[...]
```

For example, the p-value 0.0023769 tells how much better the model with line line and day can explain the data compared to a model that only takes line into account. Thus, the values assigned to variables depend on the input order.

If you use the R command drop1 with the option test="F", you get a so-called "Type II Anova", in which each line shows the influence of one variable, given the estimates of *all* other variables.

```
> drop1(model4,test="F")
[...]
      Df Sum of Sq    RSS     AIC F value    Pr(F)
<none>              15.618 -418.91
line    1   0.05860 15.677 -420.23  0.6304 0.428338
day    11   2.47080 18.089 -414.18  2.4161 0.008177 **
person  1   0.08504 15.703 -419.92  0.9147 0.340239
```

For example, the *p*-value 0.008177 says that a model that takes line, day and person into account explains the data significantly better than a model that uses only line and person.

It is often important to rescale (i.e. transform) the data. For example, if a comparison between fitted values (group means) and the residuals show that the larger values have larger standard deviations, this may mean that the random error ist rather multiplicative than additive (as it should be). In this case, a log transform may help. Other transformations are shown in the R-script. Sometimes, there is a good explantation why a certain transformation should be applied. Sometimes the Box-Cox-Transform can help, which can take various shapes, depending on a parameter to be optimized.

Nested ANOVA: What if the data are not really independent?

```
> oats.aov <- aov(Y~N*V+Error(B/V), data=oats)
> summary(oats.aov)

Error: B
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  5  15875  3175.1

Error: B:V
          Df Sum Sq Mean Sq F value Pr(>F)
V          2 1786.4  893.18  1.4853 0.2724
Residuals 10 6013.3  601.33

Error: Within
          Df  Sum Sq Mean Sq F value    Pr(>F)
N          3 20020.5  6673.5 37.6856 2.458e-12 ***
N:V        6   321.7    53.6  0.3028    0.9322
Residuals 45  7968.7   177.1
---
```

# Contents

Anova

Linear Models

Generalized Linear Models

Mixed-effects models

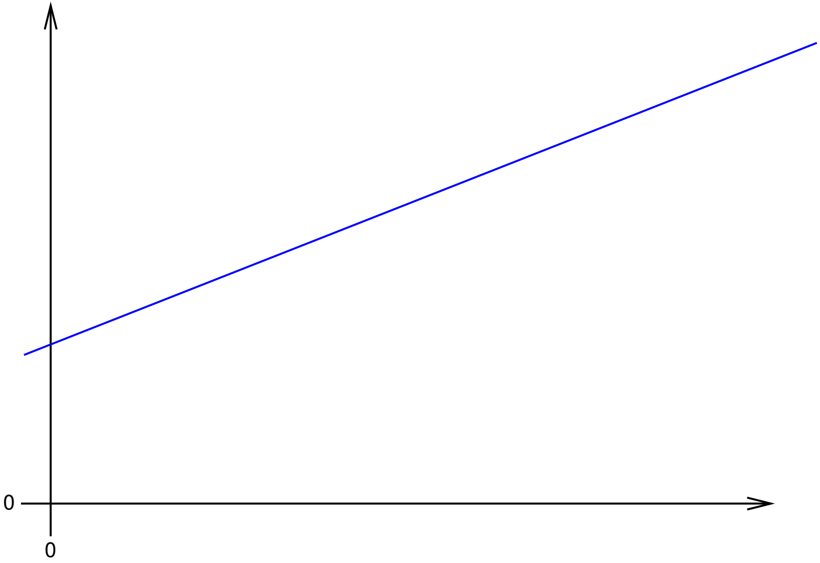
Principal Component Analysis (PCA)

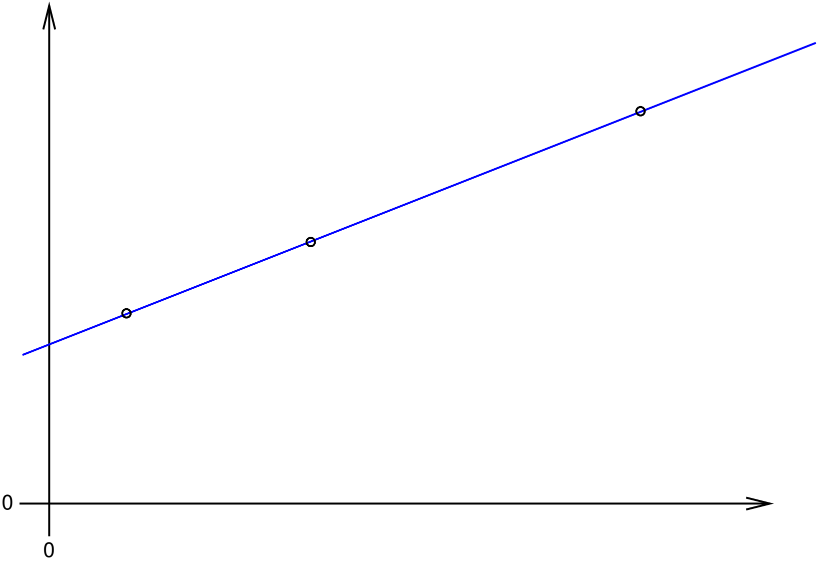
Redundancy Analysis (RDA)

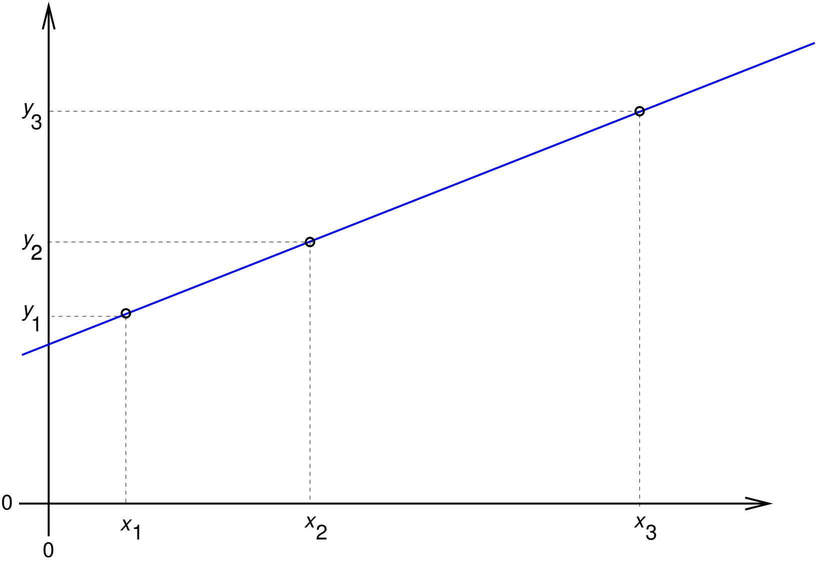
Correspondence Analysis (CA)

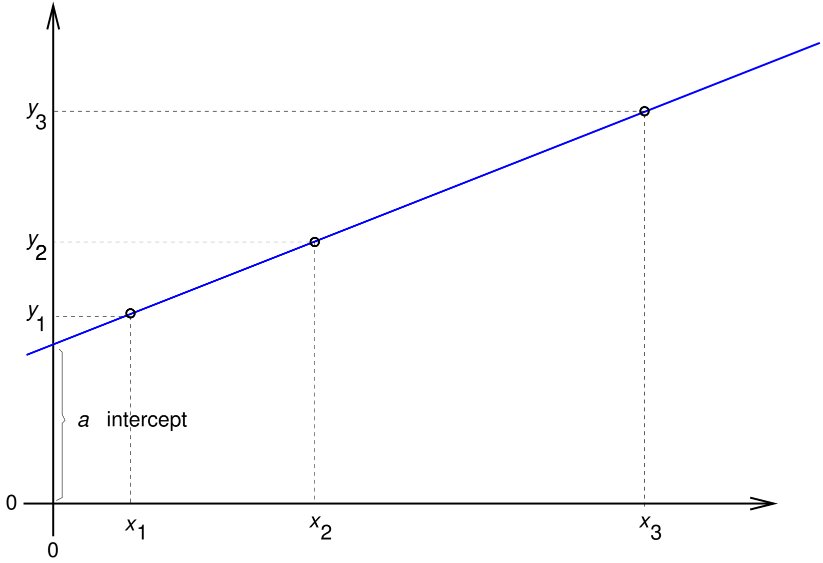
Canonical Correspondence Analysis (CCA)

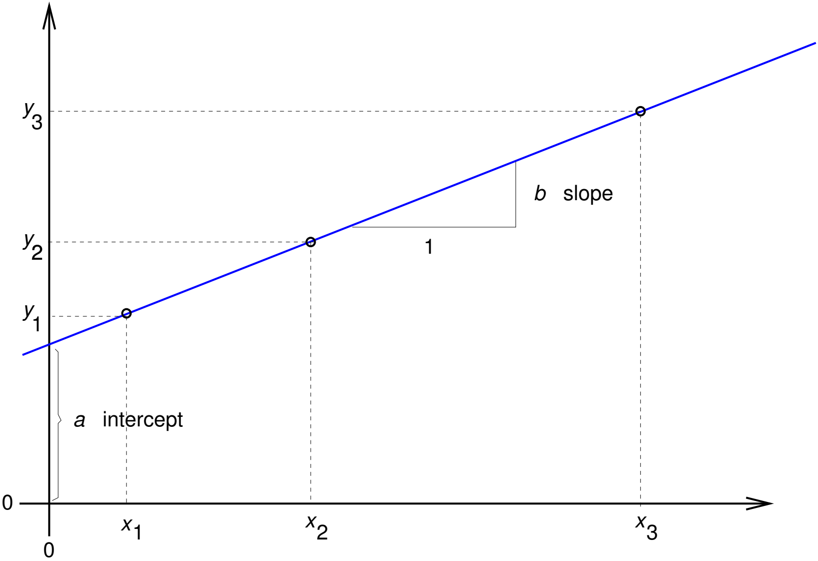
QTL Mapping

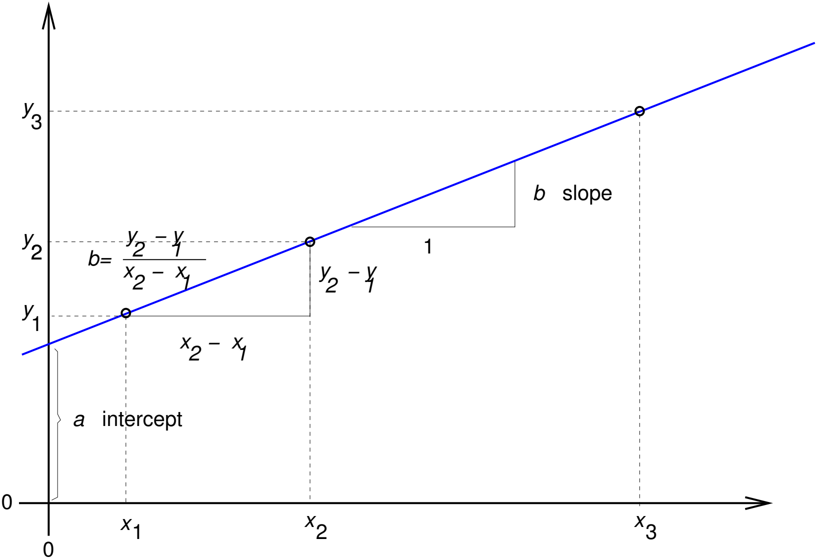

0
0

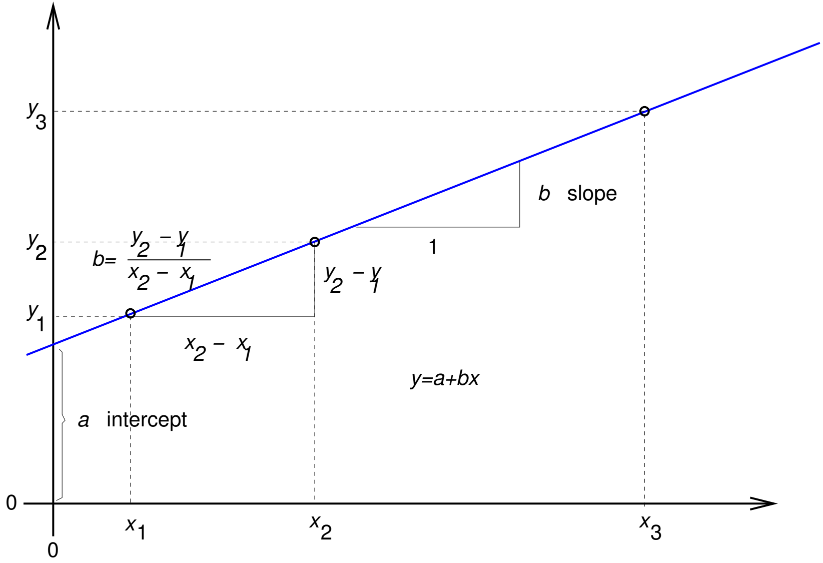

0
0


$y_3$
$y_2$
$y_1$
0
$x_1$
$x_2$
$x_3$
0


$y_3$
$y_2$
$y_1$
$a$ intercept
0
$x_1$
$x_2$
$x_3$
0

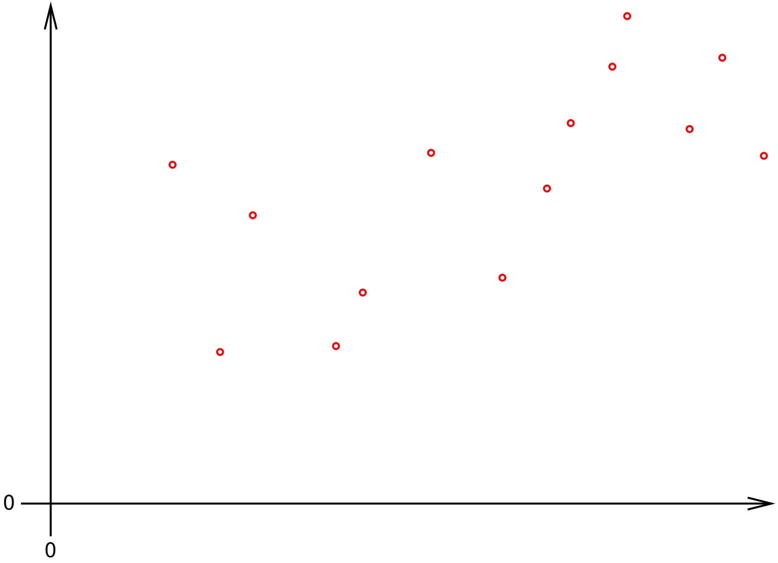

$y_3$
$b$ slope
1
$y_2$
$y_1$
$a$ intercept
0
$x_1$
$x_2$
$x_3$
0

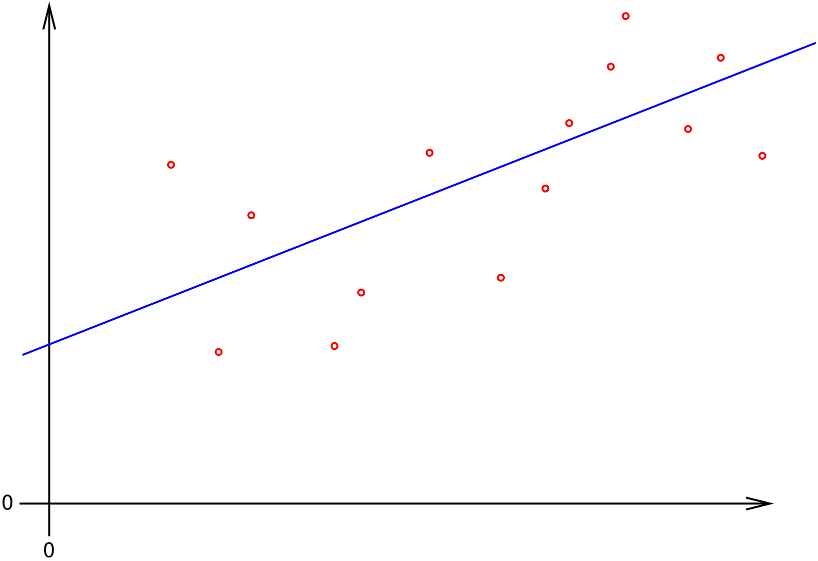

$y_3$
$y_2$
$y_1$
$b = \frac{y_2 - y_1}{x_2 - x_1}$
$b$ slope
1
$y_2 - y_1$
$x_2 - x_1$
$a$ intercept
0
0
$x_1$
$x_2$
$x_3$

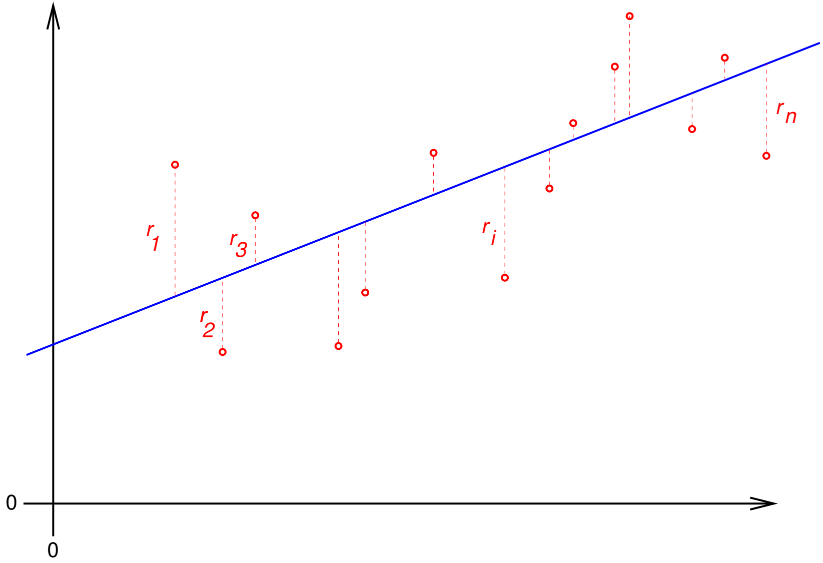


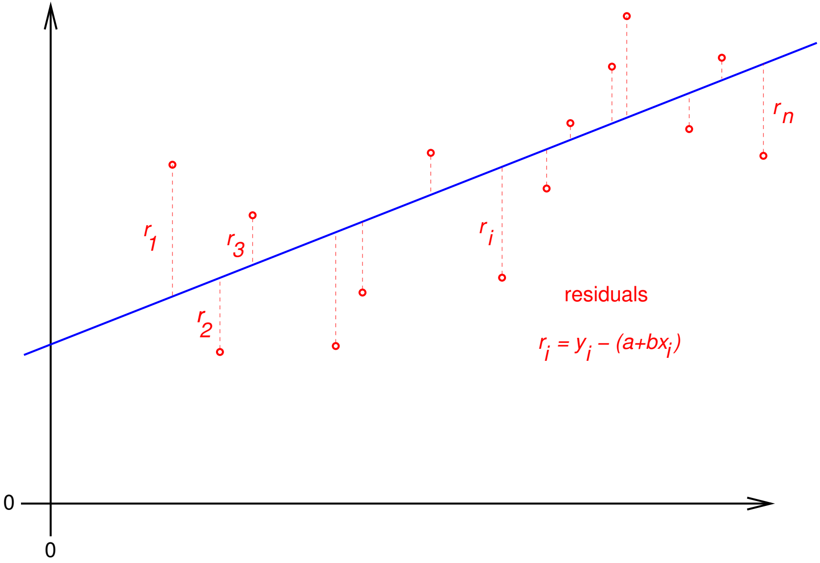
$y_3$
$y_2$
$y_1$
$b = \frac{y_2 - y_1}{x_2 - x_1}$
$b$ slope
1
$y_2 - y_1$
$x_2 - x_1$
$y = a + bx$
$a$ intercept
0
$x_1$
$x_2$
$x_3$
0

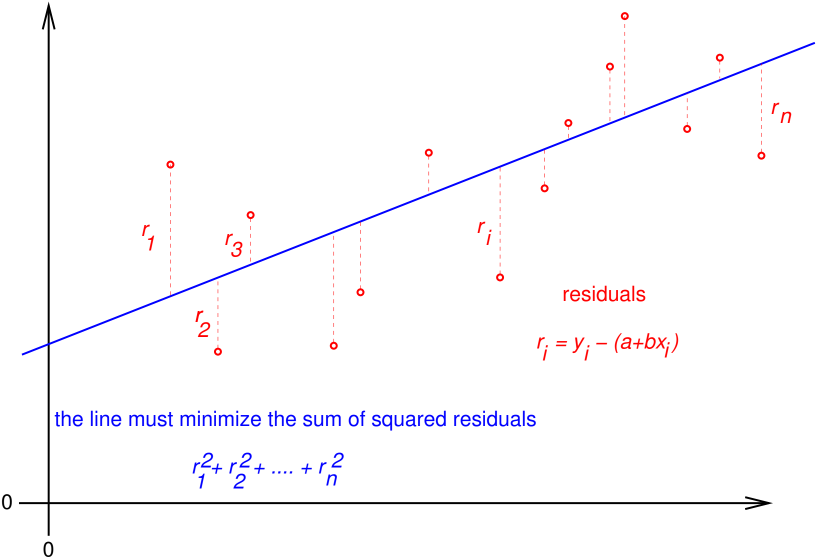
Linear Models




0
0


$r_1$
$r_2$
$r_3$
$r_i$
$r_n$
0
0


$r_1$
$r_2$
$r_3$
$r_i$
$r_n$
residuals
$r_i = y_i - (a+bx_i)$
0
0



residuals

$r_i = y_i - (a+bx_i)$

the line must minimize the sum of squared residuals

$r_1^2 + r_2^2 + .... + r_n^2$

define the regression line

$$y = \hat{a} + \hat{b} \cdot x$$

by minimizing the sum of squared residuals:

$$(\hat{a}, \hat{b}) = \arg\min_{(a,b)} \sum_i (y_i - (a + b \cdot x_i))^2$$

this is based on the model assumption that values $a, b$ exist, such that, for all data points $(x_i, y_i)$ we have

$$y_i = a + b \cdot x_i + \varepsilon_i,$$

whereas all $\varepsilon_i$ are independent and normally distributed with the same variance $\sigma^2$.

given data:

| **Y** | **X** |
|---|---|
| $y_1$ | $x_1$ |
| $y_2$ | $x_2$ |
| $y_3$ | $x_3$ |
| $\vdots$ | $\vdots$ |
| $y_n$ | $x_n$ |

given data:

| **Y** | **X** |
|---|---|
| $y_1$ | $x_1$ |
| $y_2$ | $x_2$ |
| $y_3$ | $x_3$ |
| $\vdots$ | $\vdots$ |
| $y_n$ | $x_n$ |

Model: there are values $a$, $b$, $\sigma^2$ such that

$$\begin{aligned}
y_1 &= a + b \cdot x_1 + \varepsilon_1 \\
y_2 &= a + b \cdot x_2 + \varepsilon_2 \\
y_3 &= a + b \cdot x_3 + \varepsilon_3 \\
\vdots & \quad \vdots \\
y_n &= a + b \cdot x_n + \varepsilon_n
\end{aligned}$$

given data:

| **Y** | **X** |
|---|---|
| $y_1$ | $x_1$ |
| $y_2$ | $x_2$ |
| $y_3$ | $x_3$ |
| $\vdots$ | $\vdots$ |
| $y_n$ | $x_n$ |

Model: there are values $a$, $b$, $\sigma^2$ such that

$$\begin{aligned} y_1 &= a + b \cdot x_1 + \varepsilon_1 \\ y_2 &= a + b \cdot x_2 + \varepsilon_2 \\ y_3 &= a + b \cdot x_3 + \varepsilon_3 \\ \vdots & \quad \vdots \\ y_n &= a + b \cdot x_n + \varepsilon_n \end{aligned}$$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent $\sim \mathcal{N}(0, \sigma^2)$.

given data:

| **Y** | **X** |
|---|---|
| $y_1$ | $x_1$ |
| $y_2$ | $x_2$ |
| $y_3$ | $x_3$ |
| $\vdots$ | $\vdots$ |
| $y_n$ | $x_n$ |

Model: there are values $a$, $b$, $\sigma^2$ such that

$$\begin{aligned} y_1 &= a + b \cdot x_1 + \varepsilon_1 \\ y_2 &= a + b \cdot x_2 + \varepsilon_2 \\ y_3 &= a + b \cdot x_3 + \varepsilon_3 \\ \vdots & \quad \vdots \\ y_n &= a + b \cdot x_n + \varepsilon_n \end{aligned}$$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent $\sim \mathcal{N}(0, \sigma^2)$.

$\Rightarrow y_1, y_2, \ldots, y_n$ are independent $y_i \sim \mathcal{N}(a + b \cdot x_i, \sigma^2)$.

given data:

| **Y** | **X** |
|---|---|
| $y_1$ | $x_1$ |
| $y_2$ | $x_2$ |
| $y_3$ | $x_3$ |
| $\vdots$ | $\vdots$ |
| $y_n$ | $x_n$ |

Model: there are values $a$, $b$, $\sigma^2$ such that

$$\begin{aligned} y_1 &= a + b \cdot x_1 + \varepsilon_1 \\ y_2 &= a + b \cdot x_2 + \varepsilon_2 \\ y_3 &= a + b \cdot x_3 + \varepsilon_3 \\ \vdots & \quad \vdots \\ y_n &= a + b \cdot x_n + \varepsilon_n \end{aligned}$$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent $\sim \mathcal{N}(0, \sigma^2)$.

$\Rightarrow y_1, y_2, \ldots, y_n$ are independent $y_i \sim \mathcal{N}(a + b \cdot x_i, \sigma^2)$.

$a, b, \sigma^2$ are unknown, but **not random**.

We estimate *a* and *b* by computing

$$(\hat{a}, \hat{b}) := \arg\min_{(a,b)} \sum_i (y_i - (a + b \cdot x_i))^2.$$

We estimate *a* and *b* by computing

$$(\hat{a}, \hat{b}) := \arg\min_{(a,b)} \sum_i (y_i - (a + b \cdot x_i))^2.$$

## Theorem

*Compute â and b̂ by*

$$\hat{b} = \frac{\sum_i (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i y_i \cdot (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

*and*

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}.$$

We estimate $a$ and $b$ by computing

$$(\hat{a}, \hat{b}) := \arg\min_{(a,b)} \sum_i (y_i - (a + b \cdot x_i))^2.$$

## Theorem

*Compute* $\hat{a}$ *and* $\hat{b}$ *by*

$$\hat{b} = \frac{\sum_i (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i y_i \cdot (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

*and*

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}.$$

**Please keep in mind:**
The line $y = \hat{a} + \hat{b} \cdot x$ goes through the center of gravity of the cloud of points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

