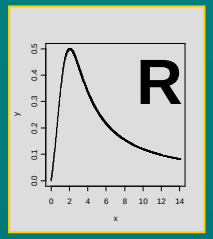


Analysis of biological data using GLM

Institute of Botany and Zoology
Masaryk University, Brno

Stano Pekár

Content

1) R Environment
   Exploratory Data Analysis
   Regression models

2) The first example
   Systematic components

3) Stochastic components
   Analyses of continual measurements

4) Analyses of continual measurements II
   Analyses of counts

5) Analyses of counts II
   Analyses of proportions

Literature

Crawley M. 2002. *Statistical computing. An Introduction to Data Analysis using S-Plus*. Wiley & Sons.

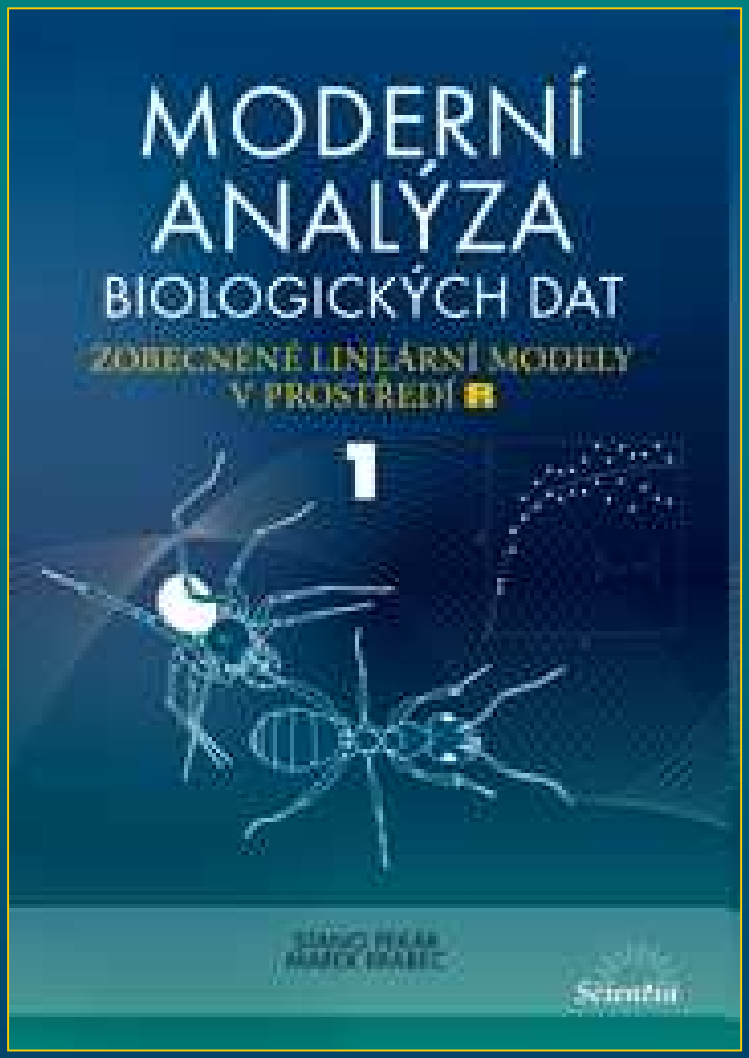
Crawley M. 2005. *Statistics: An Introduction Using R*. Wiley.

Dalgaard P. 2004. *Introductory Statistics with R*. Springer.

Faraway J.J. 2005. *Linear Models with R*. Chapman & Hall/CRC.

Fox J. 2002. *An R and S-Plus Companion to Applied Regression*. Sage.

Maindonald J. & Braun J. 2003. *Data Analysis and Graphics Using R*. Cambridge.



Pekár S. & Brabec M. 2009.
Modern Analysis of Biological Data.1. Generalised Linear Models in R.
Scientia, Praha. [in Czech]
http://www.scientia.cz/katalog.asp?idk=161&nav=&id=1931

Statistical analysis

- very fast due to use of computers
- chose statistical models that approach data characters

<u>This course</u>

- focuses on regression models in a broad sense
- only on linear models
- with only one response variable (**univariate** methods)
- with independent observations

Variables

Response variable

• (dependent) is the variable whose variation we aim to understand, the variable that we measure, it goes on ordinate

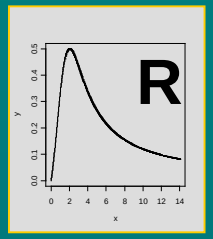
- continuous measurement, count, proportion (***y***)

Explanatory variable

• (independent) is the variable that we manipulate (select levels), interested to what extent is variation in response associated with variation in explanatory variables, displayed on abscissa

- numeric: continuous or discrete measurements (***x***) .. covariate

- categorical .. a factor (***A***, ***B***) with two or more levels (A_1, A_2, .. B_1, B_2, ..)

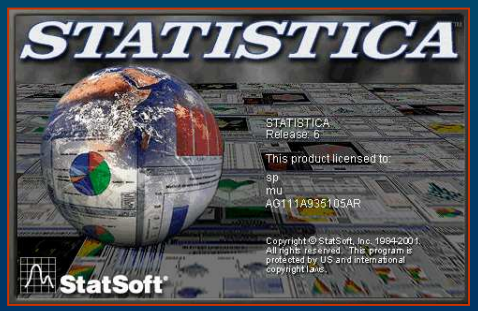


Statistical software

Stano Pekár

Software

- software packages that include GLM



What is R ?

• environment for the manipulation of objects
- data manipulation, calculation and graphical display
- a high-level programming language

• combination of S (developed at AT&T Bell Laboratories and forms the basis of the S-PLUS systems) and Scheme languages

• initially written by Gentleman & Ihaka (1996), nowadays with many contributors (R Development Core Team)
- includes about 30 standard packages
- available 2000 additional packages

• user-unfriendly (limited pull-down menus)
- based on commands
- pull-down menus only for basic commands

Why R ?

Pros

- freeware
- one of the largest statistical systems
- open environment with more dynamic development than other systems
- whereas Statistica or SAS will give copious output, R will provide minimal output
- makes you think about the analysis

Cons

- no warranty
- user-unfriendly

Instalation

- available from www.r-project.com
- copy data to C:\Program Files\R\R-2.9.0\MABD
- install sciplot package

Basic operations

+ - * / > <

== equal

!= not equal

<= less than or equal

^ power

• logical values **T** .. TRUE, **F** .. FALSE

Functions

• trigonometric

sin, cos, tan, asin, acos, atan

• logarithmic: **log**, **log2**, **log10**

• **sqrt, exp, abs, sum, prod**

• **seq, c, which, length, cbind, xbind, matrix**

• names are case sensitive

• “_” is not allowed to use

• avoid using names: **break**, **c, C, D, diff, else, F, FALSE, for, function, I, if, in Inf, mean, NA, NaN, next, NULL, pi, q, range, rank, repeat, s, sd, t, T, tree, TRUE, var, while**

• vectors: numeric, character, logical

• arguments (in parentheses): use their names or without at specified order

• centring: to subtract mean
• scaling: to divide by SD

Data frames

Created in R:

• use **`data.frame`**, **`rep`**, **`factor`**, **`levels`**, **`relevel`**

• export: **`write.table`**

Imported:

- from Excel via clipboard

```
dat <- read.delim("clipboard")
```

or via TXT file

```
dat <- read.delim("c:\\MABD\\metal.txt")
```

• data matrix:

- number of columns = number of variables

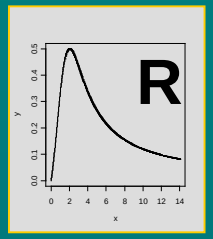
- first row contains names of variables (names without blank spaces)

- each row corresponds to an observation (trial, etc.)
- factors levels can be names or coded as numbers
- all columns must have the same number of rows
- missing data are assigned as **NA**

- `is.na`
- `$`

```
attach(dat)
names(dat)
```

soil	field	distance	amount
moist	pasture	12	0.22
moist	pasture	22	0.11
moist	pasture	43	0.29
moist	pasture	23	0.33
moist	rape	32	0.19
moist	rape	67	0.39
moist	rape	54	0.18
moist	rape	NA	0.29
dry	pasture	11	1.16
dry	pasture	33	1.03
dry	pasture	45	1.11
dry	pasture	NA	1.33
dry	rape	55	1.02
dry	rape	41	1.23
dry	rape	14	1.05
dry	rape	27	1.12



Exploratory Data Analysis

Stano Pekár

EDA

- a visual (tabular or graphical) analysis of the data

Important to
- check errors
- get an idea of the result
- suggest a model
- check assumptions for use of desired methods
- set hypotheses
- look for unexpected trends

- use expected values and variation

Expected value

• E(y), μ: theoretical long-term average of a variable

- one of a few characteristics of a distribution
- for discrete distributions E(y) might not be a possible value
- estimate of E(y) is **mean** ... `mean`
- a robust estimate for asymmetric distributions is **median**: ... `median`
- another robust estimate is **trimmed mean**: mean where α*n observations are removed from each tail ... `mean(y, trim=)`

Example

Find mean, median, and mean trimmed by 10% of the *amount* variable.

Data:

`metal.txt`

Variance

- Var(y), σ^2: a theoretical measure of the variability in a variable
- minimum and maximum ... **`range, min, max`**
- quantiles (0, 25, 50, 75, 100%) ... **`quantile`**
- estimate of Var(y) is s^2... **`var`**
- standard deviation (s) ... **`sd`**

- standard error of the mean ... $$SEM = \frac{s}{\sqrt{n}}$$

Example

Find variance, standard deviation, range and standard error of the mean for *amount*.

Confidence Intervals

• of a parameter (mean): if large number of samples is taken froma population then α% of intervals will contain mean
• based on quantiles of the t distribution **`qt`**

- lower CI_{95} $\bar{y} - t_{0.975,\nu} \times SEM$

- upper CI_{95} $\bar{y} + t_{0.975,\nu} \times SEM$

$\nu = n - 1$

• for asymmetric distributions CI_{95} is estimated on transformed values → asymmetric intervals
• from model objects use function **`confint`**

Example
Find 95% confidence intervals of mean for *amount*.

Tabular analysis

• basic summaries (min, max, Q_{25}, Q_{75}, median, mean) for all variables.. **`summary`**

• summary table for data with explanatory variable(s) .. **`tapply`**

• to count frequencies .. **`table`**

Example

Make a summary table, table of replications for *FIELD*, table of means for *SOIL* and *FIELD*, and table of SEM for *FIELD*.