```
> mod <- lm(ratiomales~rank,data=hind)
> summary(mod)
Call:
lm(formula = ratiomales ~ rank, data = hind)
Residuals:
     Min       1Q   Median       3Q      Max
-0.32798 -0.09396  0.02408  0.11275  0.37403

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20529    0.04011   5.119 4.54e-06 ***
rank         0.45877    0.06732   6.814 9.78e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.154 on 52 degrees of freedom
Multiple R-squared: 0.4717, Adjusted R-squared: 0.4616
F-statistic: 46.44 on 1 and 52 DF,  p-value: 9.78e-09
```

Model:

$$Y = a + b \cdot X + \varepsilon \qquad \text{mit } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Model:

$$Y = a + b \cdot X + \varepsilon \qquad \text{mit } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

How to compute the significance of a relationship between the *explanatory trait* $X$ and the *target variable* $Y$?

Model:

$$Y = a + b \cdot X + \varepsilon \qquad \text{mit } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

How to compute the significance of a relationship between the *explanatory trait* $X$ and the *target variable* $Y$?

In other words: How can we test the null hypothesis $b = 0$?

Model:

$$Y = a + b \cdot X + \varepsilon \qquad \text{mit } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

How to compute the significance of a relationship between the *explanatory trait* $X$ and the *target variable* $Y$?

In other words: How can we test the null hypothesis $b = 0$?

We have estimated $b$ by $\hat{b} \neq 0$. Could the true $b$ be 0?

Model:

$$Y = a + b \cdot X + \varepsilon \qquad \text{mit } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

How to compute the significance of a relationship between the *explanatory trait* $X$ and the *target variable* $Y$?

In other words: How can we test the null hypothesis $b = 0$?

We have estimated $b$ by $\hat{b} \neq 0$. Could the true $b$ be 0?

How large is the standard error of $\hat{b}$?

# t-test for $\hat{b}$

Estimate $\sigma^2$ by

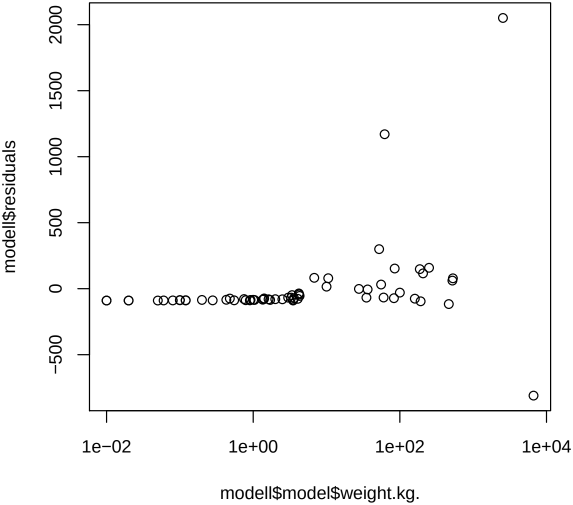
$$s^2 = \frac{\sum_i \left(y_i - \hat{a} - \hat{b} \cdot x_i\right)^2}{n-2}.$$

Then,

$$\frac{\hat{b} - b}{s \Big/ \sqrt{\sum_i (x_i - \bar{x})^2}}$$

is t-distributed with $n-2$ degrees of freedom. Thus, we can apply a t-test to test the null-hypothesis $b = 0$.