Graphics

- see **`demo(graphics)`** or **`demo(image)`**
- graphs
- basic: **`plot`**
- advanced: **`xyplot`** (library *lattice*)

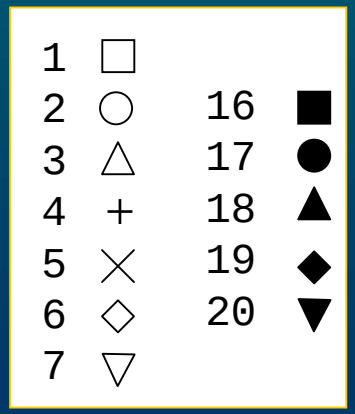
- to get all graphic parameters: **`?par`**
- to split window to subplots: **`par(mfrow)`**
- to add legend .. **`legend`**
- graph window size: **`x11`**

plot

Argument	Values
type=	Style: "**n**" (empty), "**p**" (scatter), "**l**" (lines), "**b**" (both), "**h**" (vertical)
las=	Style of axes values: **0** (parallel), **1** (horizontal) **2** (perpendicular), **3** (vertical)
xlab,ylab=	Text of axes labels: "**…**"
cex.lab=	Size of axes labels: **1,..**
xlim,ylim=	Range of axes: **c(min, max)**
cex.axis=	Size of axes values: **1,..**
log=	Logarithmic scale of **x, y** or **xy**
main=	Text of title: "**…**"
main.cex=	Size of title: **1,..**

points

Argument	Values
pch=	Type of symbols: **0,..,18**, "letters"
cex=	Size of symbols: **1**, ...
col=	Colour: **1**, **2**, **3**, **4**, **5**, **6**, **7**, 8
font=	Font type: 1, **2**, *3*, ***4***



Distribution plots

- to study distribution of a numeric (response) variable
- histogram .. **`hist`**
- stem-and-leaf plot .. **`stem`**
- q-q plots to compare distribution of two variables

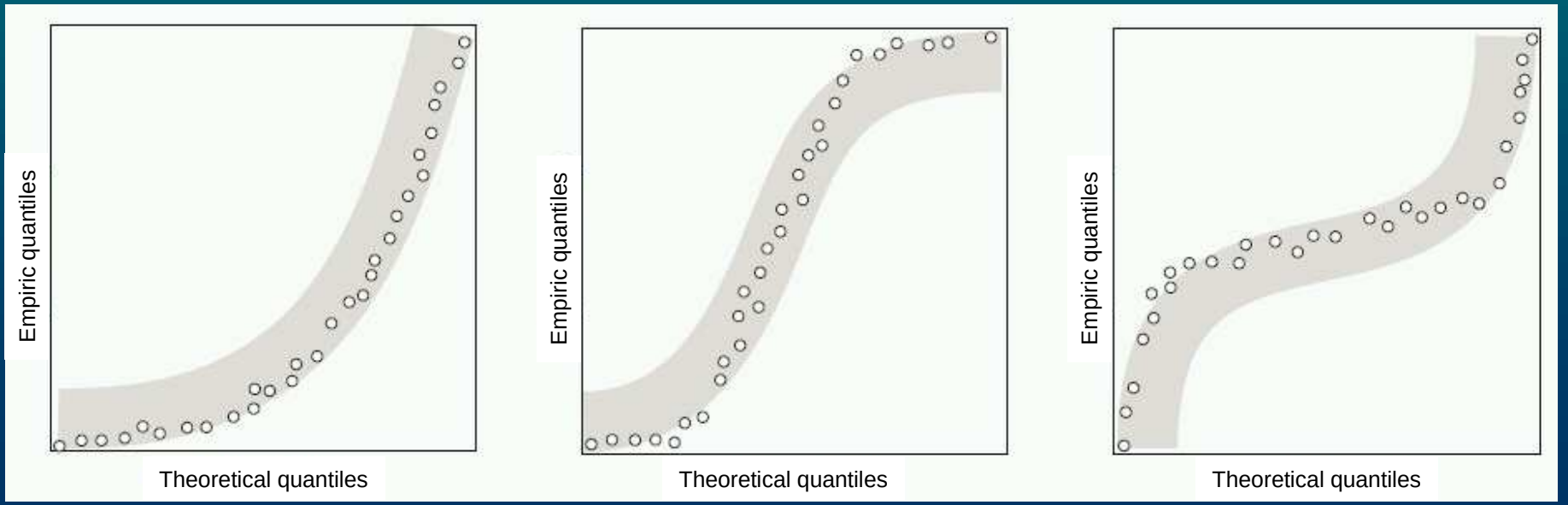
- compare a single variable with normal: **`qqnorm`**
- compare distributions of two variables: **`qqplot`**
- to add diagonal line: **`qqline`**

Example

Make histogram and q-q plot of *distance*.

Deviations from normal distribution

Asymmetric, right-skewed



Symmetric with extended tails

Symmetric with heavy tails

Scatter plots

- for data with continuous explanatory variables
- to produce plots with points: **`plot`**

Example

Make scatterplot of *distance* on *amount* without and with different points for two levels of *SOIL*.

Box plots

• when there are categorical explanatory variables

• function .. **`plot`**
- central line represents median
- box is Q_{25} and Q_{75}
- whiskers are 1.5 times interquartile range
- circles are outliers

• argument **`notch`** for boxes with CI_{95} for median
- if median of one level falls outside notches of another level, there is likely significant difference

Example

Make boxplot of *amount* for *SOIL* without and with notches.

Panel plots

• for data with both categorical and continuous explanatory variables

• **`xyplot`** from library *lattice*
- separate plots for each level of a factor: y ~ x|A
- several types: **`type="r"`** .. regression plot

Example

Make panel scatterplot regression plot of *distance* against *amount* for *SOIL*.

Interaction plot

• for data with two categorical explanatory variables
• to plot means of two factors (*A*, *B*) connected by lines

.. **`interaction.plot`**

- *A* is plotted on axis x
- *B* is in the legend

• visual assessment of interaction between factors *A*B* or *A:B*
- two factors can affect response additively or multiplicatively
- additive effect: parallel lines
- multiplicative effect: crossed lines

Example

Make interaction plot of *SOIL* and *FIELD* for *amount*.

Bar plot

- when data are counts or proportions
- data are arranged in a matrix or table
- **`barplot: beside, legend`**

Example

Make barplot of *SOIL* and *FIELD* for *amount*.

Paired plots

- when data include several continuous explanatory variables
- **`pairs`** produces matrix of all possible plots

3-dimensional plots

- when data include 2 continuous explanatory variables
- **`wireframe`** (*lattice*) produces 3-dimensional plot

Graphs with error bars

- to display error bars use vertical lines or *sciplot* package
- plot empirical means and errors
- **`bargraph.CI`**
- **`lineplot.CI`**

Example

Make barplot of *SOIL* and *amount* and line plot of *SOIL* and *FIELD* and *amount* .

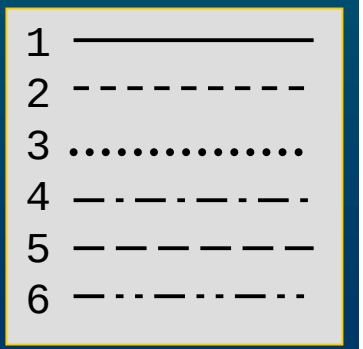
Graphs with functions

- final plot of estimated models

- **`lines`** connects points specified by coordinates
- **`abline`** produces line specified by intercept and slope

`lines`

Argument	Values
`x,y=`	Coordinates: `c( ..,..)`
`lty=`	Line type: `1,...,6`
`col=`	Colour: `1, 2, 3, 4, 5, 6, 7, 8`
`lwd=`	Width: `1, ..`



Example

Make lineplots for the following models:

inverse $y = \frac{1}{x}$

exponential $y = e^{x}$

logarithmic $y = \log(x)$

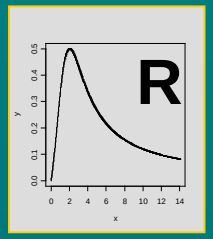
power $y = x^{3}$

logistic $y = \frac{1}{1+9e^{-0.3x}}$

squareroot $y = \sqrt{x}$

quadratic $y = 0.6 - 0.1x + 0.006x^{2}$

inverse squareroot $y = \frac{1}{\sqrt{x}}$



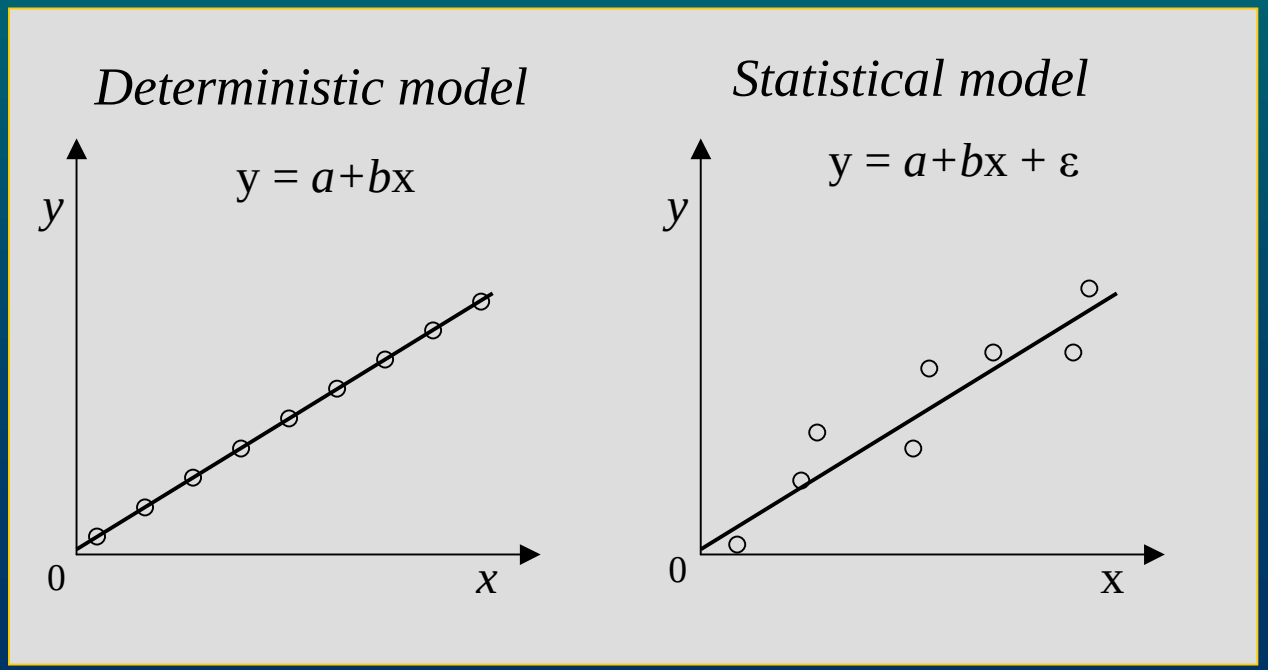
Statistical Modelling

Stano Pekár

Regression model

- includes systematic and stochastic components

$$y_i = \alpha + \beta x_i + \varepsilon_i$$



• assumptions of the stochastic component:

$$\varepsilon_i \sim N(0, \sigma^2) \qquad \mathrm{cor}(\varepsilon_i, \varepsilon_{i'}) = 0, i \neq i'$$

= variance is equal = **homoscedastic** model

To find real model we need to estimate its parameters: α, β, σ^2

as *a*, *b*, s^2 so that we get $\hat{y}(x_0) = a + bx_0$

General Linear Model

- extension of the systematic component

Simple regression

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\beta = 0$$

$$y_i = \alpha + \varepsilon_i$$

1-way ANOVA

$$y_{ij} = \alpha + \beta A_j + \varepsilon_{ij}$$

$$\beta = 0$$

$$y_i = \alpha + \varepsilon_i$$

Linear model (LM) has a general form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k + \varepsilon$$

linear predictor

x can include: u^2, $u^{1/2}$, log(*u*), exp(*u*), sin(*u*), factors

= model is linear in parameters when it includes only linear combinations of parameters

Some nonlinear relationships can be linearised

- log-transformation of both sides:

$$y = e^{a+bx_i} + e^{\varepsilon} \rightarrow \log(y) = a + bx + \varepsilon$$

$$z = \log(y) \rightarrow z = a + bx + \varepsilon$$

- e^{ε} has lognormal distribution while ε has normal distribution
- *y* has heterogenous variance *z* has homogenous variance
- e^{ε} is multiplicative while ε is additive
- curved relationship becomes linear

Other nonlinear relationships can not be linearised

$$y = \alpha(1 - \beta e^{-\gamma x})$$

use **Nonlinear regression**

Generalised Linear Model

- extension of the stochastic component
- we model transformed expected value of *y*

$$f(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k$$

$$y \sim distribution$$

f(μ) .. link function

For example,

$$\mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k$$

$$y \sim N(\mu, \sigma^2)$$

$$\varepsilon = y - \mu \sim N(0, \sigma^2)$$

GLM has 3 components:

• link function
• linear predictor
• distribution family
- Gaussian (normal), Gamma, Inverse Gaussian, Poisson, Quasipoisson, Binomial, Quasibinomial, Quasi

• measure of fit is deviance not sum of squares
- null deviance = SST
- residual deviance = SSE
- ANODEV table = ANOVA table

Good model

• a useful simplification of the reality
- should include important aspects for which it is being made and ignore aspects that we are not interested in
- like a good map

• **Principle of parsimony:** Simpler model is better if it explains study phenomenon as good as complicated model.