```
> modell <- lm(brain.weight.g~weight.kg.,subset=extinct=="no"
> summary(modell)
Call:
lm(formula = brain.weight.g ~ weight.kg., subset = extinct ==
    "no")
Residuals:
    Min      1Q  Median      3Q     Max
-809.95  -87.43  -78.55  -31.17 2051.05
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 89.91213   43.58134   2.063   0.0434 *
weight.kg.   0.96664    0.04769  20.269   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 334.8 on 60 degrees of freedom
Multiple R-squared: 0.8726, Adjusted R-squared: 0.8704
F-statistic: 410.8 on 1 and 60 DF,  p-value: < 2.2e-16
```


modell$residuals
2000
1500
1000
500
0
−500
1e−02
1e+00
1e+02
1e+04
modell$model$weight.kg.

We see that the residuals’ varaince depends on the fitted values (or the body weight): “heteroscadiscity”

We see that the residuals' varaince depends on the fitted values (or the body weight): "heteroscadiscity"
The model assumes *homoscedascity*, i.e. the random deviations must be (almost) independent of the explaining traits (body weight) and the fitted values.

We see that the residuals’ varaince depends on the fitted values (or the body weight): “heteroscadiscity”

The model assumes *homoscedascity*, i.e. the random deviations must be (almost) independent of the explaining traits (body weight) and the fitted values.

**variance-stabilizing transformation:**

can be rescale body- and brain size to make deviations independent of variables

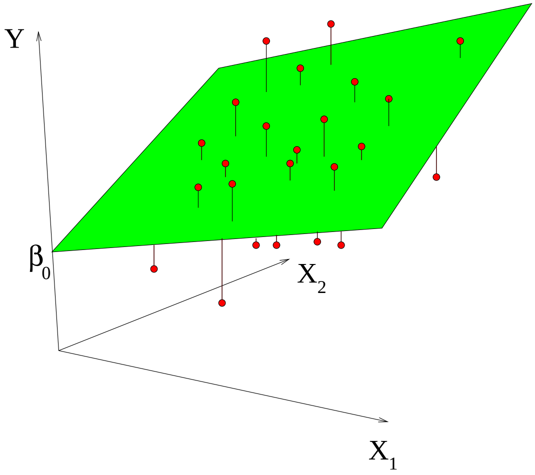
Actually not so surprising: An elephant's brain of typically 5 kg can easily be 500 g lighter or heavier from individual to individual. This can not happen for a mouse brain of typically 5 g. The latter will rather also vary by 10%, i.e. 0.5 g. Thus, the variance is not additive but rather multiplicative:

$$\text{brain mass} = (\text{expected brain mass}) \cdot \text{random}$$

We can convert this into something with additive randomness by taking the log:

$$\log(\text{brain mass}) = \log(\text{expected brain mass}) + \log(\text{random})$$

# Multivariate Regression



# Multivariate Regression

Problem: Predict $Y$ from $X_1$, $X_2$,. . . ,$X_m$.

# Multivariate Regression

Problem: Predict $Y$ from $X_1$, $X_2$,$\ldots$,$X_m$.

Observations:

$$\begin{array}{ccc}
Y_1 & , & X_{11}, X_{21}, \ldots, X_{m1} \\
Y_2 & , & X_{12}, X_{22}, \ldots, X_{m2} \\
\vdots & & \vdots \\
Y_n & , & X_{1n}, X_{2n}, \ldots, X_{mn}
\end{array}$$

# Multivariate Regression

Problem: Predict $Y$ from $X_1$, $X_2$,...,$X_m$.

Observations:

$$\begin{array}{ccc}
Y_1 & , & X_{11}, X_{21}, \dots, X_{m1} \\
Y_2 & , & X_{12}, X_{22}, \dots, X_{m2} \\
\vdots & & \vdots \\
Y_n & , & X_{1n}, X_{2n}, \dots, X_{mn}
\end{array}$$

Model: $Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_m \cdot X_m + \varepsilon$

# Multivariate Regression

Problem: Predict $Y$ from $X_1$, $X_2$,...,$X_m$.

Observations:

$$\begin{array}{ccc}
Y_1 & , & X_{11}, X_{21}, \ldots, X_{m1} \\
Y_2 & , & X_{12}, X_{22}, \ldots, X_{m2} \\
\vdots & & \vdots \\
Y_n & , & X_{1n}, X_{2n}, \ldots, X_{mn}
\end{array}$$

Model: $Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_m \cdot X_m + \varepsilon$

Equation system to determine $a$, $b_1$, $b_2$, ..., $b_m$:

$$\begin{array}{ccccccccccccc}
Y_1 & = & a & + & b_1 \cdot X_{11} & + & b_2 \cdot X_{21} & + & \ldots & + & b_m \cdot X_{m1} & + & \varepsilon_1 \\
Y_2 & = & a & + & b_1 \cdot X_{12} & + & b_2 \cdot X_{22} & + & \ldots & + & b_m \cdot X_{m2} & + & \varepsilon_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
Y_n & = & a & + & b_1 \cdot X_{1n} & + & b_n \cdot X_{2n} & + & \ldots & + & b_m \cdot X_{mn} & + & \varepsilon_n
\end{array}$$

Model:

$$\begin{array}{ccccccccccccc}
Y_1 & = & a & + & b_1 \cdot X_{11} & + & b_2 \cdot X_{21} & + & \ldots & + & b_m \cdot X_{m1} & + & \varepsilon_1 \\
Y_2 & = & a & + & b_1 \cdot X_{12} & + & b_2 \cdot X_{22} & + & \ldots & + & b_m \cdot X_{m2} & + & \varepsilon_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
Y_n & = & a & + & b_1 \cdot X_{1n} & + & b_n \cdot X_{2n} & + & \ldots & + & b_m \cdot X_{mn} & + & \varepsilon_n
\end{array}$$

target variable $Y$
explanatory variables $X_1, X_2, \ldots, X_m$
parameter to be estimated $a, b_1, \ldots, b_m$
independent normally distributed pertubations $\varepsilon_1, \ldots, \varepsilon_m$ with unknown variance $\sigma^2$.

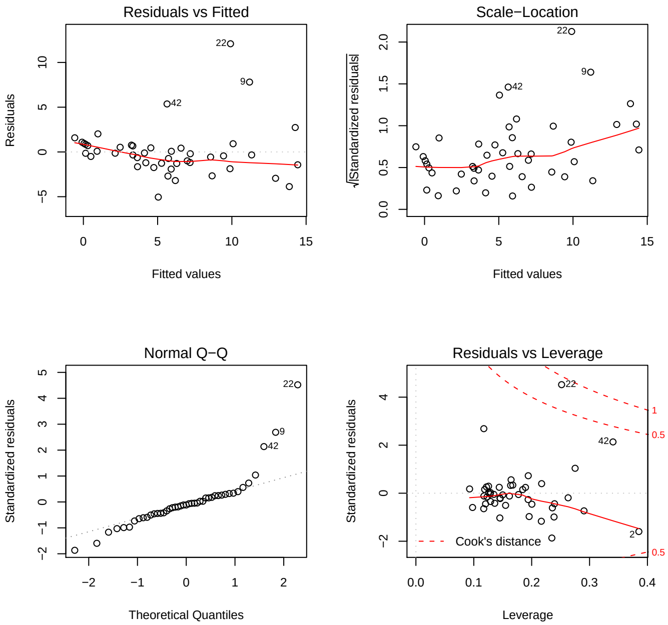
```
> modell0 <- lm(richness ~ angle2+NAP+grainsize+humus,
+                   data = rikz)
> modell <- lm(richness ~ angle2+NAP+grainsize+humus
+                            +factor(week), data = rikz)
> anova(modell0, modell)
Analysis of Variance Table

Model 1: richness ~ angle2 + NAP + grainsize + humus
Model 2: richness ~ angle2 + NAP + grainsize + humus + factor
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     40 531.17
2     37 353.66  3    177.51 6.1902 0.00162 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We reject the null hypothesis that the weeks have no effect with a $p$-value of 0.00162.

We reject the null hypothesis that the weeks have no effect with a *p*-value of 0.00162.

But wait! We can only do that if the more complex model fits well to the data. We check this graphically.

```
plot(modell)
```

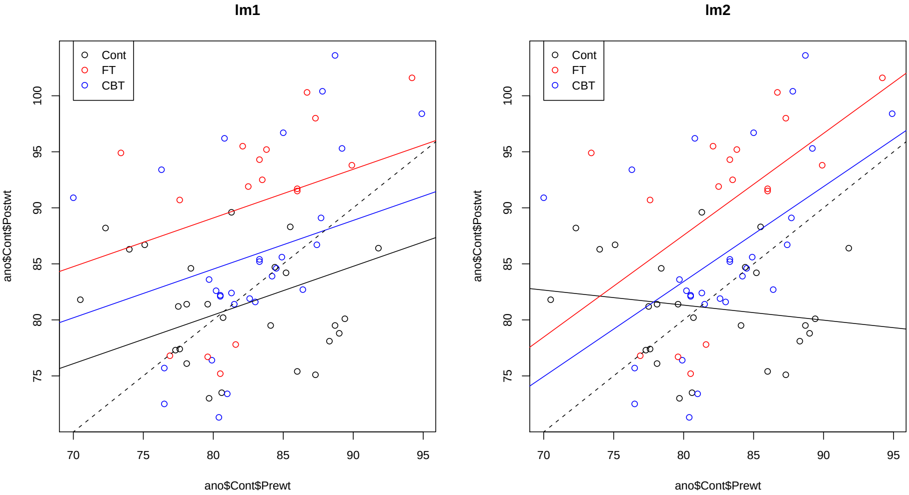

Residuals vs Fitted
Residuals
Fitted values
10
5
0
−5
0
5
10
15
22
9
42
Scale−Location
√|Standardized residuals|
Fitted values
2.0
1.5
1.0
0.5
0.0
0
5
10
15
22
9
42
Normal Q−Q
Standardized residuals
Theoretical Quantiles
5
4
3
2
1
0
−1
−2
−2
−1
0
1
2
22
9
42
Residuals vs Leverage
Standardized residuals
Leverage
4
2
0
−2
0.0
0.1
0.2
0.3
0.4
22
42
2
1
0.5
0.5
Cook's distance

# Different types of ANOVA tables

If you apply the R command `anova` to a single model, the variables are added consecutively in the same order as in the command. Each *p* value refers to the test wether the model gets significantly better by adding the variable to only those that are listed above the variable. In contrast to this, the *p* values that are given by `summary` or by `dropterm` from the MASS library always compare the model to a model where only the corresponding variable is set to 0 and all other variables can take any values. The *p* values given by `anova` thus depend on the order in which the variables are given in the command. This is not the case for `summary` and `dropterm`. The same options exist in other software packages, sometimes under the names “type I analysis” and “type II analysis”.

```
> lm1 <- lm(Postwt~Prewt+Treat,anorexia)
> lm2 <- lm(Postwt~Prewt*Treat,anorexia)
> anova(lm1,lm2)
Analysis of Variance Table

Model 1: Postwt ~ Prewt + Treat
Model 2: Postwt ~ Prewt * Treat
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     68 3311.3
2     66 2844.8  2     466.5 5.4112 0.006666 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

# Linear Models



# How to predict the winglength of a Darwin finch by its beak size?

How to predict the winglength of a Darwin finch by its beak size?
Shall we take beak height, beak length or both into account?

How to predict the winglength of a Darwin finch by its beak size?
Shall we take beak height, beak length or both into account?
Residual variance should be small....

How to predict the winglength of a Darwin finch by its beak size? Shall we take beak height, beak length or both into account? Residual variance should be small....

**Leave-one-out cross validation:** If you leave out one bird and fit the model to the others, how well can this model predict the wing span?

How to predict the winglength of a Darwin finch by its beak size? Shall we take beak height, beak length or both into account? Residual variance should be small....

**Leave-one-out cross validation:** If you leave out one bird and fit the model to the others, how well can this model predict the wing span?

```
prederrorHL <- numeric()
for (i in 1:46) {
  selection <- rep(TRUE,46)
  selection[i] <- FALSE
  modHL.R <- lm(WingL~N.UBkL+BeakH,data=finchdata,
                                   subset=selection)
  prederrorHL[i]=WingL[i]-predict(modHL.R,finchdata[i,])
}
```

| | Height | Length | Height and Length |
|---|---|---|---|
| $\sigma$(Residuals) | 3.83 | 4.78 | 3.79 |

| | Height | Length | Height and Length |
|---|---|---|---|
| $\sigma$(Residuals) | 3.83 | 4.78 | 3.79 |
| $d =$ (Number Parameters) | 2 | 2 | 3 |

| | Height | Length | Height and Length |
|---|---|---|---|
| $\sigma$(Residuals) | 3.83 | 4.78 | 3.79 |
| $d =$ (Number Parameters) | 2 | 2 | 3 |
| $\sigma(\text{Residuals})\cdot\sqrt{\frac{n-1}{n-d}}$ | 3.86 | 4.84 | 3.87 |

| | Height | Length | Height and Length |
|---|---|---|---|
| $\sigma$(Residuals) | 3.83 | 4.78 | 3.79 |
| $d =$ (Number Parameters) | 2 | 2 | 3 |
| $\sigma(\text{Residuals})\cdot\sqrt{\frac{n-1}{n-d}}$ | 3.86 | 4.84 | 3.87 |
| cross validation. | 3.96 | 4.97 | 3.977 |

| | Height | Length | Height and Length |
|---|---|---|---|
| $\sigma$(Residuals) | 3.83 | 4.78 | 3.79 |
| $d =$ (Number Parameters) | 2 | 2 | 3 |
| $\sigma$(Residuals)$\cdot\sqrt{\frac{n-1}{n-d}}$ | 3.86 | 4.84 | 3.87 |
| cross validation. | 3.96 | 4.97 | 3.977 |
| AIC | 259.0 | 279.5 | 260.1 |

| | Height | Length | Height and Length |
|---|---|---|---|
| $\sigma$(Residuals) | 3.83 | 4.78 | 3.79 |
| $d =$ (Number Parameters) | 2 | 2 | 3 |
| $\sigma$(Residuals)$\cdot\sqrt{\frac{n-1}{n-d}}$ | 3.86 | 4.84 | 3.87 |
| cross validation. | 3.96 | 4.97 | 3.977 |
| AIC | 259.0 | 279.5 | 260.1 |
| BIC | 264.4 | 285.0 | 267.4 |

| | Height | Length | Height and Length |
|---|---|---|---|
| $\sigma$(Residuals) | 3.83 | 4.78 | 3.79 |
| $d =$ (Number Parameters) | 2 | 2 | 3 |
| $\sigma$(Residuals)$\cdot\sqrt{\frac{n-1}{n-d}}$ | 3.86 | 4.84 | 3.87 |
| cross validation. | 3.96 | 4.97 | 3.977 |
| AIC | 259.0 | 279.5 | 260.1 |
| BIC | 264.4 | 285.0 | 267.4 |

Akaike's Information Criterion:

$$\mathrm{AIC} = -2 \cdot \log L + 2 \cdot (\mathrm{NumberofParameters})$$

| | Height | Length | Height and Length |
|---|---|---|---|
| $\sigma$(Residuals) | 3.83 | 4.78 | 3.79 |
| $d =$ (Number Parameters) | 2 | 2 | 3 |
| $\sigma(\text{Residuals})\cdot\sqrt{\frac{n-1}{n-d}}$ | 3.86 | 4.84 | 3.87 |
| cross validation. | 3.96 | 4.97 | 3.977 |
| AIC | 259.0 | 279.5 | 260.1 |
| BIC | 264.4 | 285.0 | 267.4 |

Akaike's Information Criterion:

$$\mathrm{AIC} = -2 \cdot \log L + 2 \cdot (\mathrm{NumberofParameters})$$

Bayesian Information Criterion:

$$\mathrm{BIC} = -2 \cdot \log L + \log(n) \cdot (\mathrm{NumberofParameters})$$

| | Height | Length | Height and Length |
|---|---|---|---|
| $\sigma$(Residuals) | 3.83 | 4.78 | 3.79 |
| $d =$ (Number Parameters) | 2 | 2 | 3 |
| $\sigma(\text{Residuals})\cdot\sqrt{\frac{n-1}{n-d}}$ | 3.86 | 4.84 | 3.87 |
| cross validation. | 3.96 | 4.97 | 3.977 |
| AIC | 259.0 | 279.5 | 260.1 |
| BIC | 264.4 | 285.0 | 267.4 |

Akaike's Information Criterion:

$$\mathrm{AIC} = -2 \cdot \log L + 2 \cdot (\mathrm{NumberofParameters})$$

Bayesian Information Criterion:

$$\mathrm{BIC} = -2 \cdot \log L + \log(n) \cdot (\mathrm{NumberofParameters})$$

# Contents

Anova

Linear Models

Generalized Linear Models

Mixed-effects models

Principal Component Analysis (PCA)

Redundancy Analysis (RDA)

Correspondence Analysis (CA)

Canonical Correspondence Analysis (CCA)

QTL Mapping

The normal distribution $\mathcal{N}(\mu, \sigma^2)$ is a continuous distribution and thus not suitable to model distributions on small numbers.

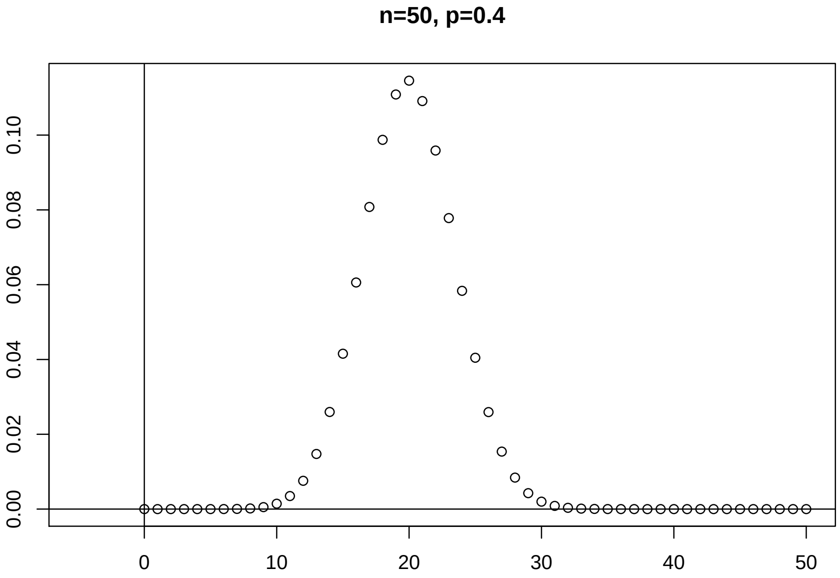
The normal distribution $\mathcal{N}(\mu, \sigma^2)$ is a continuous distribution and thus not suitable to model distributions on small numbers.

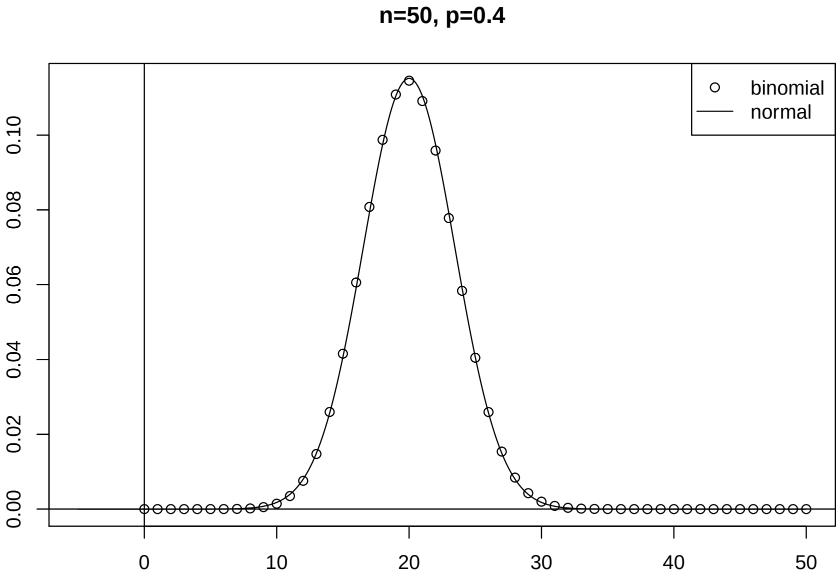
The Poisson distribution Pois($\lambda$) is a distribution on $\{0, 1, 2, 3, \dots\}$.

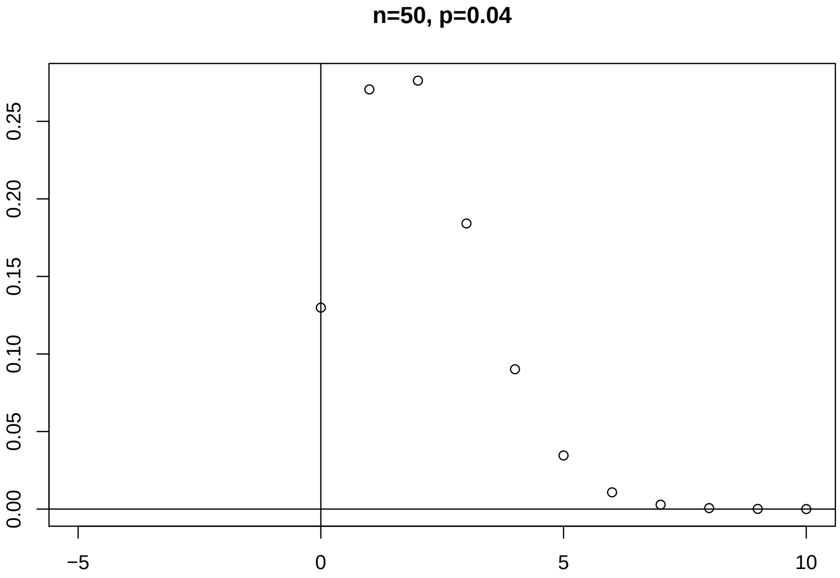
The normal distribution $\mathcal{N}(\mu, \sigma^2)$ is a continuous distribution and thus not suitable to model distributions on small numbers.

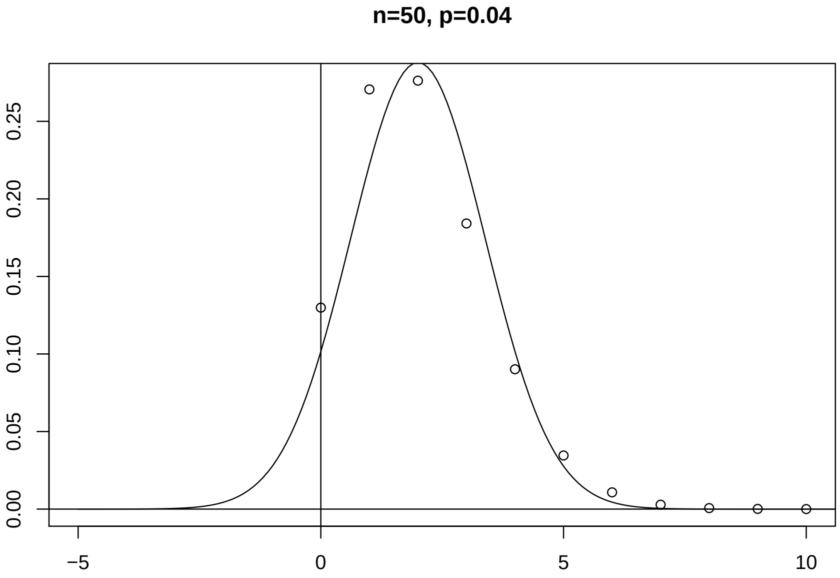
The Poisson distribution Pois($\lambda$) is a distribution on $\{0, 1, 2, 3, \dots\}$.

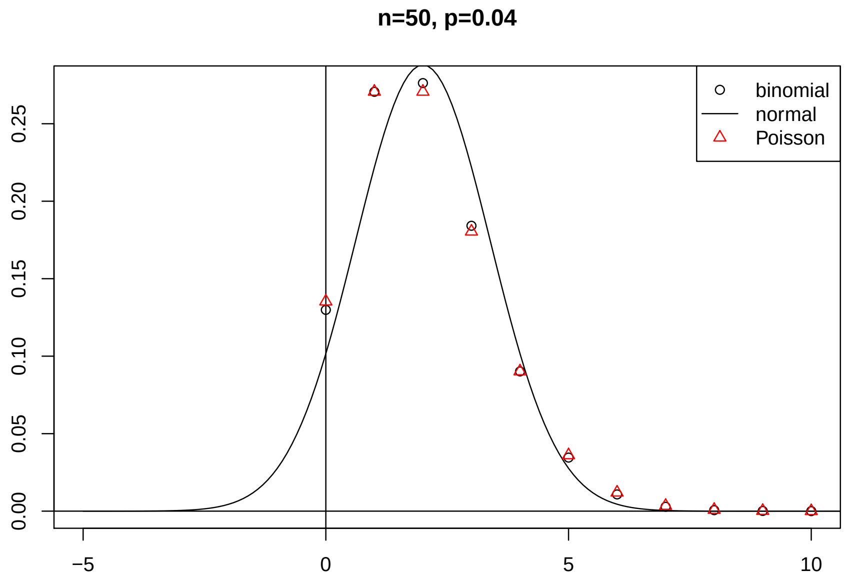
$\mathcal{N}(\mu = n \cdot p, \sigma^2 = n \cdot p \cdot (1 - p))$ approximates the binomial distribution Bin($n$,$p$) if $n \cdot p \cdot (1 - p)$ is not too small (rule of thumb: $n \cdot p \cdot (1 - p) > 9$), Pois($\lambda = n \cdot p$) gives a better approximation when $p$ is small.


n=50, p=0.4
0.00
0.02
0.04
0.06
0.08
0.10
0
10
20
30
40
50


n=50, p=0.4
binomial
normal
0.00
0.02
0.04
0.06
0.08
0.10
0
10
20
30
40
50


n=50, p=0.04
0.00
0.05
0.10
0.15
0.20
0.25
−5
0
5
10


n=50, p=0.04
0.00
0.05
0.10
0.15
0.20
0.25
−5
0
5
10


n=50, p=0.04
○ binomial
— normal
△ Poisson
0.00
0.05
0.10
0.15
0.20
0.25
−5
0
5
10

If $Y$ is Pois($\lambda$)-distributed, then

$$\begin{aligned}
\Pr(Y = k) &= \frac{\lambda^k}{k!} \cdot e^{-\lambda} \qquad \text{for } k = 0, 1, 2, \ldots \\
\mathbb{E}Y &= \lambda \\
\mathrm{Var}(Y) &= \lambda
\end{aligned}$$

If $Y$ is Pois($\lambda$)-distributed, then

$$\begin{aligned}\Pr(Y = k) &= \frac{\lambda^k}{k!} \cdot e^{-\lambda} \qquad \text{for } k = 0, 1, 2, \ldots \\ \mathbb{E}Y &= \lambda \\ \mathrm{Var}(Y) &= \lambda\end{aligned}$$

Is there a linear model with Pois($\lambda$) instead of $\mathcal{N}(\mu, \sigma^2)$?

If $Y$ is Pois($\lambda$)-distributed, then

$$\begin{aligned} \Pr(Y=k) &= \frac{\lambda^k}{k!}\cdot e^{-\lambda} \qquad \text{for } k=0,1,2,\dots \\ \mathbb{E}Y &= \lambda \\ \mathrm{Var}(Y) &= \lambda \end{aligned}$$

Is there a linear model with Pois($\lambda$) instead of $\mathcal{N}(\mu,\sigma^2)$?

Yes, the **Generalized Linear Model (GLM) of type Poisson**.

Remeber the normal linear model:

$$Y_i = b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i} + \varepsilon_i \qquad \text{with } \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Remeber the normal linear model:

$$Y_i = b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i} + \varepsilon_i \qquad \text{with } \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

or equivalently:

$$\begin{aligned} \eta_i &= b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i} \\ Y_i &\sim \mathcal{N}(\eta_i, \sigma^2) \end{aligned}$$

$\eta$ is called the *linear predictor*.

Remeber the normal linear model:

$$Y_i = b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i} + \varepsilon_i \qquad \text{with } \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

or equivalently:

$$\begin{aligned} \eta_i &= b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i} \\ Y_i &\sim \mathcal{N}(\eta_i, \sigma^2) \end{aligned}$$

$\eta$ is called the *linear predictor*.

This also works for the Poisson distribution:

$$\begin{aligned} \eta_i &= b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i} \\ Y_i &\sim \mathrm{Pois}(\eta_i) \end{aligned}$$

(but note that the additional $\sigma^2$ is missing!)

Instead of using $\eta$ directly as parameter of the Poisson distribution, it is common to apply a transformation:

$$\begin{aligned} \ell(\mu_i) = \eta_i &= b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i} \\ Y_i &\sim \mathrm{Pois}(\mu_i) \end{aligned}$$

$\ell(.)$ is called the *link function*.

Instead of using $\eta$ directly as parameter of the Poisson distribution, it is common to apply a transformation:

$$\begin{aligned} \ell(\mu_i) \; = \; \eta_i \; &= \; b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i} \\ Y_i \; &\sim \; \mathrm{Pois}(\mu_i) \end{aligned}$$

$\ell(.)$ is called the *link function*.
The default link function for Poisson GLMs is log, the natural logarithm.

Instead of using $\eta$ directly as parameter of the Poisson distribution, it is common to apply a transformation:

$$\begin{aligned} \ell(\mu_i) \;=\; \eta_i \;&=\; b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i} \\ Y_i \;&\sim\; \mathrm{Pois}(\mu_i) \end{aligned}$$

$\ell(.)$ is called the *link function*.
The default link function for Poisson GLMs is log, the natural logarithm.
Thus,

$$\mathbb{E}\,Y_i = \mu_i = e^{\eta_i} = e^{b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i}} = e^{b_0} \cdot e^{b_1 \cdot X_{1,i}} \cdots e^{b_k \cdot X_{k,i}}$$

and the Poisson GLM with this default link is multiplicative model rather than an additive one.

```
> pmod1 <- glm(counts~foodlevel+species,data=daph,
                                        family=poisson)
> summary(pmod1)
[...]
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.1166     0.1105  28.215  < 2e-16 ***
foodlevellow  -1.1567     0.1298  -8.910  < 2e-16 ***
speciesmagna   0.9794     0.1243   7.878 3.32e-15 ***
[...]
```

Note that the Poisson model has log as its default link function. Thus, the model pmod1 assumes that the number of Daphnia in row $i$ is Poisson distributed with mean $\lambda_i$, i.e. $\Pr(X = k) = \frac{\lambda_i^k}{k!} e^{-\lambda}$, and

$$\log\left(\lambda_i\right) \approx 3.12 - 1.15 \cdot I_{\text{lowfoodlevel}} + 0.979 \cdot I_{\text{magna}}$$

Note that the Poisson model has log as its default link function. Thus, the model pmod1 assumes that the number of Daphnia in row $i$ is Poisson distributed with mean $\lambda_i$, i.e. $\Pr(X = k) = \frac{\lambda_i^k}{k!} e^{-\lambda}$, and

$$\log(\lambda_i) \approx 3.12 - 1.15 \cdot I_{\mathrm{lowfoodlevel}} + 0.979 \cdot I_{\mathrm{magna}}$$

or, equivalently,

$$\lambda_i \approx e^{3.12} \cdot {e^{-1.15}}^{I_{\mathrm{lowfoodlevel}}} \cdot {e^{0.979}}^{I_{\mathrm{magna}}} \approx 22.6 \cdot 0.317^{I_{\mathrm{lowfoodlevel}}} \cdot 2.66^{I_{\mathrm{magna}}}$$

Thus, this Poisson model assumes multiplicative effects.