G. E. P. Box: „All models are wrong. But some of them are useful.“

Modelling procedure

Bottom -up or forward selection

• building up a model by adding available variables

Top-down or backward selection

• reducing maximal (saturated) model

1. Fit maximal model- all main effects and interactions
2. Remove insignificant interactions and main effects
3. Group together similar factor levels
4. Check diagnostic plots
5. Alter model if necessary
6. Achieve minimal adequate model

- contains only terms in which all parameters are significantly different

Model criticism

• to assess model quality and assumptions
- study of both systematic and stochastic components
- we can never prove that model is adequate

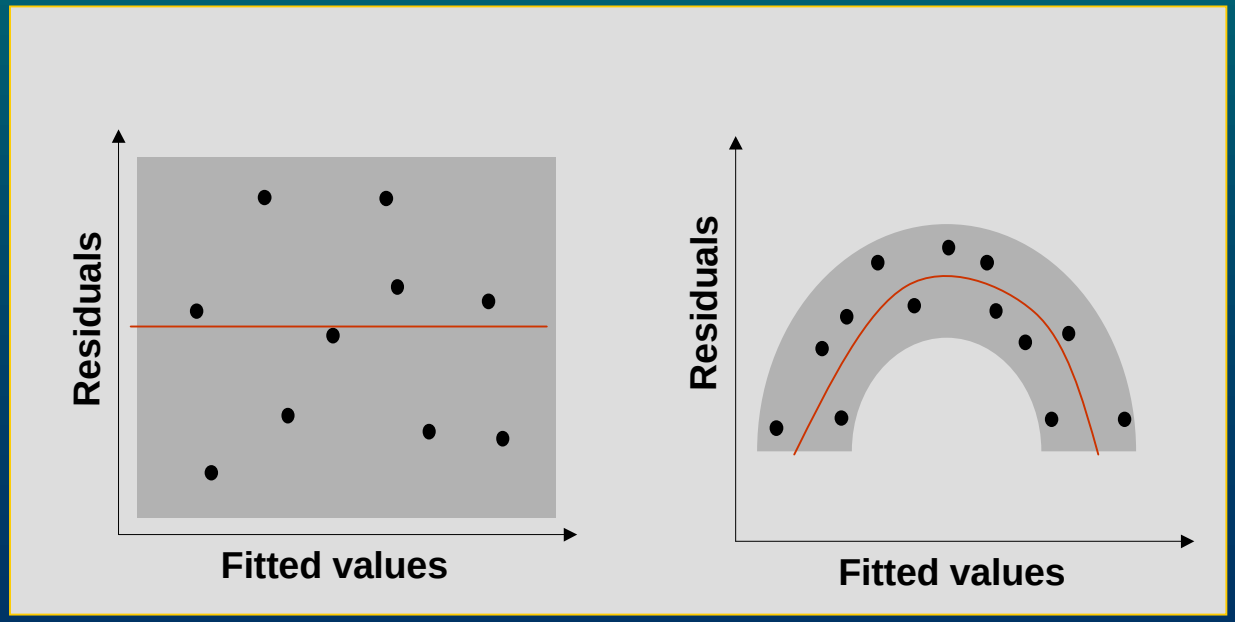
Residuals $\varepsilon_i \sim N(0, \sigma^2)$ $\text{cor}(\varepsilon_i, \varepsilon_{i'}) = 0, i \neq i'$

should not
• make trends when plotted against explanatory or response variables
• be heteroscdeastic
• have unusual distribution
• be interdependent

Checking assumptions
• informal using plots - **`plot`** produces 6 plots
• formal using tests

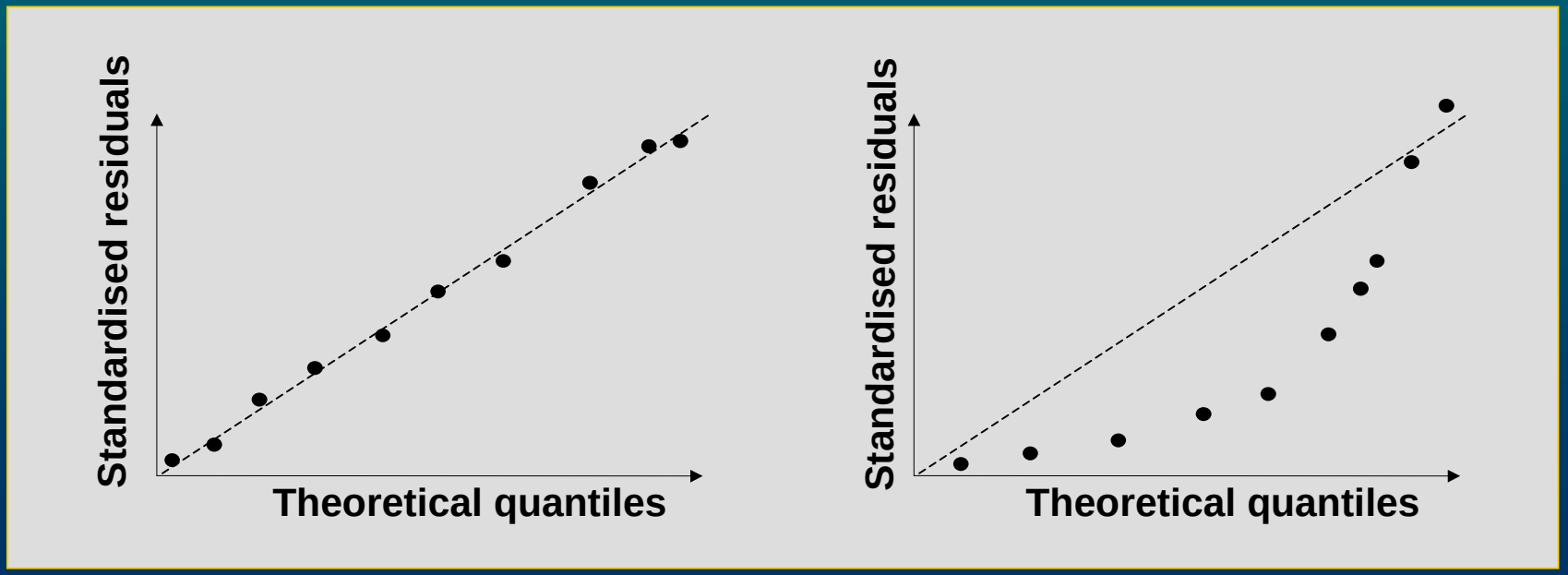
Predictor's adequacy

- raw (LM) or deviance (GLM) residuals against fitted values
- curved pattern suggests lack of polynomial term



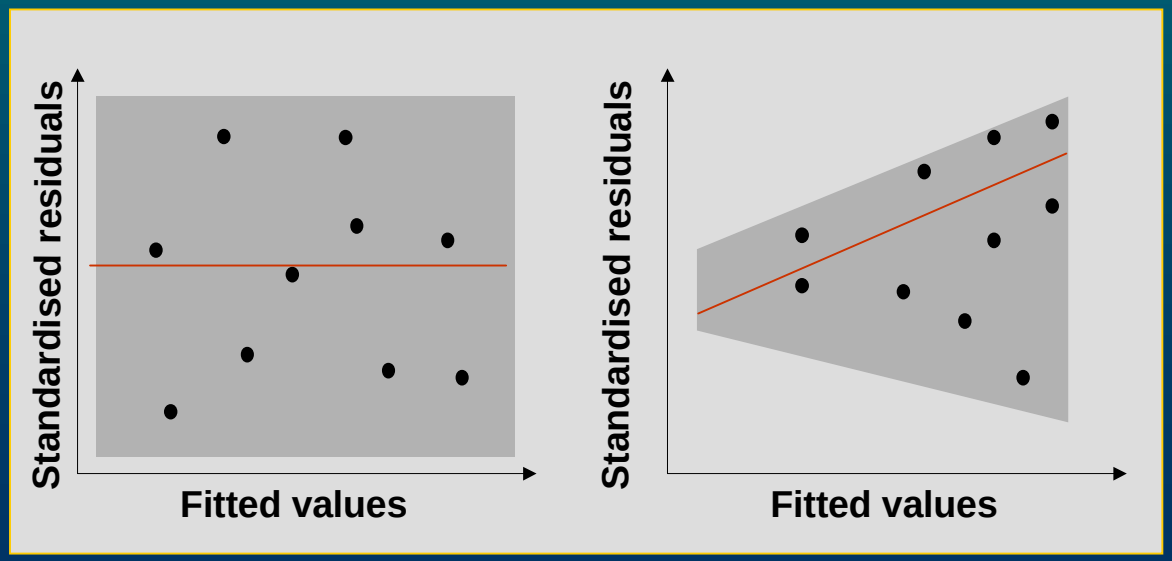
Normality

- q-q plot of standardised (LM) standardised deviance (GLM) residuals
- data from other than normal distribution can not have normally distributed residuals
- when the pattern is “J” or “S” shaped change link function or transform the variable



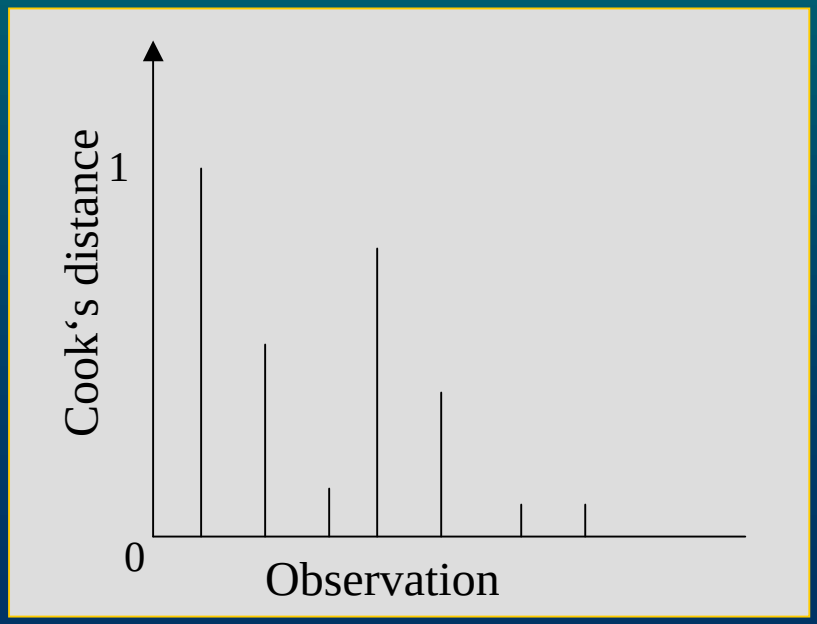
Variance homogeneity

• plot of standardised (LM) standardised deviance (GLM) residuals against fitted/predicted values
• when variance increases with the mean use Poisson or gamma distribution or log transformation