```
> pmod1 <- glm(counts~foodlevel+species,
                              data=daph,family=poisson)
> pmod2 <- glm(counts~foodlevel*species,
                              data=daph,family=poisson)
> anova(pmod1,pmod2,test="F")
> pmod1 <- glm(counts~foodlevel+species,
                              data=daph,family=poisson)
> pmod2 <- glm(counts~foodlevel*species,
                              data=daph,family=poisson)
> anova(pmod1,pmod2,test="F")

Analysis of Deviance Table

Model 1: counts ~ foodlevel + species
Model 2: counts ~ foodlevel * species
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1         9     6.1162
2         8     6.0741  1 0.042071 0.0421 0.8375
Warning message:
F-Test not appropriate for family ’poisson’
```

Note:

- The anova command gives us an “analysis of deviance” instead of an analysis of variance!

Note:

- The anova command gives us an “analysis of deviance” instead of an analysis of variance!
- What is a deviance?

Note:

- The anova command gives us an “analysis of deviance” instead of an analysis of variance!
- What is a deviance?
- There is a Warning “F-Test not appropriate for family ’poisson’ ”.

Note:

- The anova command gives us an “analysis of deviance” instead of an analysis of variance!
- What is a deviance?
- There is a Warning “F-Test not appropriate for family ’poisson’ ”.
- Why?

Note:

- The anova command gives us an “analysis of deviance” instead of an analysis of variance!
- What is a deviance?
- There is a Warning “F-Test not appropriate for family ’poisson’ ”.
- Why?
- Which test should we apply?

# What is the deviance?

Let $\widehat{b_0}, \ldots, \widehat{b_k}$ be our fitted model coefficients and

$$\widehat{\mu}_i = \ell^{-1}\left(\widehat{b_0} + \widehat{b_1} X_{1i} + \cdots + \widehat{b_k} X_{ki}\right)$$

be the predicted means for all observations. The Likelihood of the fitted parameter values is the probability of the observations assuming the fitted parameter values:

$$L(\widehat{\mu}) = \frac{\widehat{\mu_1}^{Y_1}}{Y_1!} e^{-\widehat{\mu_1}} \cdot \frac{\widehat{\mu_2}^{Y_2}}{Y_2!} e^{-\widehat{\mu_2}} \cdots \frac{\widehat{\mu_k}^{Y_k}}{Y_k!} e^{-\widehat{\mu_k}}$$

Now we compare this to a *saturated* Poisson GLM model, i.e. a model with so many parameters such that we can get a perfect fit of $\widetilde{\mu}_i = Y_i$. This leads to the highest possible likelihood $L(\widetilde{\mu})$.

# What is the deviance?

Let $\widehat{b}_0, \ldots, \widehat{b}_k$ be our fitted model coefficients and

$$\widehat{\mu}_i = \ell^{-1}\left(\widehat{b}_0 + \widehat{b}_1 X_{1i} + \cdots + \widehat{b}_k X_{ki}\right)$$

be the predicted means for all observations. The Likelihood of the fitted parameter values is the probability of the observations assuming the fitted parameter values:

$$L(\widehat{\mu}) = \frac{\widehat{\mu_1}^{Y_1}}{Y_1!} e^{-\widehat{\mu_1}} \cdot \frac{\widehat{\mu_2}^{Y_2}}{Y_2!} e^{-\widehat{\mu_2}} \cdots \frac{\widehat{\mu_k}^{Y_k}}{Y_k!} e^{-\widehat{\mu_k}}$$

Now we compare this to a *saturated* Poisson GLM model, i.e. a model with so many parameters such that we can get a perfect fit of $\widetilde{\mu}_i = Y_i$. This leads to the highest possible likelihood $L(\widetilde{\mu})$.In practice such a model is not desirable because it leads to overfitting.

# What is the deviance?

$$\begin{aligned}
\text{our model: } L(\widehat{\mu}) &= \frac{\widehat{\mu_1}^{Y_1}}{Y_1!}e^{-\widehat{\mu_1}} \cdot \frac{\widehat{\mu_2}^{Y_2}}{Y_2!}e^{-\widehat{\mu_2}} \cdots \frac{\widehat{\mu_k}^{Y_k}}{Y_k!}e^{-\widehat{\mu_k}} \\
\text{saturated model: } L(\widetilde{\mu}) &= \frac{Y_1^{Y_1}}{Y_1!}e^{-Y_1} \cdot \frac{Y_2^{Y_2}}{Y_2!}e^{-Y_2} \cdots \frac{Y_k{}^{Y_k}}{Y_k!}e^{-Y_k}
\end{aligned}$$

# What is the deviance?

$$\begin{aligned}\text{our model: } L(\widehat{\mu}) &= \frac{\widehat{\mu_1}^{Y_1}}{Y_1!}e^{-\widehat{\mu_1}} \cdot \frac{\widehat{\mu_2}^{Y_2}}{Y_2!}e^{-\widehat{\mu_2}} \ldots \frac{\widehat{\mu_k}^{Y_k}}{Y_k!}e^{-\widehat{\mu_k}} \\ \text{saturated model: } L(\widetilde{\mu}) &= \frac{Y_1^{Y_1}}{Y_1!}e^{-Y_1} \cdot \frac{Y_2^{Y_2}}{Y_2!}e^{-Y_2} \ldots \frac{Y_k^{Y_k}}{Y_k!}e^{-Y_k}\end{aligned}$$

The *residual deviance* of our model is defined as

$$2 \cdot \left[\log\left(L(\widehat{\mu})\right) - \log\left(L(\widetilde{\mu})\right)\right].$$

# What is the deviance?

$$
\begin{aligned}
\text{our model: } L(\widehat{\mu}) &= \frac{\widehat{\mu_1}^{Y_1}}{Y_1!}e^{-\widehat{\mu_1}} \cdot \frac{\widehat{\mu_2}^{Y_2}}{Y_2!}e^{-\widehat{\mu_2}} \cdots \frac{\widehat{\mu_k}^{Y_k}}{Y_k!}e^{-\widehat{\mu_k}} \\
\text{saturated model: } L(\widetilde{\mu}) &= \frac{Y_1^{Y_1}}{Y_1!}e^{-Y_1} \cdot \frac{Y_2^{Y_2}}{Y_2!}e^{-Y_2} \cdots \frac{Y_k^{Y_k}}{Y_k!}e^{-Y_k}
\end{aligned}
$$

The *residual deviance* of our model is defined as

$$
2 \cdot \left[\log\left(L(\widehat{\mu})\right) - \log\left(L(\widetilde{\mu})\right)\right].
$$

It measures how far our model is away from the theoretical optimum.

- The deviance is approximately $\chi^2_{\mathrm{df}}$ distributed, where df is the degrees of freedom of our model.

- The deviance is approximately $\chi^2_{\mathrm{df}}$ distributed, where df is the degrees of freedom of our model.
- Thus, the deviance should be of the same order of magnitude as df.

- The deviance is approximately $\chi^2_{\mathrm{df}}$ distributed, where df is the degrees of freedom of our model.
- Thus, the deviance should be of the same order of magnitude as df.
- Check this to assess the fit of the model!

- The deviance is approximately $\chi^2_{\mathrm{df}}$ distributed, where df is the degrees of freedom of our model.
- Thus, the deviance should be of the same order of magnitude as df.
- Check this to assess the fit of the model!

**Analysis of deviance:**

If $D_1$ and $D_2$ are the deviances of models $M_1$ with $p_1$ parameters and $M_2$ with $p_2$ parameters, and $M_1$ is nested in $M_2$ (i.e. the parameters of $M_1$ are a subset of the parameters of $M_2$), then $D_1 - D_2$ is approximately $\chi^2_{p_2-p_1}$-distributed.

- The deviance is approximately $\chi^2_{\mathrm{df}}$ distributed, where df is the degrees of freedom of our model.
- Thus, the deviance should be of the same order of magnitude as df.
- Check this to assess the fit of the model!

**Analysis of deviance:**
If $D_1$ and $D_2$ are the deviances of models $M_1$ with $p_1$ parameters and $M_2$ with $p_2$ parameters, and $M_1$ is nested in $M_2$ (i.e. the parameters of $M_1$ are a subset of the parameters of $M_2$), then $D_1 - D_2$ is approximately $\chi^2_{p_2-p_1}$-distributed.
This Test is the classical likelihood-ratio test. (Note that $D_1 - D_2$ is 2x the log of the likelihood-ratio of the two models.)

```
> pmod1 <- glm(counts~foodlevel+species,
                               data=daph,family=poisson)
> pmod2 <- glm(counts~foodlevel*species,
                               data=daph,family=poisson)
> anova(pmod1,pmod2,test="Chisq")

Analysis of Deviance Table

Model 1: counts ~ foodlevel + species
Model 2: counts ~ foodlevel * species
  Resid. Df Resid. Dev Df Deviance    P(>|Chi|)
1         9     6.1162
2         8     6.0741  1 0.042071    0.8375
```

# Why not the *F*-test?

Why not the $F$-test?
Remember that we did not estimate a variance $\sigma^2$ for the Poisson distribution.

Why not the $F$-test?
Remember that we did not estimate a variance $\sigma^2$ for the Poisson distribution.
There is an $F$-distribution approximation of a rescaled $D_1 - D_2$ for GLMs in which an extra variance parameter is estimated.

Why not the *F*-test?
Remember that we did not estimate a variance $\sigma^2$ for the Poisson distribution.
There is an *F*-distribution approximation of a rescaled $D_1 - D_2$ for GLMs in which an extra variance parameter is estimated.

Example: *overdispersed Poisson*, also called *quasipoisson* GLM. Here, $\mathbb{E}Y_i = \mu_i$ but $\mathrm{Var}Y_i = \phi \cdot \mu_i$ with the dispersion parameter $\phi > 1$.

Why not the *F*-test?
Remember that we did not estimate a variance $\sigma^2$ for the Poisson distribution.
There is an *F*-distribution approximation of a rescaled $D_1 - D_2$ for GLMs in which an extra variance parameter is estimated.

Example: *overdispersed Poisson*, also called *quasipoisson* GLM. Here, $\mathbb{E}\,Y_i = \mu_i$ but $\mathrm{Var}\,Y_i = \phi \cdot \mu_i$ with the dispersion parameter $\phi > 1$.
This is often used to model the influence of unknown external factors.

Why not the $F$-test?
Remember that we did not estimate a variance $\sigma^2$ for the Poisson distribution.
There is an $F$-distribution approximation of a rescaled $D_1 - D_2$ for GLMs in which an extra variance parameter is estimated.

Example: *overdispersed Poisson*, also called *quasipoisson* GLM. Here, $\mathbb{E}Y_i = \mu_i$ but $\mathrm{Var}\,Y_i = \phi \cdot \mu_i$ with the dispersion parameter $\phi > 1$.
This is often used to model the influence of unknown external factors.
Since the dispersion parameter is estimated, one can apply an $F$ approximation in the analysis of deviance.

Why not the $F$-test?
Remember that we did not estimate a variance $\sigma^2$ for the Poisson distribution.
There is an $F$-distribution approximation of a rescaled $D_1 - D_2$ for GLMs in which an extra variance parameter is estimated.

Example: *overdispersed Poisson*, also called *quasipoisson* GLM. Here, $\mathbb{E}\,Y_i = \mu_i$ but $\mathrm{Var}\,Y_i = \phi \cdot \mu_i$ with the dispersion parameter $\phi > 1$.
This is often used to model the influence of unknown external factors.
Since the dispersion parameter is estimated, one can apply an $F$ approximation in the analysis of deviance. But also $\chi^2$ is still an option.

Since the variance is proportional to the expectation value in the Poisson model, usual residuals are not so informatative.

Since the variance is proportional to the expectation value in the Poisson model, usual residuals are not so informatative.

Instead use *deviance residuals*. Let $d_i$ be the contribution of observation $i$ (row $i$ in the data table) to the Deviance, then the deviance residual of observation $i$ is

$$\mathrm{sign}(Y_i - \widehat{\mu}_i) \cdot \sqrt{d_i}.$$

Since the variance is proportional to the expectation value in the Poisson model, usual residuals are not so informatative.

Instead use *deviance residuals*. Let $d_i$ be the contribution of observation $i$ (row $i$ in the data table) to the Deviance, then the deviance residual of observation $i$ is

$$\operatorname{sign}(Y_i - \widehat{\mu}_i) \cdot \sqrt{d_i}.$$

The deviance residuals are the default residuals given by R for GLMs. They have similar properties as the standard residuals in the normal linear model.

Since the variance is proportional to the expectation value in the Poisson model, usual residuals are not so informatative.

Instead use *deviance residuals*. Let $d_i$ be the contribution of observation $i$ (row $i$ in the data table) to the Deviance, then the deviance residual of observation $i$ is

$$\mathrm{sign}(Y_i - \widehat{\mu}_i) \cdot \sqrt{d_i}.$$

The deviance residuals are the default residuals given by R for GLMs. They have similar properties as the standard residuals in the normal linear model.
In the following plot obtained with plot(pmod1) the word “residual” always refers to deviance residuals.

# Binomial GLM / logistic regression

In experiment $i$ (row $i$ of the data table) there are $n_i$ flies. Each of these flies decided independently of all other to go to the odorant with probability $p_i$ and, thus, to go to the fresh air with probability $(1 - p_i)$.

# Binomial GLM / logistic regression

In experiment $i$ (row $i$ of the data table) there are $n_i$ flies. Each of these flies decided independently of all other to go to the odorant with probability $p_i$ and, thus, to go to the fresh air with probability $(1 - p_i)$.

Thus, the number $Y_i$ of flies which went to the odorant is binomially distributed:

$$Y_i \sim \mathrm{bin}(n_i, p_i)$$

# Binomial GLM / logistic regression

In experiment $i$ (row $i$ of the data table) there are $n_i$ flies. Each of these flies decided independently of all other to go to the odorant with probability $p_i$ and, thus, to go to the fresh air with probability $(1 - p_i)$.

Thus, the number $Y_i$ of flies which went to the odorant is binomially distributed:

$$\begin{aligned} Y_i &\sim \mathrm{bin}(n_i, p_i) \\ \Pr(Y_i = k) &= \binom{n_i}{k} \cdot p_i^k \cdot (1 - p_i)^{n_i - k} \end{aligned}$$

# Binomial GLM / logistic regression

In experiment $i$ (row $i$ of the data table) there are $n_i$ flies. Each of these flies decided independently of all other to go to the odorant with probability $p_i$ and, thus, to go to the fresh air with probability $(1 - p_i)$.

Thus, the number $Y_i$ of flies which went to the odorant is binomially distributed:

$$\begin{aligned} Y_i &\sim \mathrm{bin}(n_i, p_i) \\ \Pr(Y_i = k) &= \binom{n_i}{k} \cdot p_i^k \cdot (1 - p_i)^{n_i - k} \\ \mathbb{E} Y_i &= n_i \cdot p_i \end{aligned}$$

# Binomial GLM / logistic regression

In experiment $i$ (row $i$ of the data table) there are $n_i$ flies. Each of these flies decided independently of all other to go to the odorant with probability $p_i$ and, thus, to go to the fresh air with probability $(1 - p_i)$.

Thus, the number $Y_i$ of flies which went to the odorant is binomially distributed:

$$\begin{aligned} Y_i &\sim \mathrm{bin}(n_i, p_i) \\ \Pr(Y_i = k) &= \binom{n_i}{k} \cdot p_i^k \cdot (1 - p_i)^{n_i - k} \\ \mathbb{E}\, Y_i &= n_i \cdot p_i \\ \mathrm{Var}\, Y_i &= n_i \cdot p_i \cdot (1 - p_i) \end{aligned}$$

# Binomial GLM / logistic regression

In experiment $i$ (row $i$ of the data table) there are $n_i$ flies. Each of these flies decided independently of all other to go to the odorant with probability $p_i$ and, thus, to go to the fresh air with probability $(1 - p_i)$.

Thus, the number $Y_i$ of flies which went to the odorant is binomially distributed:

$$\begin{aligned} Y_i &\sim \mathrm{bin}(n_i, p_i) \\ \Pr(Y_i = k) &= \binom{n_i}{k} \cdot p_i^k \cdot (1 - p_i)^{n_i - k} \\ \mathbb{E} Y_i &= n_i \cdot p_i \\ \mathrm{Var}\, Y_i &= n_i \cdot p_i \cdot (1 - p_i) \end{aligned}$$

How does $p_i$ depend on the odorant and on the species?

# Binomial GLM with logit link

Similar as in Poisson GLMs we assume:

$$\ell(p_i) \;=\; \eta_i \;=\; b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i}$$

# Binomial GLM with logit link

Similar as in Poisson GLMs we assume:

$$\ell(p_i) \;=\; \eta_i \;=\; b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i}$$

The default link of the Binomial GLM is the logit link:

$$\eta \;=\; \mathrm{logit}(p) \;=\; \log(p/(1-p))$$

# Binomial GLM with logit link

Similar as in Poisson GLMs we assume:

$$\ell(p_i) \;=\; \eta_i \;=\; b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i}$$

The default link of the Binomial GLM is the logit link:

$$\eta \;=\; \mathrm{logit}(p) \;=\; \log(p/(1-p))$$

Its inverse is the logistic function

$$p \;=\; \frac{1}{1+e^{-\eta}}$$

# Binomial GLM with logit link

Similar as in Poisson GLMs we assume:

$$\ell(p_i) \;=\; \eta_i \;=\; b_0 + b_1 \cdot X_{1,i} + \cdots + b_k \cdot X_{k,i}$$

The default link of the Binomial GLM is the logit link:

$$\eta \;=\; \mathrm{logit}(p) \;=\; \log(p/(1-p))$$

Its inverse is the logistic function

$$p \;=\; \frac{1}{1+e^{-\eta}}$$

Binomial GLM with the logit link is also called *logistic regression*.

# Likelihood and Deviance

If $\widehat{p}_1, \ldots, \widehat{p}_k$ are the estimated $p_i$ in our model, then the likelihood of the fitted parameters is

$$L(\widehat{p}) = \binom{n_1}{Y_1} \widehat{p}_1{}^{Y_1} (1 - \widehat{p}_1)^{n_1 - Y_1} \cdot \binom{n_2}{Y_2} \widehat{p}_2{}^{Y_2} (1 - \widehat{p}_2)^{n_2 - Y_2} \cdots \\ \cdots \binom{n_k}{Y_k} \widehat{p}_k{}^{Y_k} (1 - \widehat{p}_k)^{n_k - Y_k}$$

Using this likelihood, the *deviance* and the deviance residuals are defined like in the Poisson GLM.

# Analysis of deviance and overdispersion

Note that, like in the Poisson model, $\operatorname{Var} Y_i = n_i \cdot p_i \cdot (1 - p_i)$ is fixed for given $\mathbb{E} Y_i = n_i p_i$. Thus, the $\chi^2$ approximation should be used in the anaysis of deviance.

# Analysis of deviance and overdispersion

Note that, like in the Poisson model, $\mathrm{Var}\, Y_i = n_i \cdot p_i \cdot (1 - p_i)$ is fixed for given $\mathbb{E}\, Y_i = n_i p_i$. Thus, the $\chi^2$ approximation should be used in the anaysis of deviance.

There is an overdispersed binomial GLM (available in R with the option family=quasibinomial) with an additional dispersion parameter. For these models one can use both $\chi^2$ approximation and $F$ approximations in analyses of deviance.

A residual deviance of 1187.1 on 639 degrees of freedom (as observed in one of the example datasets) is very high and indicates that the model parameters cannot fully explain the data.

A residual deviance of 1187.1 on 639 degrees of freedom (as observed in one of the example datasets) is very high and indicates that the model parameters cannot fully explain the data.

$\Rightarrow$ Fit an overdispersed model!

A residual deviance of 1187.1 on 639 degrees of freedom (as observed in one of the example datasets) is very high and indicates that the model parameters cannot fully explain the data.

$\Rightarrow$ Fit an overdispersed model!

There is a price we have to pay for overdispersion: Since it is not a clearly defined distribution, AIC is not available for model selection.

A residual deviance of 1187.1 on 639 degrees of freedom (as observed in one of the example datasets) is very high and indicates that the model parameters cannot fully explain the data.

$\Rightarrow$ Fit an overdispersed model!

There is a price we have to pay for overdispersion: Since it is not a clearly defined distribution, AIC is not available for model selection.

Select parameters

1. that seem important to you from the biological context
2. or have low *p*-values.

# Compute an approx. 95% confidence range