Influence

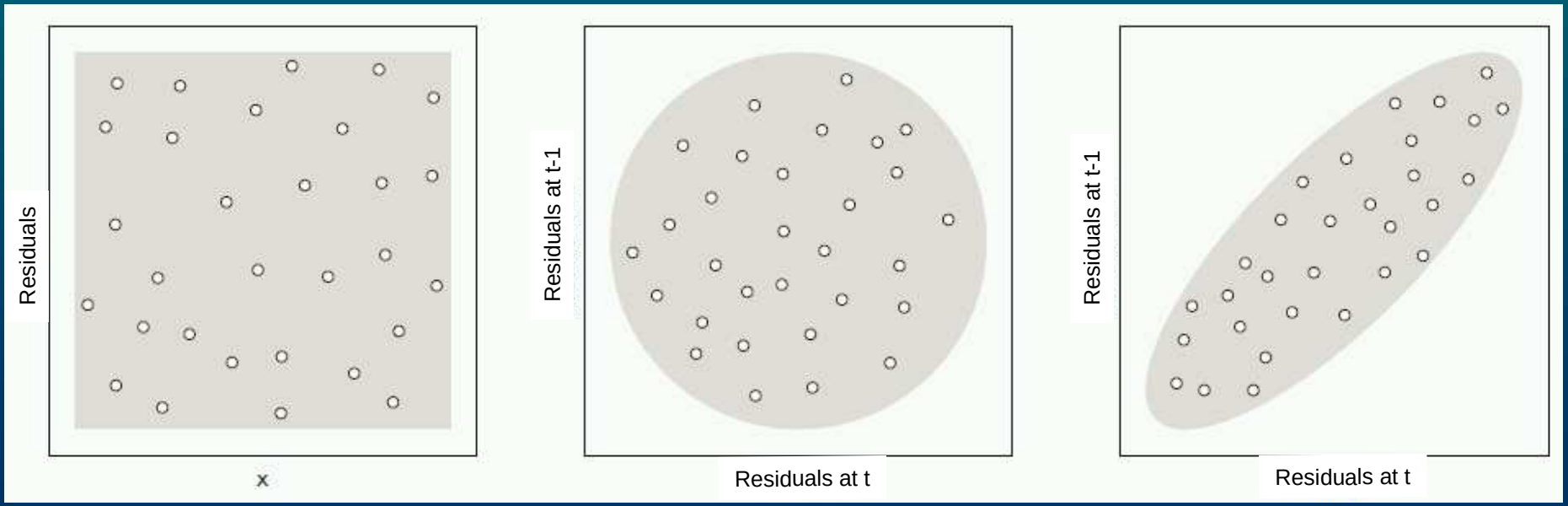
• plot of Cook's distance for each observation shows the influence of individual observations on the model fit
• values of influential observations are close to 1 and higher
• check for errors in the data
• omit influential observations or transform the explanatory variables (using log, power, reciprocal)



Independence

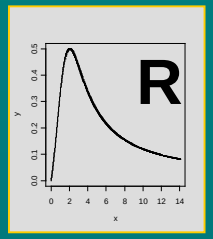
- dependence on continual explanatory variable
- using standardised (LM) or Pearson residues (GLM)
- serial dependence if explanatory variable is time or space

Independence on x



Serial independence

Serial dependence



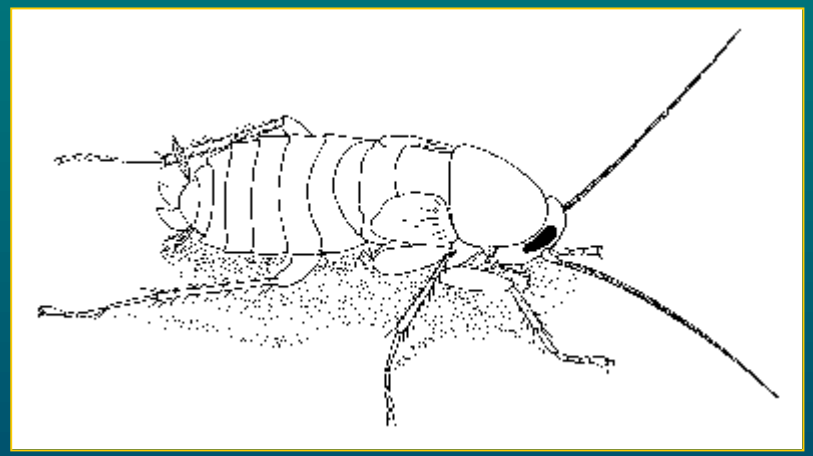
The first trial

Stano Pekár

2-way ANCOVA

Background

Nutritional quality of the diet affects growth of organisms in various ways. To find optimal diet for cockroaches the following experiments was performed.



Design

Effect of five diet types (control, lipid1, lipid2, protein1, protein2) was tested on body weight [g] of male and female cockroaches. For each diet 10 females and 7 males were used. Their body weight [g] was recorded before and after the experiment.

Hypotheses

Is weight influenced by the diet type?

If so which diet resulted in largest weight?

Is weight on diets similar for males and females?

Variables

DIET: control, lipid1, lipid2, protein1, protein2

SEX: male, female

start

weight

Data

`cockroach.txt`

ANOVA Table

• **anova** uses Type I Sum of Squares
- sequential assessment of effects according to the given order
- at first main effects are assessed then interactions
- in orthogonal the order is not important
- if data are unortogonal it is more appropriate to use Type III SS

Ortogonality
• independent variables are orthogonal - effects are straightforward
• correlated variables are unorthogonal - effects are complicated
- when there are missing values or unequal number of observations *per* treatment

Quadratic term

• check for curvature by fitting a separate quadratic term for continuous explanatory variables

$$y = \alpha + \beta x + \gamma x^2 + \varepsilon$$

• quadratic model - a simple description of nonmonotonous trend
• use either **`poly(x,2)`** or **`x + I(x^2)`**

Removing terms

• remove insignificant interactions

- begin with the higher order terms because main effects are marginal to interactions

- intercept is marginal to slope and both are marginal to the quadratic term

• remove insignificant main effects

Criteria

• test (F or χ^2) and a given p-value (**anova**)

• **Akaike Information Criterion** (AIC):

$$\mathrm{AIC} = -2LogLik + 2p$$

- the more there are parameters in the model the better fit but worse explanatory power of the model

- the lower AIC the better model

Comparisons

- compare individual differences between factor levels
- comparisons are valid only if a factor is significant

Options:

- *Apriori* contrasts (before analysis)
- *Posteriori* simplification (after analysis)
- Multiple comparisons (after analysis)

- *apriori* contrasts are preferred to avoid excess of significant results

Contrasts

For a model $y_{ij} = A_j + \varepsilon_{ij}$

a contrast will be $K = \sum_{j=1}^{J} w_j A_j$

where A_j .. mean value of a level, w_j .. contrast coefficient

Creating contrasts
- levels lumped together get the same sign
- levels contrasted get opposite sign
- levels excluded get 0

.. so that sum of each contrast $\sum_{j=1}^{J} w_j = 0$

Contrasts are arranged in a matrix

• only *k*-1 (*k* .. number of levels) contrasts are orthogonal, i.e. each level (combination) is compared only once

... products of any two contrasts = 0

• specified by function **`contrasts`** prior to analysis

Pre-specified contrasts:

• **Treatment** (default in R) - compare specific level with the reference level

• **Helmert** - compare specific level with the average of previous levels

• **Sum** - compare specific level with the grand mean

• **Textbook** - compare each level with 0

Simplification

- levels of a factor are compared using Wald statistics from ouput
- similar factor levels are the grouped together
- test each grouping by **anova**
- compare the final model with the first one

Diagnosis

We should check as many aspects as possible

- use diagnostic plots
- use formal tests:

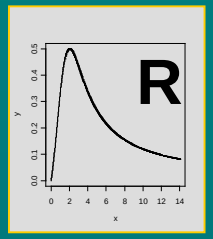
- Bartlett test to compare variances
- Shapiro-Wilk test of normality

Analysis

```
dat<-read.delim("cockroach.txt"); attach(dat); names(dat)
plot(diet,weight)
interaction.plot(diet,sex,weight)
library(lattice)
xyplot(weight~start|diet,groups=sex,pch=1:2)
m1<-lm(weight~diet*sex*start)
anova(m1)
m2<-lm(weight~diet*sex*poly(start,2))
anova(m1,m2)
m3<-update(m1,~.-diet:sex:start)
anova(m1,m3)
anova(m3)
m4<-update(m3,~.-diet:start)
anova(m4)
m5<-update(m4,~.-sex:start)
anova(m5)
m6<-update(m5,~.-diet:sex)
anova(m6)
m7<-update(m6,~.-start)
anova(m7)
m8<-update(m7,~.-sex)
anova(m8)
```

```
summary(m8)
levels(diet)
contrasts(diet)<-cbind(c(1,-1/4,-1/4,-1/4,-1/4),c(0,-1/2,-1/2,1/2,1/2),
c(0,0,0,1/2,-1/2),c(0,-1/2,1/2,0,0))
contrasts(diet)
summary(lm(weight~diet))
contrasts(diet)<-'contr.helmert'
summary(lm(weight~diet))
contrasts(diet)<-'contr.sum'
summary(lm(weight~diet))
diet1<-diet
levels(diet1)
levels(diet1)[4:5]<-"prot"
levels(diet1)
contrasts(diet1)<-'contr.treatment'
m9<-lm(weight~diet1)
anova(m8,m9)
diet2<-diet1
levels(diet2)[2:3]<-"lipid"
m10<-lm(weight~diet2)
anova(m9,m10)
summary(m10)
diet3<-diet2
levels(diet3)[2:3]<-"other"
m11<-lm(weight~diet3)
```

```
anova(m10,m11)
anova(m10,m1)
plot(m10,which=1:4)
shapiro.test(resid(m10))
library(sciplot)
lineplot.CI(diet2,weight,ylab="Weight",xlab="Diet")
```



Systematic componet

Stano Pekár

Analytical methods

$$y_i = a + bx_i + \varepsilon_i$$

• the same explanatory variable can be taken once as continuous other time as categorical: e.g. two levels of concentration
• continuous variable allows interpolation and extrapolation

Key to methods:

Explanatory variable(s)	**Method**
Continuous	Regression
Categorical	ANOVA
Continuous and categorical	ANCOVA

Linear predictor can include various terms:

- intercept .. α estimated as a
- linear term .. βx with b as coefficient of linear trend
- quadratic term .. γx^2 with c as coefficient of quadratic trend
- cubic term .. τx^3 with t as coefficient of cubic trend
- main effect .. A
- interaction between factors .. $A{:}B$
- interaction between continuous variables $x_1{:}x_2$
- linear interaction .. $A{:}x$
- quadratic interaction .. $A{:}x^2$

Regression

- **simple regression** ... 1 explanatory variable
- **multiple regression** .. 2 and more explanatory variabels

General linear predictor of multiple regression

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k$$

α .. intercept
β_k .. linear coefficients of x_k
x .. may represent polynomic functions (x^3), interactions ($x_1.x_2$)

- rule of thumb: less than $n/3$ parameters in model at any time
- number of combinations of explanatory variables will often exceed the number of data so we can not include all terms

Simplification

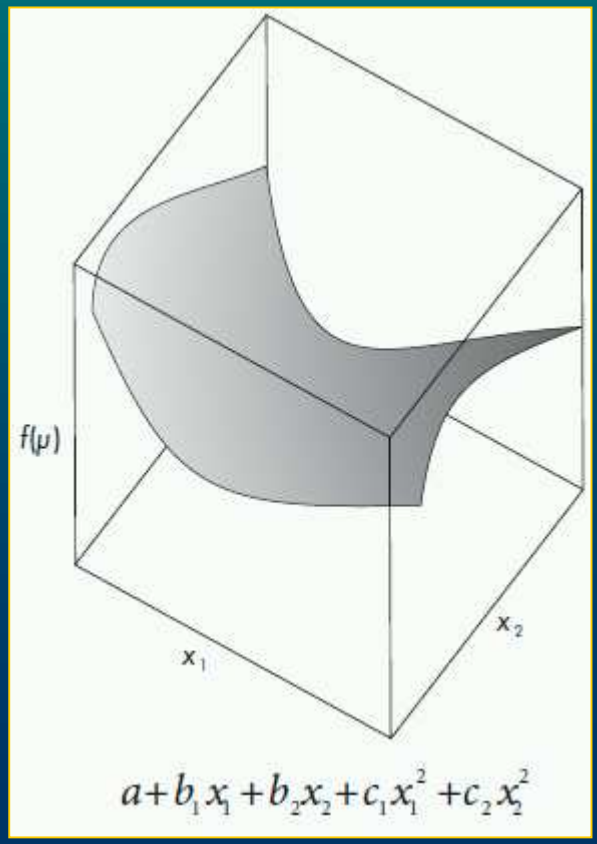
• linear predictor with 2 explanatory variables (x_1, x_2) should include all main effects, all interactions, and quadratic terms