```
> case <- data.frame(species="mel",odorant="CO2",sex="males")
> (pred <- predict(model4,case,type="link",se.fit=TRUE) )
$fit
-1.593086
$se.fit
[1] 0.1327248
$residual.scale
[1] 1.328106
> invlink <- function(x) {    ## inverse link function
+   1/(1+exp(-x))
+ }
> invlink(pred$fit)     ## prediction
0.1689501
> invlink(pred$fit-2*pred$se.fit)   ## lower bound
0.1348738
> invlink(pred$fit+2*pred$se.fit)   ## upper bound
0.2095506
```

This can be done simultaneously for a whole data frame (e.g. newdata) instead just for one on case (in our example mel/CO2/males)

This can be done simultaneously for a whole data frame (e.g. newdata) instead just for one on case (in our example mel/CO2/males)

Should be done on the linear predictor (“link”) scale and not on the response scale because it is based on a normal distribution approximation, which is only (more or less) valid on the linear predictor scale. (Remember: for a normal distribution, $>$ 95% are within the 2$\sigma$-bounds around the mean.)

# Model selection when AIC is not available.

- Apply backward model selection strategy: apply drop1 and remove the variable with the highest p-value. Apply drop1 on the reduced model and repeat this again and again until you only variables are left which are significant or almost significant.
- Variables will not be removed if they are involved in interactions, because drop1 wont show those variables.
- Do not a variable if there is a good biological reason why it should be in the model.

# GLMs and their links (canonical links first)

| | |
|---|---|
| Poisson | log($\mu$), $\mu$, $\sqrt{\mu}$ |
| binomial | logit, probit, cloglog |
| gaussian | $\mu$ |
| Gamma | $-1/\mu$, $\mu$, log($\mu$) |
| inverse gaussian | $-2/\mu^2$ |

Also interesting: **negative binomial** as alternative to overdispersed Poisson.

# Contents

Anova

Linear Models

Generalized Linear Models

Mixed-effects models

Principal Component Analysis (PCA)

Redundancy Analysis (RDA)

Correspondence Analysis (CA)

Canonical Correspondence Analysis (CCA)

QTL Mapping

We revisit the RIKZ dataset.

We revisit the RIKZ dataset.

Species abundance and many other covariates were measured at 9 beaches.

We revisit the RIKZ dataset.

Species abundance and many other covariates were measured at 9 beaches.

On every beach, 5 plots were sampled in the intertidal range.

We revisit the RIKZ dataset.

Species abundance and many other covariates were measured at 9 beaches.

On every beach, 5 plots were sampled in the intertidal range.

Each plot was sampled only once. Thus, each line in the data table corresponds to one plot.

- We are not interested in the precise effect of each beach

- We are not interested in the precise effect of each beach
- We do not want to estimate 8 extra paramters for the beaches

- We are not interested in the precise effect of each beach
- We do not want to estimate 8 extra paramters for the beaches
- Is there another way to take the difference between the beaches into account?

- We are not interested in the precise effect of each beach
- We do not want to estimate 8 extra paramters for the beaches
- Is there another way to take the difference between the beaches into account?
- Assume that the effect $\alpha_k$ of beach $k$ is random. Do not estimate all $\alpha_k$ but only their standard deviation $\sigma_\alpha$.

# Mixed-effects model

Let $S_i$ and $N_i$ be the ShannonW and the NAP observed at plot $i$, which is on beach $k$.

$$S_i = a + b \cdot N_i + \alpha_k + \varepsilon_i$$

# Mixed-effects model

Let $S_i$ and $N_i$ be the ShannonW and the NAP observed at plot $i$, which is on beach $k$.

$$S_i = a + b \cdot N_i + \alpha_k + \varepsilon_i$$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{45}$ are independently $\mathcal{N}(0, \sigma^2)$-distributed.

# Mixed-effects model

Let $S_i$ and $N_i$ be the ShannonW and the NAP observed at plot $i$, which is on beach $k$.

$$S_i = a + b \cdot N_i + \alpha_k + \varepsilon_i$$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{45}$ are independently $\mathcal{N}(0, \sigma^2)$-distributed.
$\alpha_1, \alpha_2, \ldots, \alpha_9$ are independently $\mathcal{N}(0, \sigma_\alpha^2)$-distributed.

# Mixed-effects model

Let $S_i$ and $N_i$ be the ShannonW and the NAP observed at plot $i$, which is on beach $k$.

$$S_i = a + b \cdot N_i + \alpha_k + \varepsilon_i$$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{45}$ are independently $\mathcal{N}(0, \sigma^2)$-distributed.
$\alpha_1, \alpha_2, \ldots, \alpha_9$ are independently $\mathcal{N}(0, \sigma_\alpha^2)$-distributed.
Mixed-effects: $a$ and $b$ are *deterministic*, $\alpha_1, \alpha_2, \ldots, \alpha_9$ are *random*.

# Mixed-effects model

Let $S_i$ and $N_i$ be the ShannonW and the NAP observed at plot $i$, which is on beach $k$.

$$S_i = a + b \cdot N_i + \alpha_k + \varepsilon_i$$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{45}$ are independently $\mathcal{N}(0, \sigma^2)$-distributed.
$\alpha_1, \alpha_2, \ldots, \alpha_9$ are independently $\mathcal{N}(0, \sigma_\alpha^2)$-distributed.
Mixed-effects: *a* and *b* are *deterministic*, $\alpha_1, \alpha_2, \ldots, \alpha_9$ are *random*.

To be estimated: *a*,*b*, $\sigma_\alpha$, $\sigma$.

```
> summary(mmod0)
Linear mixed model fit by REML
Formula: ShannonW ~ 1 + NAP + (1 | Beach)
   Data: rikz
   AIC   BIC logLik deviance REMLdev
 4.968 12.19  1.516   -12.27  -3.032
Random effects:
 Groups   Name        Variance Std.Dev.
 Beach    (Intercept) 0.017595 0.13264
 Residual             0.036504 0.19106
Number of obs: 45, groups: Beach, 9

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.46722    0.05366   8.707
NAP         -0.21380    0.03060  -6.987

Correlation of Fixed Effects:
    (Intr)
NAP -0.198
```

```
> summary(mmod0)
Linear mixed model fit by REML
Formula: ShannonW ~ 1 + NAP + (1 | Beach)
   Data: rikz
   AIC   BIC logLik deviance REMLdev
 4.968 12.19  1.516   -12.27  -3.032
Random effects:
 Groups   Name        Variance Std.Dev.
 Beach    (Intercept) 0.017595 0.13264
 Residual             0.036504 0.19106
Number of obs: 45, groups: Beach, 9

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.46722    0.05366   8.707
NAP         -0.21380    0.03060  -6.987

Correlation of Fixed Effects:
    (Intr)
NAP -0.198
```

What is REML?

```
> summary(mmod0)
Linear mixed model fit by REML
Formula: ShannonW ~ 1 + NAP + (1 | Beach)
   Data: rikz
   AIC   BIC logLik deviance REMLdev
 4.968 12.19  1.516   -12.27  -3.032
Random effects:
 Groups   Name        Variance Std.Dev.
 Beach    (Intercept) 0.017595 0.13264
 Residual             0.036504 0.19106
Number of obs: 45, groups: Beach, 9

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.46722    0.05366   8.707
NAP         -0.21380    0.03060  -6.987

Correlation of Fixed Effects:
    (Intr)
NAP -0.198
```

What is REML?

Why are there *t*-values but no *p*-values?

# REML vs. ML

- ML (Maximum Likelihood): estimate all parameters (here *a*, *b*, $\sigma_\alpha$, $\sigma$) by maximizing their joint likelihood.

# REML vs. ML

- ML (Maximum Likelihood): estimate all parameters (here *a*, *b*, $\sigma_\alpha$, $\sigma$) by maximizing their joint likelihood.
- REML (Restricted Maximum Likelihood): first estimate variance parameters (here $\sigma_\alpha$, $\sigma$) from the components of the response space that are orthogonal on components that can be explained by fixed effects. Using these estimates, the coefficients of the fixed effects (here *a* and *b*) are estimated with ML.

# REML vs. ML

- ML (Maximum Likelihood): estimate all parameters (here $a$, $b$, $\sigma_\alpha$, $\sigma$) by maximizing their joint likelihood.
- REML (Restricted Maximum Likelihood): first estimate variance parameters (here $\sigma_\alpha$, $\sigma$) from the components of the response space that are orthogonal on components that can be explained by fixed effects. Using these estimates, the coefficients of the fixed effects (here $a$ and $b$) are estimated with ML.
- Comparable to estimation of $\sigma^2$ from sample $X_1, \ldots, X_n$ by $\frac{1}{n-1}\sum_i(\mu_X - X_i)^2$ instead of the biased ML estimator $\frac{1}{n}\sum_i(\mu_X - X_i)^2$

# REML vs. ML

- ML (Maximum Likelihood): estimate all parameters (here *a*, *b*, $\sigma_\alpha$, $\sigma$) by maximizing their joint likelihood.
- REML (Restricted Maximum Likelihood): first estimate variance parameters (here $\sigma_\alpha$, $\sigma$) from the components of the response space that are orthogonal on components that can be explained by fixed effects. Using these estimates, the coefficients of the fixed effects (here *a* and *b*) are estimated with ML.
- Comparable to estimation of $\sigma^2$ from sample $X_1, \ldots, X_n$ by $\frac{1}{n-1}\sum_i(\mu_X - X_i)^2$ instead of the biased ML estimator $\frac{1}{n}\sum_i(\mu_X - X_i)^2$
- Also for fitting parameters of mixed-effects models, ML estimation is biased and REML is usually preferred.

# REML vs. ML

- ML (Maximum Likelihood): estimate all parameters (here $a$, $b$, $\sigma_\alpha$, $\sigma$) by maximizing their joint likelihood.
- REML (Restricted Maximum Likelihood): first estimate variance parameters (here $\sigma_\alpha$, $\sigma$) from the components of the response space that are orthogonal on components that can be explained by fixed effects. Using these estimates, the coefficients of the fixed effects (here $a$ and $b$) are estimated with ML.
- Comparable to estimation of $\sigma^2$ from sample $X_1, \ldots, X_n$ by $\frac{1}{n-1}\sum_i(\mu_X - X_i)^2$ instead of the biased ML estimator $\frac{1}{n}\sum_i(\mu_X - X_i)^2$
- Also for fitting parameters of mixed-effects models, ML estimation is biased and REML is usually preferred.
- ML estimation should be used when a likelihood ratio test shall be applied to models with different fixed effects and the same random effects.

# Why no *p*-values for the *t*-values?

- The *t*-values computed like in the usual linear model, but in the case of mixed-effects models they are in general not *t*-distributed (under the null hypothesis). Thus, it is not clear how to get *p*-values from the *t*-values.

# Why no *p*-values for the *t*-values?

- The *t*-values computed like in the usual linear model, but in the case of mixed-effects models they are in general not *t*-distributed (under the null hypothesis). Thus, it is not clear how to get *p*-values from the *t*-values.
- Some other programs give *p*-values which can be very imprecise.

# Why no *p*-values for the *t*-values?

- The *t*-values computed like in the usual linear model, but in the case of mixed-effects models they are in general not *t*-distributed (under the null hypothesis). Thus, it is not clear how to get *p*-values from the *t*-values.
- Some other programs give *p*-values which can be very imprecise.
- Exception: small balanced datasets. Here, *t*-values are approximately *t*-distributed and $|t| > 2$ usually indicates significance on the 5% level.

One possibility to visualize the estimations for the parameter and to assess their significance is based on sampling parameter values from their posterior distribution by an MCMC method.

One possibility to visualize the estimations for the parameter and to assess their significance is based on sampling parameter values from their posterior distribution by an MCMC method.

In contrast to most other methods discussed in this lecture, this is a Bayesian approach and thus needs prior distributions for the parameter values (or at least pseudo priors).

# General considerations of model selection

- What is the purpose of the model?
  1. Making predictions as precise as possible
  2. or to understand what the most influential paramters are?

# General considerations of model selection

- What is the purpose of the model?
  1. Making predictions as precise as possible
  2. or to understand what the most influential paramters are?
- In the first case AIC may be appropriate.

# General considerations of model selection

- What is the purpose of the model?
  1. Making predictions as precise as possible
  2. or to understand what the most influential paramters are?
- In the first case AIC may be appropriate.
- In the second case it may be better to use likelihood-ratio tests and remove all parameters which do not significantly improve the fit.

# General considerations of model selection

- What is the purpose of the model?
  1. Making predictions as precise as possible
  2. or to understand what the most influential paramters are?
- In the first case AIC may be appropriate.
- In the second case it may be better to use likelihood-ratio tests and remove all parameters which do not significantly improve the fit.
- Variable selection should not only depend on statistics but also on the relevance of the parameter for the biological question.

When random and fixed parameters have to be selected we apply the following strategy:

1. Start with a model that contains as many of the relevant parameters and interactions as possible.
2. First select random parameters. To decide between models which have different random parameters, fit models with REML and choose model of minimal AIC.
3. Now select fixed parameters. This can be done with the help of AIC or with likelihood ratio tests. If likelihood ratio tests are used, apply ML to fit the models to the data.
4. Never remove covariates that are still involved in interactions.
5. Fit the final model with REML.

Next, we fit a model where there is not only a random intercept for every beach but also a random coefficient of NAP. Again, let $S_i$ and $N_i$ be the ShannonW and the NAP observed at plot $i$, which is on beach $k$. The model says

$$S_i = a + [\text{fixed effects terms}] + \alpha_k + \beta_k \cdot N_i + \varepsilon_i.$$

$\varepsilon_1, \dots, \varepsilon_{45}$ are independently $\mathcal{N}(0, \sigma^2)$-distributed,
$\alpha_1, \dots, \alpha_9$ are independently $\mathcal{N}(0, \sigma_\alpha^2)$-distributed,
$\beta_1, \dots, \beta_9$ are independently $\mathcal{N}(0, \sigma_\beta^2)$-distributed,

Next, we fit a model where there is not only a random intercept for every beach but also a random coefficient of NAP. Again, let $S_i$ and $N_i$ be the ShannonW and the NAP observed at plot $i$, which is on beach $k$. The model says

$$S_i = a + [\text{fixed effects terms}] + \alpha_k + \beta_k \cdot N_i + \varepsilon_i.$$

$\varepsilon_1, \dots, \varepsilon_{45}$ are independently $\mathcal{N}(0, \sigma^2)$-distributed,
$\alpha_1, \dots, \alpha_9$ are independently $\mathcal{N}(0, \sigma_\alpha^2)$-distributed,
$\beta_1, \dots, \beta_9$ are independently $\mathcal{N}(0, \sigma_\beta^2)$-distributed,

Besides the fixed-effects coefficients we have to estimate $\sigma$, $\sigma_\alpha$ and $\sigma_\beta$.

Don’t trust the *p*-values on the previous slide! The problem is not only that the models were fitted with REML. The main problem ist that the null hypotheses (e.g. $\sigma_\beta = 0$ in the case of B2/B3) are on the boundary of the parameter space. $\sigma_\beta$ can only be $\geq 0$, and deviations from $\sigma_\beta = 0$ are thus only possible in one direction. The $\chi^2$-approximation of likelihood ratio tests are only reliable when deviations from the expectation under the null hypothesis are possible in all directions, for example if the null hypothesis $\theta = 0$ is tested for some parameter $\theta$, and estimates of $\theta$ can lead to positive as well as negative values.

Don't trust the *p*-values on the previous slide! The problem is not only that the models were fitted with REML. The main problem ist that the null hypotheses (e.g. $\sigma_\beta = 0$ in the case of B2/B3) are on the boundary of the parameter space. $\sigma_\beta$ can only be $\geq 0$, and deviations from $\sigma_\beta = 0$ are thus only possible in one direction. The $\chi^2$-approximation of likelihood ratio tests are only reliable when deviations from the expectation under the null hypothesis are possible in all directions, for example if the null hypothesis $\theta = 0$ is tested for some parameter $\theta$, and estimates of $\theta$ can lead to positive as well as negative values.

Thus, we rather base our decision on the AIC values. This is, of course, also not stringent. However, in our case, all criteria favor model B2.

- Generalized linear mixed-effects models can be fitted with the glmer command in the lme4 package.

- Generalized linear mixed-effects models can be fitted with the glmer command in the lme4 package.
- REML is not applied, more complex algorithms are applied to fit models.

- Generalized linear mixed-effects models can be fitted with the glmer command in the lme4 package.
- REML is not applied, more complex algorithms are applied to fit models.
- All *p*-values can be very imprecise, so do not trust them too much, especially if they are close to the significance level.

- Generalized linear mixed-effects models can be fitted with the glmer command in the lme4 package.
- REML is not applied, more complex algorithms are applied to fit models.
- All *p*-values can be very imprecise, so do not trust them too much, especially if they are close to the significance level.
- Mcmc methods or other nice methods to visualize the results of a mixed-effects GLM are not yet implemented in lme4.

- Generalized linear mixed-effects models can be fitted with the glmer command in the lme4 package.
- REML is not applied, more complex algorithms are applied to fit models.
- All *p*-values can be very imprecise, so do not trust them too much, especially if they are close to the significance level.
- Mcmc methods or other nice methods to visualize the results of a mixed-effects GLM are not yet implemented in lme4.
- As an example we fit an overdispersed Poisson model to the RIKZ data with Richness as the response variable.

# Contents

Anova

Linear Models

Generalized Linear Models

Mixed-effects models

Principal Component Analysis (PCA)

Redundancy Analysis (RDA)

Correspondence Analysis (CA)

Canonical Correspondence Analysis (CCA)

QTL Mapping

# Reading biplots

Distance biplot (scale=0)

- Angles between lines are meaningless.
- The lines are projections of length 1 vectors into the plane of the first two principal components. So the length indicates how well the corresponding variable is represented by the first two components.
- Distances between points/labels approximate distances of the observations for different objects.
- The projection of a point onto a vector at right angle approximates the position of the corresponding object along the corresponding variable.

Correlation biplot (scale=1)

- The cosine of the angle between two lines is approximately equal to the correlation between the corresponding variables.
- If the PCA used `scale=FALSE`, then the length of a line is approximately $\sqrt{N-1}$ times the estimated standard deviation of the corresponding variable. If the PCA used `scale=TRUE`, then the lines are projections of length $\sqrt{N-1}$ vectors into the plane of the first two principal components. So the length indicates how well the corresponding variable is represented by the first two components.
- Distances between points/labels are meaningless.
- The projection of a point onto a vector at right angle approximates the position of the corresponding object along the corresponding variable.

One problem with PCA is to decide how many components to present, and there are various rules of thumb.

- 80%-rule: Present the first $k$ axes that explain 80% of the total variation.

One problem with PCA is to decide how many components to present, and there are various rules of thumb.

- 80%-rule: Present the first $k$ axes that explain 80% of the total variation.
- ellbow-rule: Plot the eigenvalues as vertical lines or bars next to each other. Use $k$ axes if the 'elbow' is at $k + 1$.

One problem with PCA is to decide how many components to present, and there are various rules of thumb.

- 80%-rule: Present the first $k$ axes that explain 80% of the total variation.
- ellbow-rule: Plot the eigenvalues as vertical lines or bars next to each other. Use $k$ axes if the 'elbow' is at $k+1$.
- broken-stick-rule: If a stick of unit length is broken at random in $p$ pieces, then the expected length of piece number $j$ is given by

  $$L_j = \frac{1}{p}\sum_{i=j}^{p}\frac{1}{i} \tag{1}$$

  If the eigenvalue of the $j$-th axis is larger than $L_j$, then it can be considered as important.

One problem with PCA is to decide how many components to present, and there are various rules of thumb.

- 80%-rule: Present the first $k$ axes that explain 80% of the total variation.
- ellbow-rule: Plot the eigenvalues as vertical lines or bars next to each other. Use $k$ axes if the 'elbow' is at $k+1$.
- broken-stick-rule: If a stick of unit length is broken at random in $p$ pieces, then the expected length of piece number $j$ is given by

  $$L_j = \frac{1}{p}\sum_{i=j}^{p}\frac{1}{i} \tag{1}$$

  If the eigenvalue of the $j$-th axis is larger than $L_j$, then it can be considered as important.

The broken-stick-model is the most reliable rule of thumb.

In many cases the different variables are on different scales. Then you are recommended to scale the variables with their standard deviations, that is, to use the correlation matrix rather than the covariance matrix.
Otherwise the first principal component might be dominated by the variable with the largest scale.

In many cases the different variables are on different scales. Then you are recommended to scale the variables with their standard deviations, that is, to use the correlation matrix rather than the covariance matrix.
Otherwise the first principal component might be dominated by the variable with the largest scale.
For you this means to use the argument `scale=TRUE` in the `prcomp()` command.

In many cases the different variables are on different scales. Then you are recommended to scale the variables with their standard deviations, that is, to use the correlation matrix rather than the covariance matrix.
Otherwise the first principal component might be dominated by the variable with the largest scale.
For you this means to use the argument `scale=TRUE` in the `prcomp()` command.

If the values of the variables are of comparable order, then it is also fine to not scale the variables, that is, to apply PCA to the covariance matrix.
In `R` this means to use the argument `scale=FALSE`.

# Summary

Principal component analysis is a transformation
(rotation and reflection) of the data such that
most of the variation is on the first axis,
the second most variation is on the second axis...

# Summary

Principal component analysis is a transformation
(rotation and reflection) of the data such that
most of the variation is on the first axis,
the second most variation is on the second axis...

Used for:

- Visualizing multi-variate data (we have no better method)

# Summary

Principal component analysis is a transformation
(rotation and reflection) of the data such that
most of the variation is on the first axis,
the second most variation is on the second axis...

Used for:

- Visualizing multi-variate data (we have no better method)
- Get a feeling on the dependencies

# Summary

Principal component analysis is a transformation
(rotation and reflection) of the data such that
most of the variation is on the first axis,
the second most variation is on the second axis...

Used for:

- Visualizing multi-variate data (we have no better method)