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 x_1 + \gamma_2 x_2 + \delta x_1 x_2$$

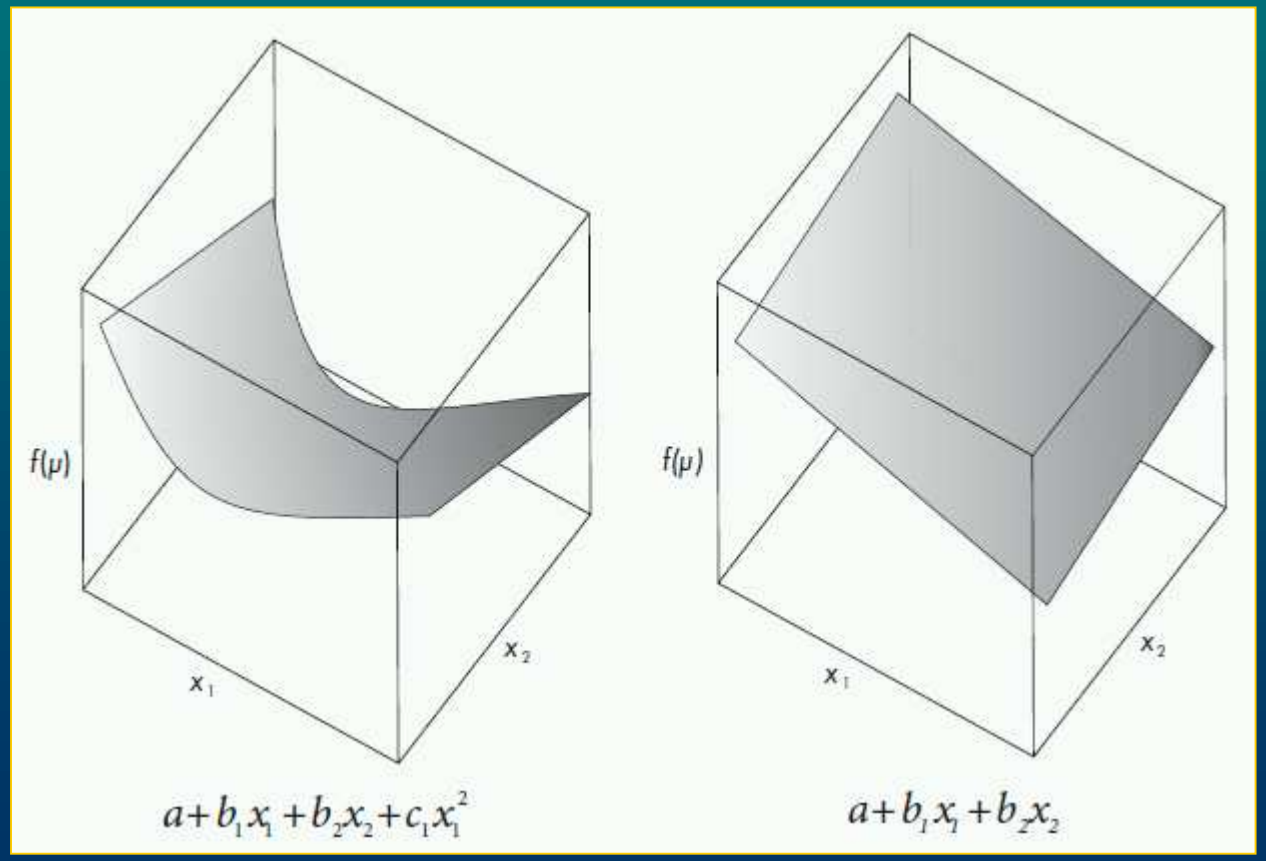
with estimates a, b_1, b_2, c_1,c_2, d

Nested models are:

• 5 parameters (a, b_1, b_2, c_1, c_2), at least c_1 and c_2 are significantly different

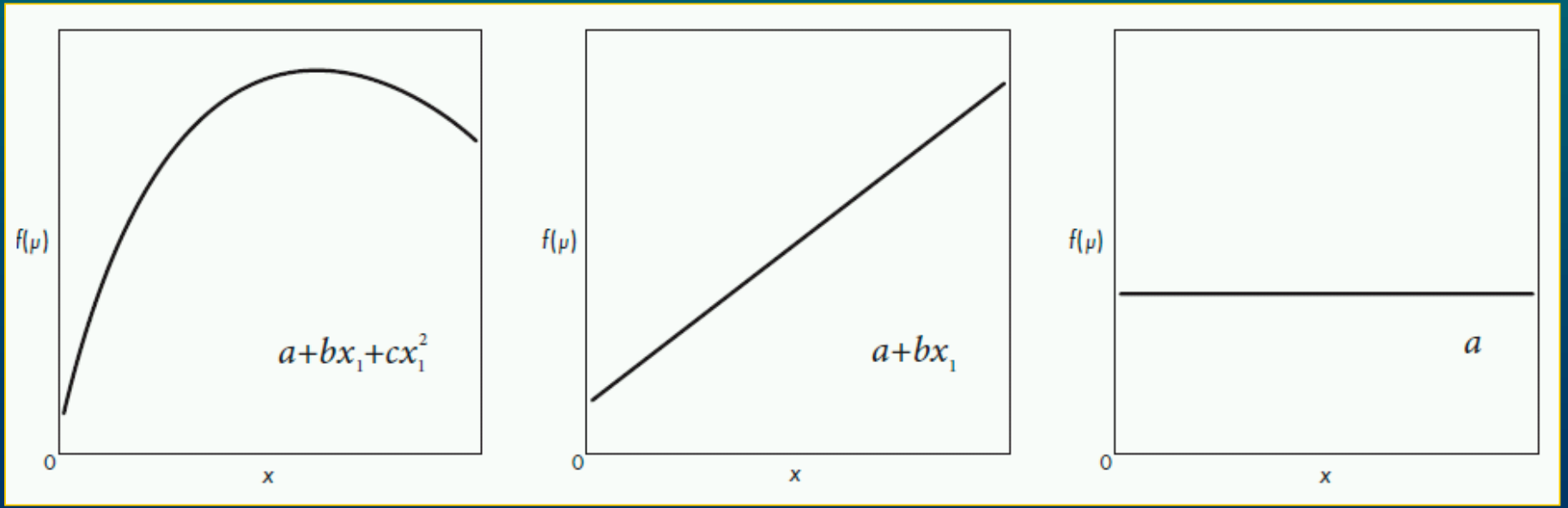


- 4 parameters (a, b_1, b_2, c_1), at least c_1 is significantly different
- 3 parameters (a, b_1, b_2), at least b_1 and b_2 are significantly different



If one explanatory variable (x_2) turns out to be insignificant:

- 3 parameters (a, b, c), at least c is significantly different
- 2 parameters (a, b), at least b is significantly different
- 1 parameter (a) that is significantly different



ANOVA

- 1-way ANOVA .. 1 factor
- 2-way ANOVA .. 2 factors
- k-way ANOVA .. k factors

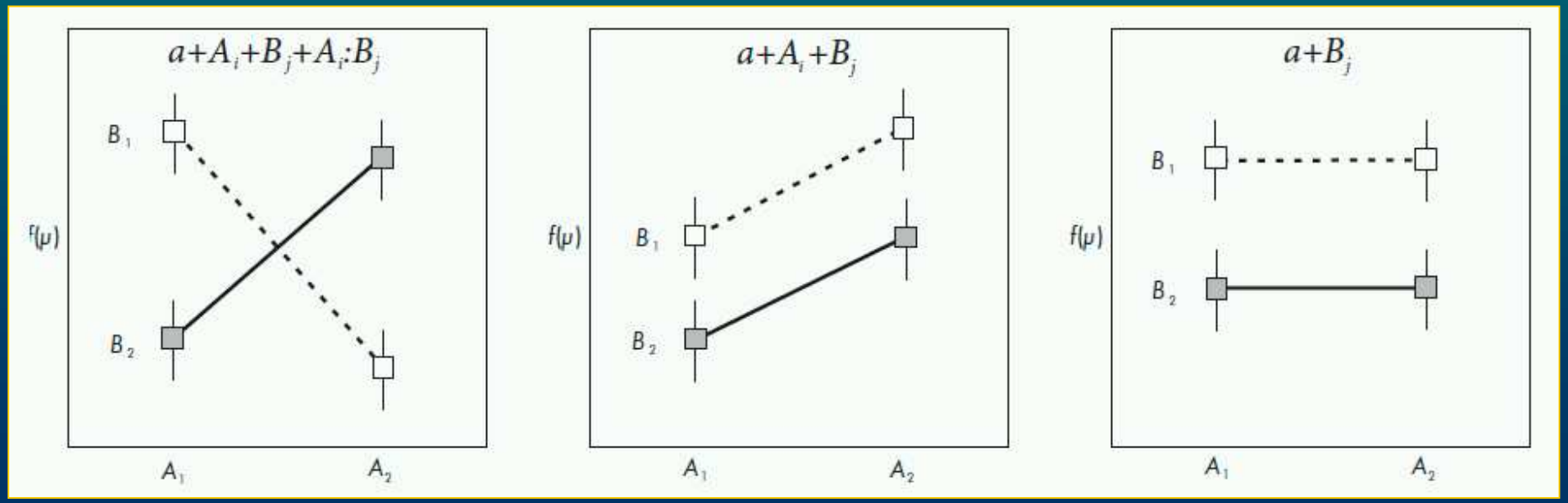
- k-way ANOVA might be with our without interactions

Given 2 categorical variables A and B each with 2 levels (A_1, A_2, and B_1, B_2) model with treatment contrasts is

$$\alpha + A_i + B_j + A:B_{ij}$$

α .. mean of A_1B_1, A_i and B_j .. main effects, A:B_{ij} .. interaction

• 4 parameters (A_1B_1, $A_2B_1–A_1B_1$, $A_1B_2–A_1B_1$ a $A_2B_2–A_1B_2$): interaction is significant
• 3 parameters (A_1B_1, $A_2B_1–A_1B_1$, $B_2–B_1$): only A and B are significant
• 2 parameters (B_1, $B_2–B_1$): only B is significant
• 1 parameter (grand mean): null model



ANCOVA

- combination of regression and ANOVA
- continuous variable = covariate

Given 1 factor (A_j) and 1 covariate (x) linear predictor is:

$$\alpha + A_j + \beta x + \delta_j x$$

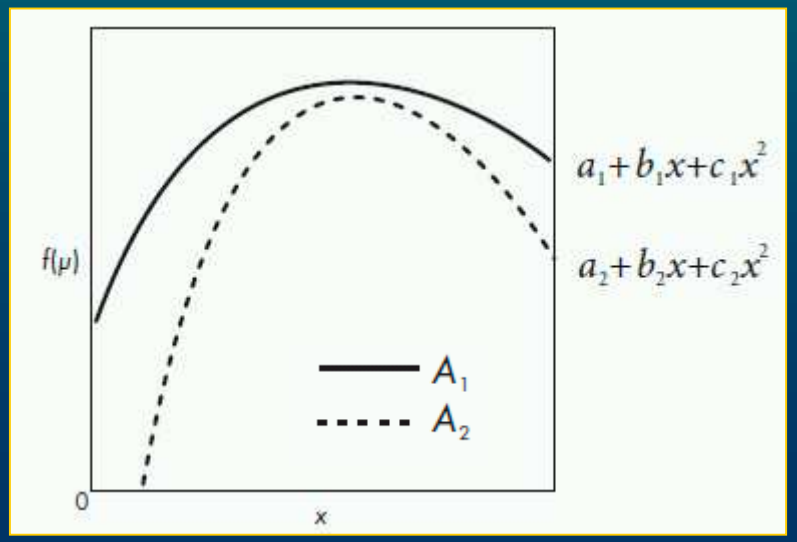
α .. intercept, A_j .. effect of factor, β .. slope, δ .. effect of interaction

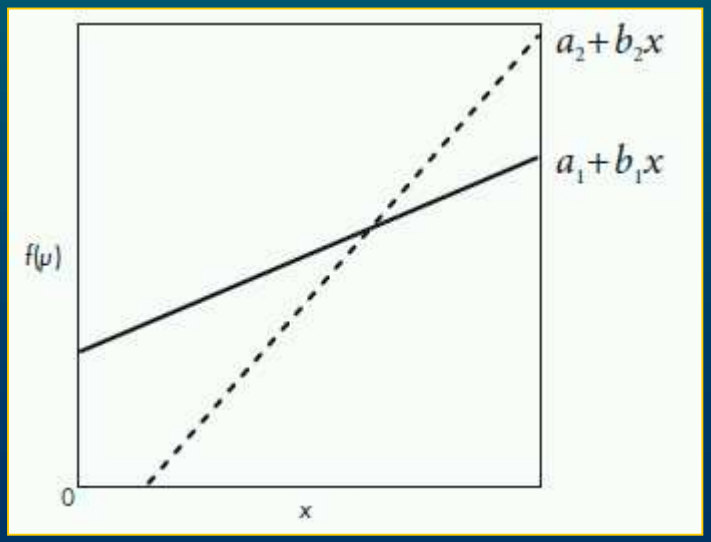
Given 1 categorical variable A with 2 levels (A_1, A_2) and 1 continual x, the linear predictor will be

$$\alpha + A_j + \beta x + \delta_j x + \gamma x^2 + \omega_j x^2$$

• 6 parameters - 2 intercepts (a_1, a_2–a_1), 2 slopes (b_1, b_2–b_1), 3 quadratic (c_1, c_2–c_1) - interaction A:x^2 is significant

• 4 parameters - 2 intercepts (a_1, a_2–a_1), 2 slopes (b_1, b_2–b_1) - interaction A:x is significant, but quadratic terms are not significant

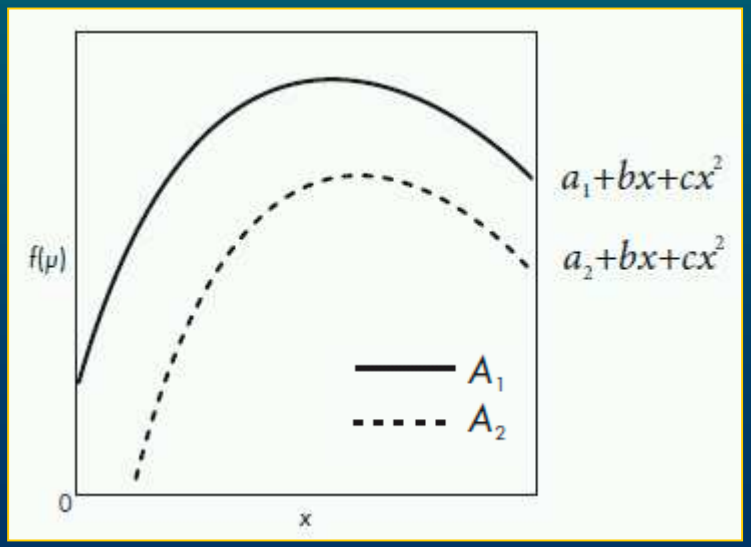


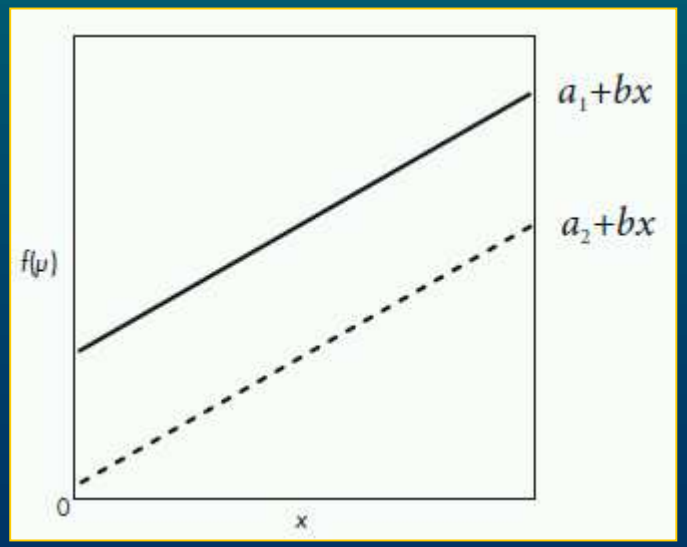


• 4 parameters - 2 intercepts (a_1, a_2–a_1), 1 slope (b), 1 quadratic (c) - interactions A:x^2 and A:x are not significant, but A and quadratic terms are significant

• 3 parameters - 2 intercepts (a_1, a_2–a_1), 1 slope (b) - only main effects (A and x) are significant

• Further simplification → 1-way ANOVA or simple regression





Model formulae

response variable ~ *explanatory variable*(s)

• Operators:
- on left side any mathematical operator can be used
- on the right side only few:

+ .. add
- .. delete
: .. interaction
***** .. all terms
1 .. intercept
I .. interpreter that translates operators into mathematical meaning
/ .. nested
| .. conditioned

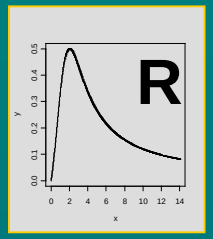
Model formula	Description	
`y ~ 1`	Null model	$f(\mu_i) = \alpha$
`y ~ x`	Linear model with 1 explanatory variable	$f(\mu_i) = \alpha + \beta x_i$
`log(y) ~ x -1`	Linear model with 1 explanatory variable, without intercept and with log-transformed response	$\log(\mu_i) = \beta x_i$
`y ~ x + I(x^2)` `y ~ poly(x,2)`	Quadratic model with 1 explanatory variable	$f(\mu_i) = \alpha + \beta x_i + \gamma x_i^2$
`y ~ x1 + x2`	Linear model with 2 explanatory variables	$f(\mu_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i}$

Model formula	Description
`y ~ A*B*C`	3-way ANOVA with three main effects, two 2-way interactions and one 3-way interaction
`y ~ (A+B+C)^2`	3-way ANOVA with only three 2-way interactions
`y ~ x*A`	1-way ANCOVA

$$f(\mu_{ijk}) = \alpha + A_i + B_j + C_k + A:B_{ij} + A:C_{ik} + B:C_{jk} + A:B:C_{ijk}$$

$$f(\mu_{ijk}) = \alpha + A_i + B_j + C_k + A:B_{ij} + A:C_{ik} + B:C_{jk}$$

$$f(\mu_{ij}) = \alpha + A_j + \beta x_i + \delta_j x_i$$



Stochastic componet

Stano Pekár

$$y_i = a + bx_i + \varepsilon_i$$

- choose distribution if using GLM
- there are many distributions but only some are available for GLM
- decision should be based upon theoretical models or previous experience

Response variable can be

- continuous measurements
- counts
- proportions

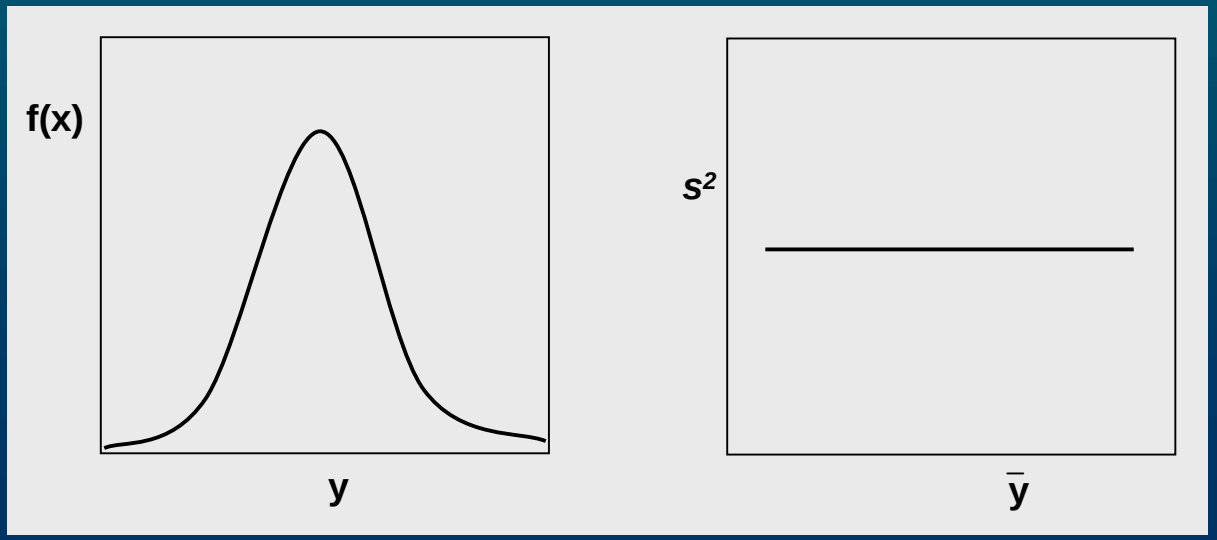
Continuous measurements

- measurements that can be made with infinite precision

Gauss (normal) distribution

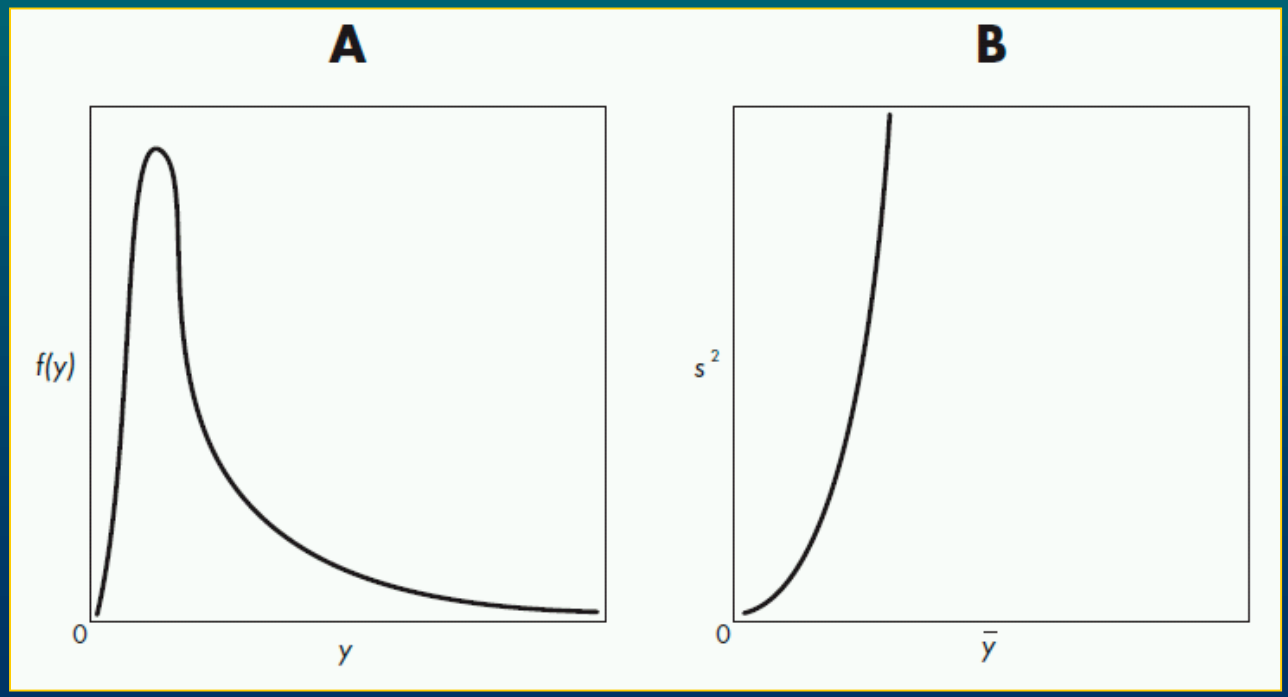
- bell-shaped, symmetric around mean
- mean = median = modus
- parameters: μ, σ^2