- Get a feeling on the dependencies
- Find clusters in the variables
  (e.g. $\{X1, X2\}$ and $\{X3, X4\}$ in the EWU data set)

# Summary

Principal component analysis is a transformation
(rotation and reflection) of the data such that
most of the variation is on the first axis,
the second most variation is on the second axis...

Used for:

- Visualizing multi-variate data (we have no better method)
- Get a feeling on the dependencies
- Find clusters in the variables
  (e.g. $\{X1, X2\}$ and $\{X3, X4\}$ in the EWU data set)
- Find clusters in the set of objects/individuals
  (e.g. girls and guys in the height and weight data)

Be aware:

- Principal components can often not be interpreted
  $2 * \text{shoe} + 3 * \text{height}$ is a measure for size
  But how shall we interpret $2 * \text{shoe} - \text{height}$?

Be aware:

- Principal components can often not be interpreted
  $2 * \text{shoe} + 3 * \text{height}$ is a measure for size
  But how shall we interpret $2 * \text{shoe} - \text{height}$?
- If first two principal components explain less than 70%, then consider forgetting PCA

Be aware:

- Principal components can often not be interpreted
  $2 * \text{shoe} + 3 * \text{height}$ is a measure for size
  But how shall we interpret $2 * \text{shoe} - \text{height}$?
- If first two principal components explain less than 70%, then consider forgetting PCA
- Biplots are easily misread. Be careful!

Be aware:

- Principal components can often not be interpreted
  $2 * \text{shoe} + 3 * \text{height}$ is a measure for size
  But how shall we interpret $2 * \text{shoe} - \text{height}$?
- If first two principal components explain less than 70%, then consider forgetting PCA
- Biplots are easily misread. Be careful!
- It’s spelled ’principal’ (main, Haupt-),
  not ’principle’ (Prinzip, Grundsatz)

# Contents

Anova

Linear Models

Generalized Linear Models

Mixed-effects models

Principal Component Analysis (PCA)

Redundancy Analysis (RDA)

Correspondence Analysis (CA)

Canonical Correspondence Analysis (CCA)

QTL Mapping

Given: Data frames/matrices $Y$ and $X$
The variables in $X$ are called explanatory variables
The variables in $Y$ are called response variables

Given: Data frames/matrices $Y$ and $X$
The variables in $X$ are called explanatory variables
The variables in $Y$ are called response variables

Goal: Find those components of $Y$ which are linear combinations of $X$ and (among those) represent as much variance of $Y$ as possible.

Given: Data frames/matrices $Y$ and $X$
The variables in $X$ are called explanatory variables
The variables in $Y$ are called response variables

Goal: Find those components of $Y$ which are linear combinations of $X$ and (among those) represent as much variance of $Y$ as possible.

Assumption: There is a linear dependence of the response variables in $Y$ on the explanatory variables in $X$.

Given: Data frames/matrices $Y$ and $X$
The variables in $X$ are called explanatory variables
The variables in $Y$ are called response variables

Goal: Find those components of $Y$ which are linear combinations of $X$ and (among those) represent as much variance of $Y$ as possible.

Assumption: There is a linear dependence of the response variables in $Y$ on the explanatory variables in $X$.

The idea behind redundancy analysis is to apply linear regression in order to represent $Y$ as linear function of $X$ and then to use PCA in order to visualize the result.

Given: Data frames/matrices $Y$ and $X$
The variables in $X$ are called explanatory variables
The variables in $Y$ are called response variables

Goal: Find those components of $Y$ which are linear combinations of $X$ and (among those) represent as much variance of $Y$ as possible.

Assumption: There is a linear dependence of the response variables in $Y$ on the explanatory variables in $X$.

The idea behind redundancy analysis is to apply linear regression in order to represent $Y$ as linear function of $X$ and then to use PCA in order to visualize the result.

Among those components of $Y$ which can be linearly explained with $X$ (multivariate linear regression) take those components which represent most of the variance.

Before applying RDA:

- Is $Y$ increasing with increasing values of $X$?
- If the variables in $X$ are twice as high, are the variables in $Y$ also approximately twice as high?

These questions are to check the assumption of linear dependence.

The graphical output of RDA consists of two biplots on top of each other and is called `triplot`.
You produce a triplot with `plot(rda.object)` (which itself calls `plot.cca()`).

The graphical output of RDA consists of two biplots on top of each other and is called `triplot`.
You produce a triplot with `plot(rda.object)` (which itself calls `plot.cca()`).

There are three components in a triplot:

- Continuous explanatory variables (numeric values) are represented by lines. Nominal explanatory variables (factor object) (coded $0 - 1$) by squares (or triangles) (one for each level). The square is plotted at the centroid of the observations that have the value 1.

The graphical output of RDA consists of two biplots on top of each other and is called `triplot`.
You produce a triplot with `plot(rda.object)` (which itself calls `plot.cca()`).

There are three components in a triplot:

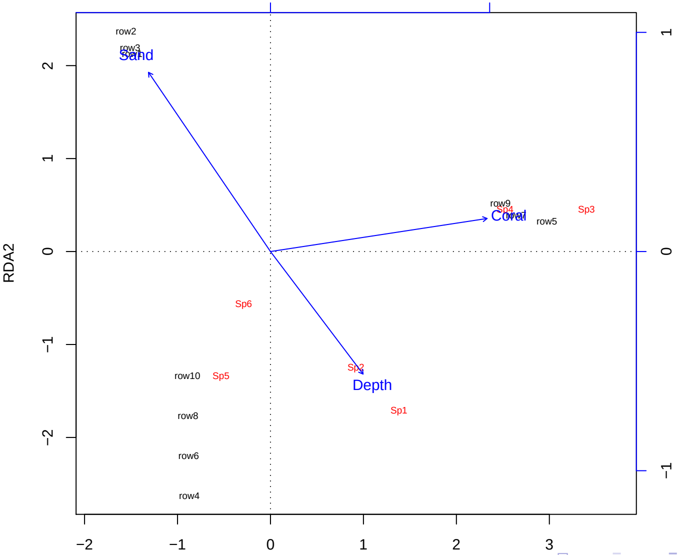
- Continuous explanatory variables (numeric values) are represented by lines. Nominal explanatory variables (factor object) (coded $0 - 1$) by squares (or triangles) (one for each level). The square is plotted at the centroid of the observations that have the value 1.
- The response variables by labels or lines.

The graphical output of RDA consists of two biplots on top of each other and is called `triplot`.
You produce a triplot with `plot(rda.object)` (which itself calls `plot.cca()`).

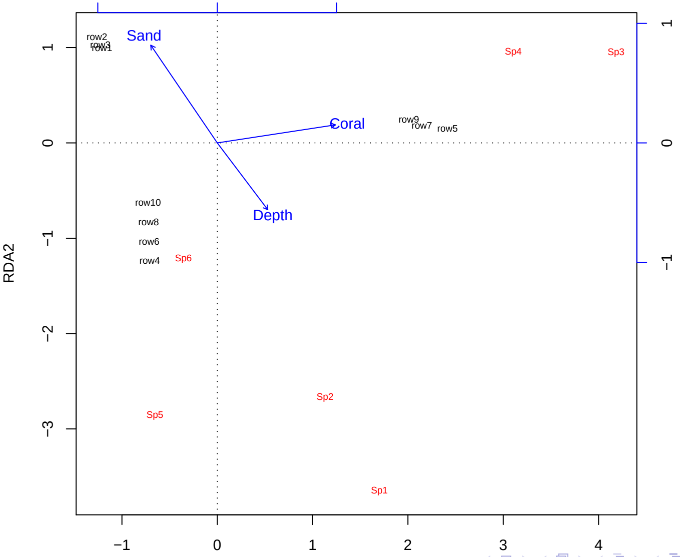
There are three components in a triplot:

- Continuous explanatory variables (numeric values) are represented by lines. Nominal explanatory variables (factor object) (coded $0 - 1$) by squares (or triangles) (one for each level). The square is plotted at the centroid of the observations that have the value 1.
- The response variables by labels or lines.
- The observations by points or labels.

# Correlation triplot



# Distance triplot



Distance triplot (scaling=1)

- Distances between points (observations), between squares or between points and squares approximate distances of the observations (or the centroid of the nominal explanatory variable).
- Angles between lines of response variables and lines of explanatory variables represent a two-dimensional approximation of correlations.
- Other angles between lines are meaningless.

Distance triplot (scaling=1)

- Distances between points (observations), between squares or between points and squares approximate distances of the observations (or the centroid of the nominal explanatory variable).
- Angles between lines of response variables and lines of explanatory variables represent a two-dimensional approximation of correlations.
- Other angles between lines are meaningless.
- The projection of a point onto the line of a response variable at right angle approximates the position of the corresponding object along the corresponding variable.
- Squares/triangles cannot be compared with lines of qualitatively explanatory variables.

Correlation triplot (scaling=2)

- The cosine of the angle between lines (of response variable or of explanatory variable) is approximately equal to the correlation between the corresponding variables.

Correlation triplot (scaling=2)

- The cosine of the angle between lines (of response variable or of explanatory variable) is approximately equal to the correlation between the corresponding variables.
- Distances are meaningless.

Correlation triplot (scaling=2)

- The cosine of the angle between lines (of response variable or of explanatory variable) is approximately equal to the correlation between the corresponding variables.
- Distances are meaningless.
- The projection of a point onto a line (response variable or explanatory variable) at right angle approximates the value of the corresponding variable of this observation.

Correlation triplot (scaling=2)

- The cosine of the angle between lines (of response variable or of explanatory variable) is approximately equal to the correlation between the corresponding variables.
- Distances are meaningless.
- The projection of a point onto a line (response variable or explanatory variable) at right angle approximates the value of the corresponding variable of this observation.
- The length of lines are not important.

# Contents

Anova

Linear Models

Generalized Linear Models

Mixed-effects models

Principal Component Analysis (PCA)

Redundancy Analysis (RDA)

Correspondence Analysis (CA)

Canonical Correspondence Analysis (CCA)

QTL Mapping

# Correspondence analysis

Given: Data frame/matrix $Y$

$Y[i, \cdot]$ are the observations of species `i`

$Y[\cdot, j]$ are the observations at site `j`

# Correspondence analysis

Given: Data frame/matrix $Y$
$Y[i, \cdot]$ are the observations of species `i`
$Y[\cdot, j]$ are the observations at site `j`

Goal: Find associations of species and sites

# Correspondence analysis

Given: Data frame/matrix $Y$

$Y[i, \cdot]$ are the observations of species `i`

$Y[\cdot, j]$ are the observations at site `j`

Goal: Find associations of species and sites

Assumption: There is a niche dependence of the species on the environmental variables

# Correspondence analysis

Given: Data frame/matrix $Y$
$Y[i,\cdot]$ are the observations of species `i`
$Y[\cdot,j]$ are the observations at site `j`

Goal: Find associations of species and sites

Assumption: There is a niche dependence of the species on the environmental variables

The setting is formulated here in terms of species and sites.
If you have measured quantities (variables) of some objects, then replace 'species' by 'object' and 'site' by 'variable'.

Instead of frequencies we now consider probabilities

$$p[i,k] := Y[i,k]/n$$

and define a matrix $Q$ with entries

$$Q[i,k] := \frac{p[i,k] - p[i,+] \cdot p[+,k]}{\sqrt{p[i,+]p[+,k]}}$$

Now all further steps are just as in PCA with the centred/normalized matrix $Y$ replaced by the association matrix $Q$. Again we get a distance biplot and a correlation biplot.

Instead of frequencies we now consider probabilities

$$p[i,k] := Y[i,k]/n$$

and define a matrix $Q$ with entries

$$Q[i,k] := \frac{p[i,k] - p[i,+] \cdot p[+,k]}{\sqrt{p[i,+]p[+,k]}}$$

Now all further steps are just as in PCA with the centred/normalized matrix $Y$ replaced by the association matrix $Q$. Again we get a distance biplot and a correlation biplot.

Correspondence analysis assesses
the association between species and sites
(or objects and variables)

The position of a species represents the optimum value in terms of the Gausian response model (niche) along the first and second axes. For this reason, species scores are represented as labels or points.

The position of a species represents the optimum value in terms of the Gausian response model (niche) along the first and second axes. For this reason, species scores are represented as labels or points.

Site conditional biplot (scaling=1)

- The sites are the centroids of the species, that is, sites are plotted close to the species which occur at those sites.
- Distances between sites are two-dimensional approximations of their Chi-square distances. So sites close to each other are similar in terms of the Chi-square distance.

Species conditional biplot (scaling=2)

- The species are the centroids of the sites, that is, species are plotted close to the sites where they occur.
- Distances between species are two-dimensional approximations of their Chi-square distances. So species close to each other are similar in terms of the Chi-square distance.

There is also a joint plot of species and site scores (scaling=3). In this plot distances between sites and distances between species can be interpreted as the approximations of the respective Chi-square distances. However the relative positions of sites and frequencies cannot be interpreted. So this biplot is to be used with care if used at all.

Note:

- The total inertia (or total variance) in correspondence analysis is defined as the Chi-square statistic of the site-by-species table divided by the total number of observations.
- Points further away from the origin in a biplot are the most interesting as these points make a relatively high contribution to the Chi-square statistic. So the further away from the origin a site is plotted, the more different it is from the average site.

# Contents

Anova

Linear Models

Generalized Linear Models

Mixed-effects models

Principal Component Analysis (PCA)

Redundancy Analysis (RDA)

Correspondence Analysis (CA)

Canonical Correspondence Analysis (CCA)

QTL Mapping

Given: Data frames/matrices $Y$ and $X$
$Y[i,\cdot]$ are the observations of species `i`
$Y[\cdot,j]$ are the observations at site `j`
$X$ are the explanatory variables

Given: Data frames/matrices $Y$ and $X$
$Y[i, \cdot]$ are the observations of species `i`
$Y[\cdot, j]$ are the observations at site `j`
$X$ are the explanatory variables

Goal: Find associations of species abundancies and sites with each environmental condition on a site being a linear combination of the environmental variables of $X$.

Given: Data frames/matrices $Y$ and $X$
$Y[i, \cdot]$ are the observations of species `i`
$Y[\cdot, j]$ are the observations at site `j`
$X$ are the explanatory variables

Goal: Find associations of species abundancies and sites with each environmental condition on a site being a linear combination of the environmental variables of $X$.

Assumption: There is a niche dependence of the species on environmental factors

The species scores, the site scores and the environmental scores are plotted in a figure called a triplot (confer with triplots in RDA). These triplots are the biplots from CA with additionally the explanatory variables plotted as lines.
Again the position of a species represents the optimum value in terms of the Gausian response model (niche) along the first and second axes. For this reason, species scores are represented as labels or points.

The species scores, the site scores and the environmental scores are plotted in a figure called a triplot (confer with triplots in RDA). These triplots are the biplots from CA with additionally the explanatory variables plotted as lines.
Again the position of a species represents the optimum value in terms of the Gausian response model (niche) along the first and second axes. For this reason, species scores are represented as labels or points.

In addition: Species can be projected perpendicular (=orthogonally) on the lines showing the species optima of the respective explanatory variables (in the respective scaling). Projecting sites perpendicular on the lines results in the values of the respective environmental variable at those sites.

The angle between lines does NOT represent correlation between the variables. Instead if the tip of a line is projected on another line or an axis then the resulting value represents a weighted correlation.

# When PCA, RDA, CA, CCA?

Summary of methods:

- Relationships in PCA and RDA are linear.
- In RDA and CCA two sets of variables are used, and a cause-effect relationship is asssumed.

# When PCA, RDA, CA, CCA?

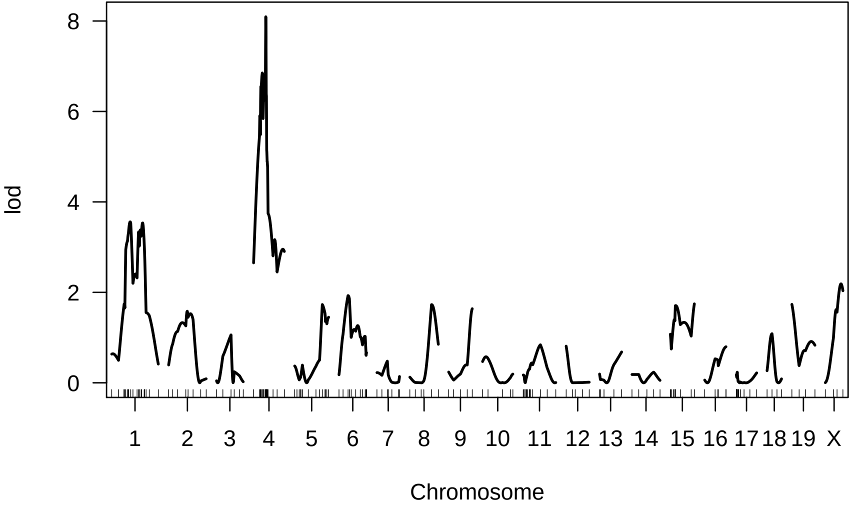
Summary of methods:

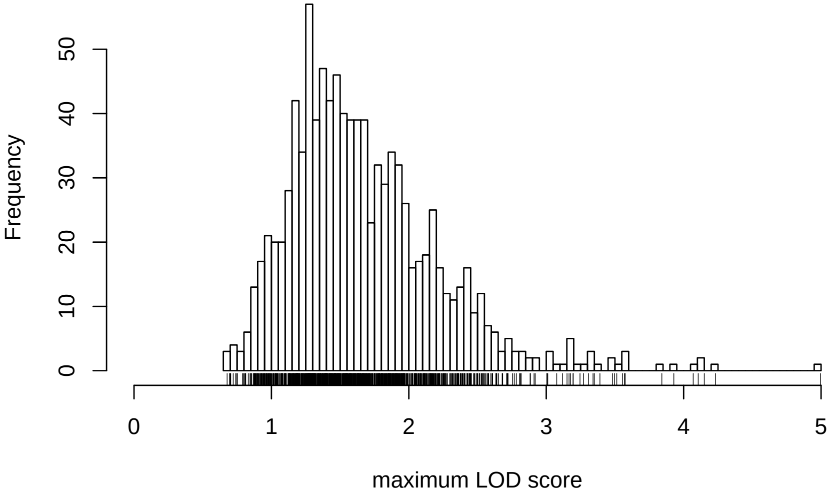
- Relationships in PCA and RDA are linear.
- In RDA and CCA two sets of variables are used, and a cause-effect relationship is asssumed.

| | Pure ordination | Cause-effect relation |
|---|---|---|
| Linear model | PCA | RDA |
| Unimodal model | CA | CCA |

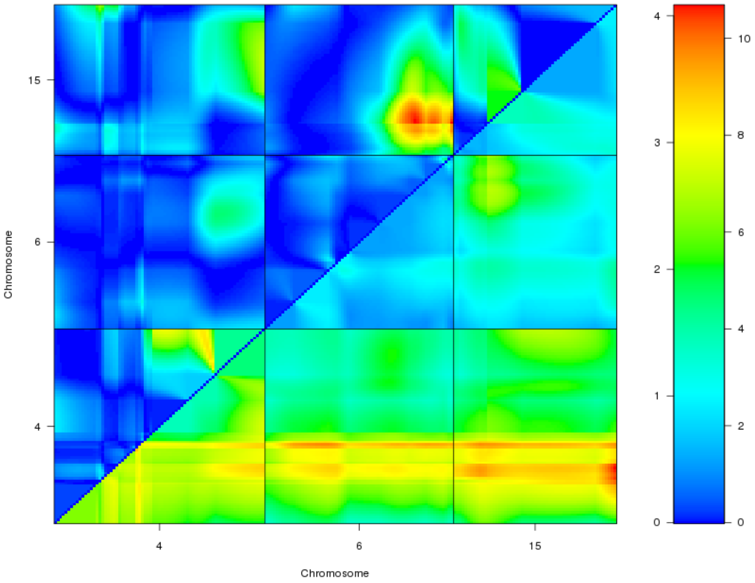
# Contents

- Anova
- Linear Models
- Generalized Linear Models
- Mixed-effects models
- Principal Component Analysis (PCA)
- Redundancy Analysis (RDA)
- Correspondence Analysis (CA)
- Canonical Correspondence Analysis (CCA)
- QTL Mapping


8
6
4
2
0
lod
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Chromosome


Frequency
0
10
20
30
40
50
0
1
2
3
4
5
maximum LOD score

# QTL Mapping



- Candidate loci and interactions found by scanone and scantwo can then be used in multiple QTL analysis.
- Then, p-values from multiple QTL analysis are not reliable because not multiple-testing corrected. Massive multiple-testing problem is caused by preselection by scanone and scantwo.
- If two QTL are close to each other with only few marker loci inbetween, scanone may falsely indicate strong evidence for one QTL between the two.