- s^2 is independent of mean



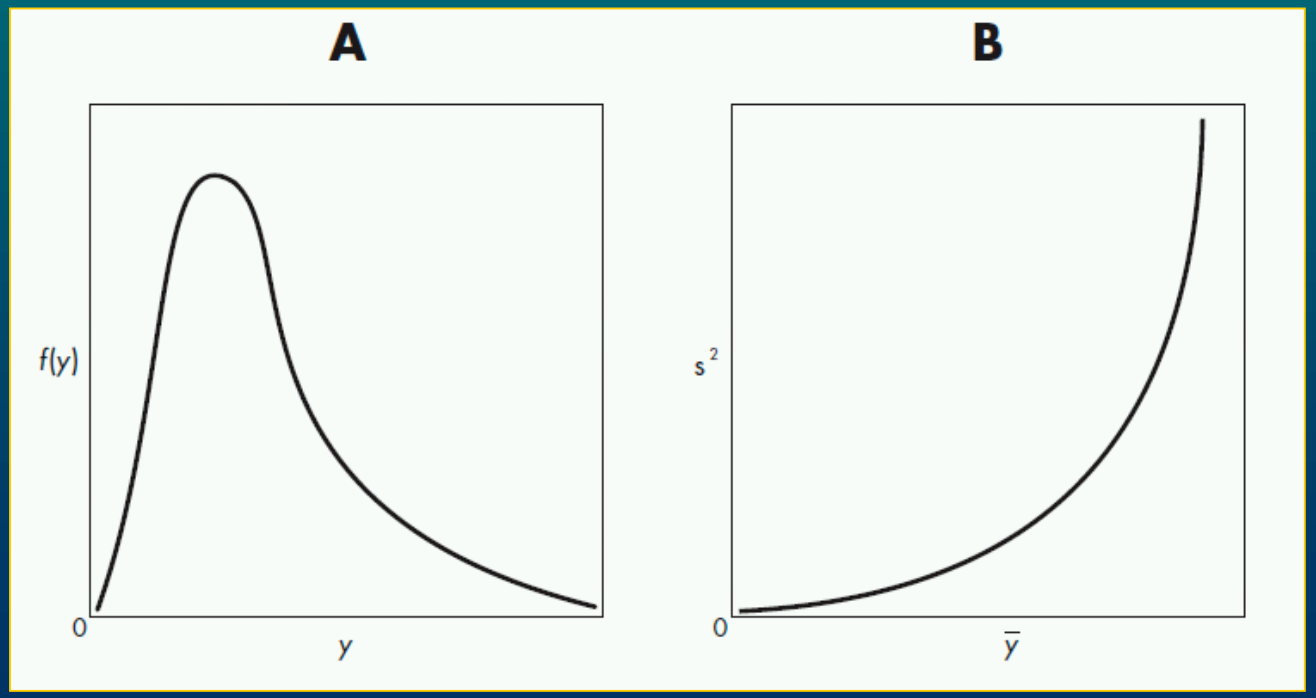
Lognormal distribution

- discrete values, made of integers
- asymmetric, skewed to the right
- variance increases with mean at quadratic trend
- after logarithmic transformation variances are similar



Gamma distribution

- positive real values
- asymmetric, skewed to the right
- variance increases with mean at a quadratic trend



Other distributions

- Inverse Gaussian distribution

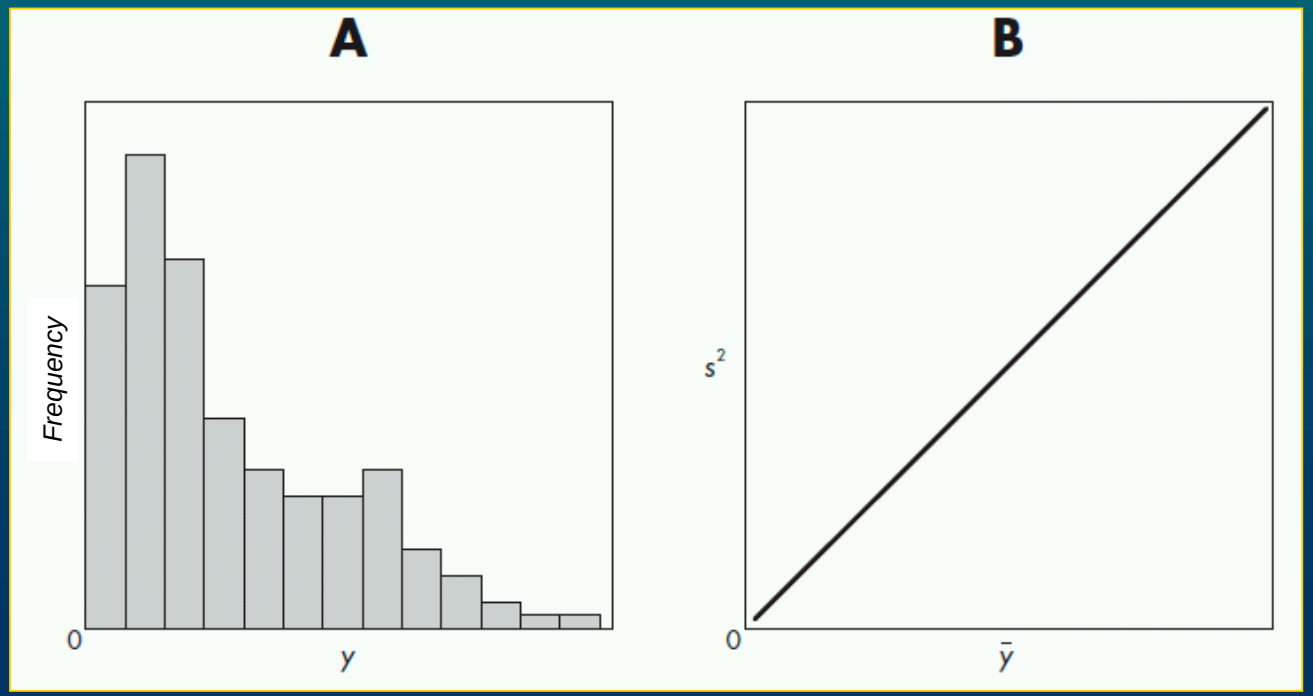
- used to model diffusion processes
- variance increases steeply with mean

Counts

Poisson distribution

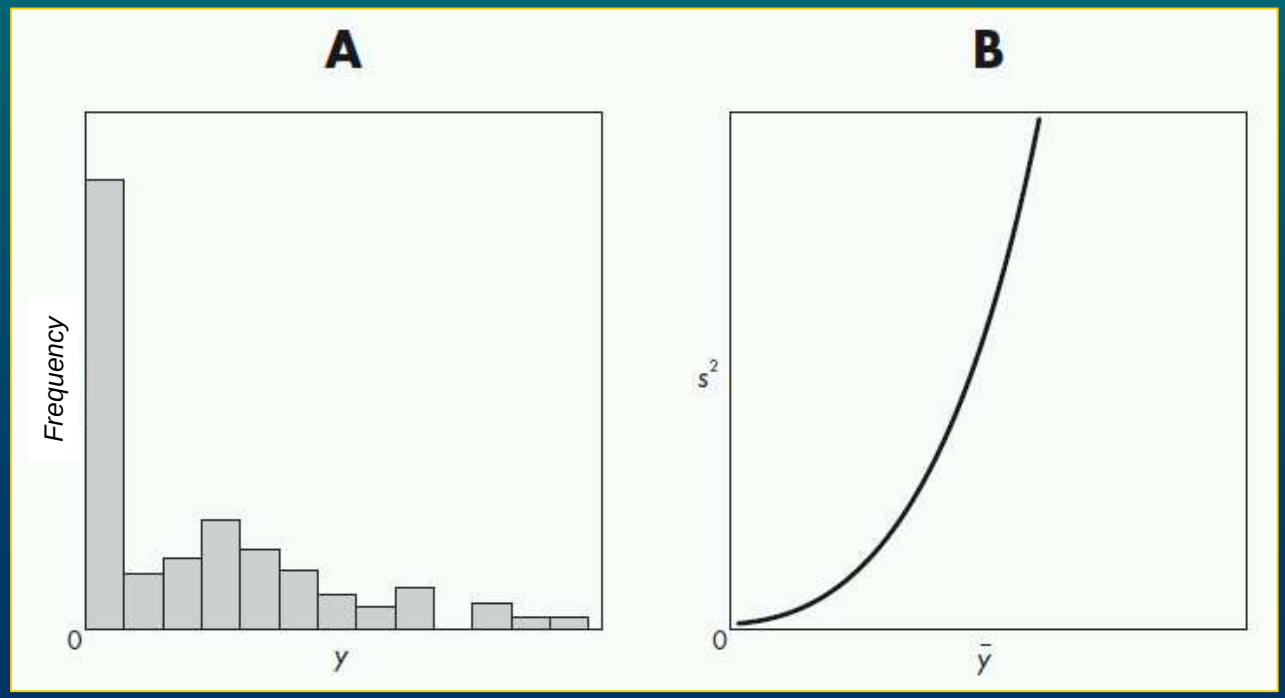
- discrete values, made of integers
- asymmetric, skewed to the right
- variance is equal to expected value

- variance increases with mean



Negative-binomial distribution

- discrete values, made of integers
- asymmetric, strongly skewed to the right
- variance is larger than expected value
- variance increases with mean at a parabolic trend



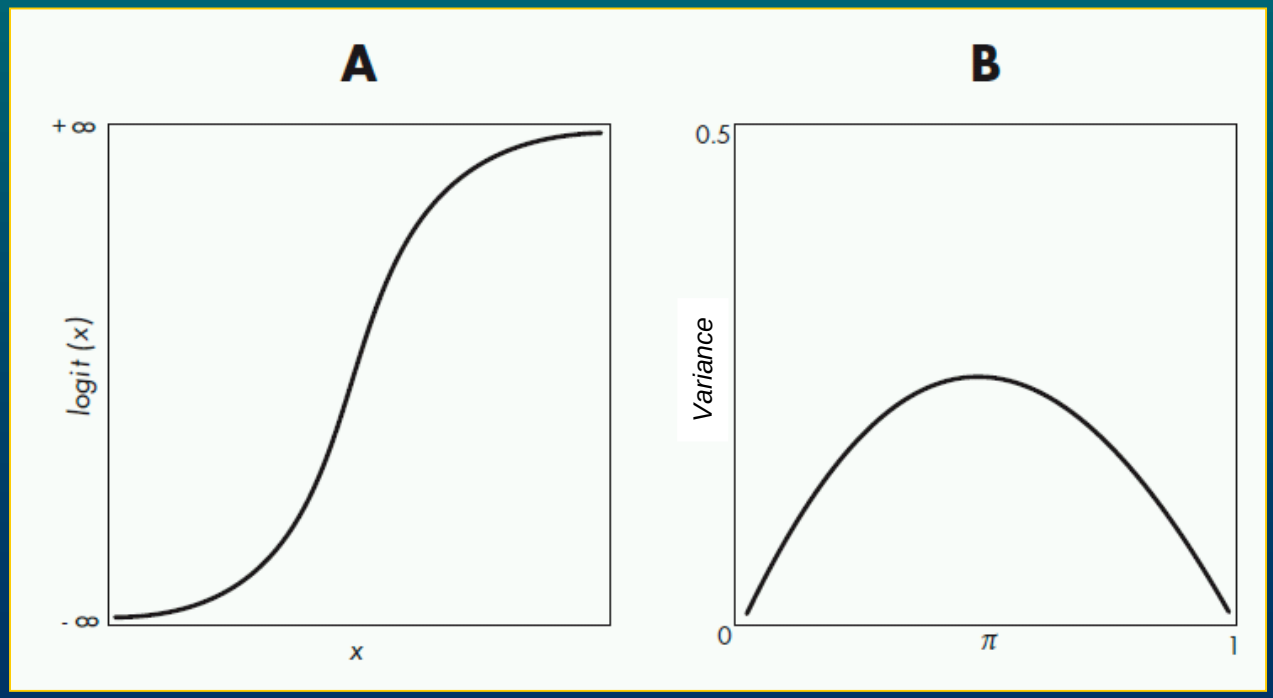
Proportions

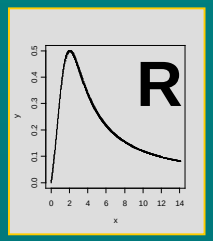
- arise when we counts events (y) from a whole population (n)
- p .. relative frequency = y/n
- we study only qualitative character of an event not its quantitative aspect
- p is an estimate of a theoretical value π
- based on logit transformation

$$\log\left(\frac{p}{1-p}\right)$$

Binomial & Binary distributions

- measurements (y) are integers of n independent trials
- π .. a single parameter showing probability of event occurrence
- $0 \leq \pi \leq 1$
- variance of π is maximal at 0.5





Analyses of continuous \

Stano Pekár

Gaussian (normal) distribution

- response variable is continuous
- measurements of length, width, distance, concentration, pH, etc.
- data are real numbers
- distribution is symmetric $(-\infty, +\infty)$
- parameters: μ, σ^2 independent of each other

Analytical methods

- **t-test** (`t.test`) to compare one or two means
- **Linear model** (`lm`) to study effect of categorical and continuous variables
  - inference is exact, reliable for each *n*
- **GLM** (`glm`) to study effect of categorical and continuous variables
  - Gaussian family (default)
  - link: identity
  - inference is asymptotic, valid only for large *n*

```
glm(formula, family=Gaussian)
```

Simple Regression

Background

The number of grains in ears affects the yield of cereals.

Design

On 20 plots mean number of seeds per oat ear was estimated.
Then at harvest the yield [t/ha] for each plot was estimated.

Hypotheses

Is number of seeds related to the yield?

What is the predictive model of this relationship?

Variables

grain

yield

Data

`oat.txt`

Analysis

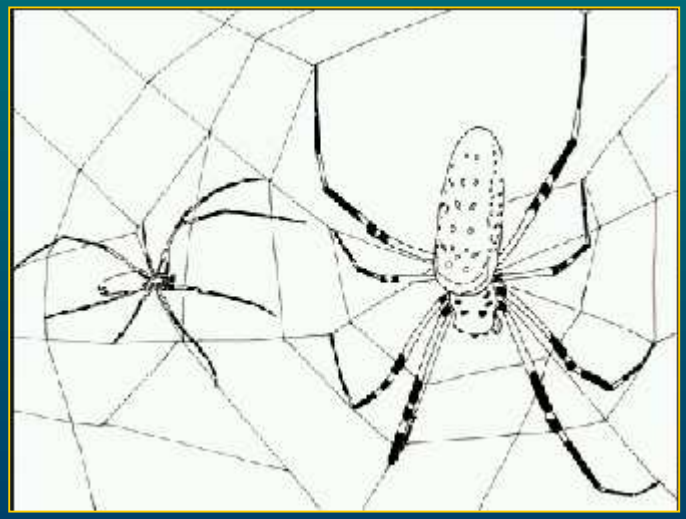
```
dat<-read.delim("oat.txt"); attach(dat); names(dat)
plot(grain,yield)
m1<-lm(yield~poly(grain,2))
summary(m1)
m2<-lm(yield~grain)
summary(m2)
m3<-update(m2,~.-1)
summary(m3)
AIC(m2,m3)
plot(grain,yield,xlim=c(0,30),ylim=c(0,6))
abline(m2)
abline(m3,lty=2)
legend(18,3,c("m2","m3"),lty=1:2)
plot(m2,which=1)
sr<-rstandard(m2); plot(grain,sr)
0.07617+c(-0.0111,0.0111)*qt(0.975,19)
2*(1-pt((0.1-0.07617)/0.0111,18))
```

Weighted Regression

Weighting

- to increase/decrease effect of some measurements
- only positive values are allowed
- instead of least squares weighted least squares are used

$$\frac{\sigma^2}{n}$$



Background

Sexual size dimorphism may increases with ambient temperature in spiders.

Design

Males and females of *Zodarion* spiders were sampled on 13 sites with a different temperature [ºC]. Of the average size of males and females a size ratio was calculated for each site. The number of individuals varied between sites (2 to 62 specimens).

Hypotheses

Is there relationship between the ratio and the temperature?
What is the model?

Variables

temp
number
ratio

Data

`zodarion.txt`

Analysis

```
dat<-read.delim("zodarion.txt"); attach(dat); names(dat)
plot(temp,ratio)
m1<-lm(ratio~poly(temp,2))
summary(m1)
m2<-lm(ratio~temp)
summary(m2)
m3<-update(m2,weights=number)
summary(m3)
plot(temp,ratio,xlab="Temperature",ylab="Size ratio")
abline(m2)
abline(m3,lty=2)
legend(6,1.15,c("m2","m3"),lty=1:2)
```

Multiple Regression

Background

Yield of cereals is determined by a number of variables. To predict yield with high accuracy, various effects have to be studied.

Design

On 100 plots, the yield of wheat [t/ha] was estimated together with six other variables: 1. number of overwintering plants, 2. number of ears/m^2, 3. pH of soil, 4. content of phosphorus [mg/kg], 5. content of potassium [mg/kg], 6. content of magnesium [mg/kg].

Hypotheses

Did any of six variables affect the yield?

If so which ones?

What is the model for prediction of yield?

Variables

winter

ears

pH

P

K

Mg

yield

Data

`wheat.txt`

Analysis

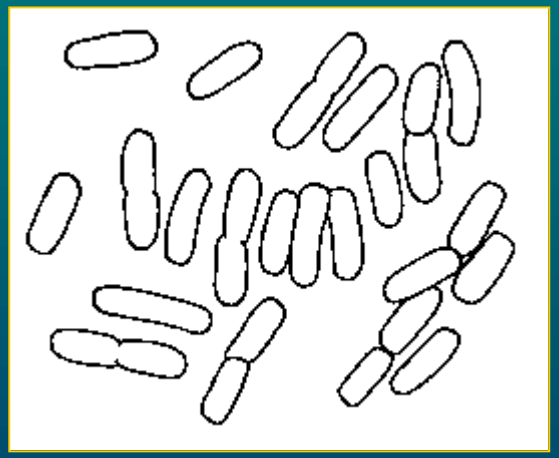
```
dat<-read.delim("wheat.txt"); attach(dat); names(dat)
pairs(yield~winter+ears+pH+P+K+Mg,panel=panel.smooth)
m1<-lm(yield~(winter+ears+pH+P+K+Mg)^2+I(winter^2)+I(ears^2)+I(pH^2)+
I(P^2)+I(K^2)+I(Mg^2))
anova(m1)
m2<-step(m1)
anova(m2)
summary(m2,corr=T)
w1<-scale(winter); e1<-scale(ears); pH1<-scale(pH)
P1<-scale(P); K1<-scale(K); Mg1<-scale(Mg)
m3<-lm(yield~(w1+e1+pH1+P1+K1+Mg1)^2+I(w1^2)+I(e1^2)+I(pH1^2)+
I(P1^2)+I(K1^2)+I(Mg1^2))
anova(m3)
m4<-update(m3,~.-w1:pH1); anova(m4)
m5<-update(m4,~.-e1:K1); anova(m5)
m26<-lm(yield~w1+pH1+K1+I(pH1^2))
anova(m26)
mean(winter);sd(winter)
mean(pH);sd(pH)
mean(K);sd(K)
summary(m26,corr=T)
```

```
plot(m26,which=1)
plot(m26,which=2)
plot(m26,which=3)
sr<-rstandard(m26)
plot(w1,sr)
plot(pH1,sr)
plot(K1,sr)
range(winter)
range(K)
plot(pH,yield,type="n")
phh<-seq(4,8,0.2)
y1<-8.71416+0.28494*(162-275.6)/50.94-0.01134*(phh-5.852)/0.381-
0.1888*((phh-5.852)/0.381)^2-0.09666*(60-106.7)/40.39
lines(phh,y1)
y2<-8.71416+0.28494*(162-275.6)/50.94-0.01134*(phh-5.852)/0.381-
0.1888*((phh-5.852)/0.381)^2-0.09666*(320-106.7)/40.39
lines(phh,y2,lty=2)
y3<-8.71416+0.28494*(400-275.6)/50.94-0.01134*(phh-5.852)/0.381-
0.1888*((phh-5.852)/0.381)^2-0.09666*(320-106.7)/40.39
lines(phh,y3,lty=3)
legend(6,9.5,c("w=162,K=60","w=162,K=320","w=400,K=320"),lty=1:3)
```

2-way ANOVA

Background

The carcinogenic disease is related to the production of toxins by certain bacteria in the body of patients. Presence of toxins can be used as an indicator of certain carcinogenic disease.



Design

In a clinical study, the amount of a toxin [units/µl] produced by four bacteria species was measured in patients with two carcinogenic and two non-carcinogenic diseases. For each disease there were 20 patients. In each patient only a single bacterial toxin was measured so there were 5 replications per bacteria species.

Hypotheses

Is the amount of toxin similar for four bacteria species and four diseases?

If not what is the difference?

Which species can be used as an indicator?

Variables

*SPECIES:*bacterA, bacterB, bacterC, bacterD

*DIAGNOSIS:*carc.rectum, carc.intestine, apendicitis, skin.absces

toxin

Data

`bacteria.txt`

Analysis

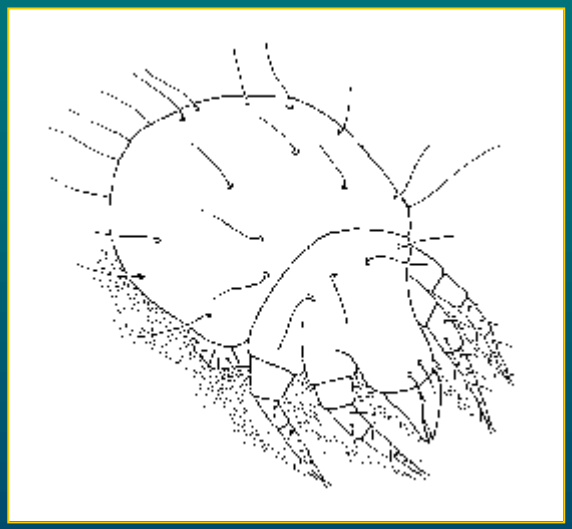
```
dat<-read.delim("bacteria.txt"); attach(dat); names(dat)
interaction.plot(species,diagnosis,toxin)
m1<-lm(toxin~species*diagnosis)
anova(m1)
summary(m1)
tapply(predict(m1),list(species,diagnosis),mean)
diagnosis1<-c(rep("carc",40),rep("non",40))
diagnosis1<-factor(diagnosis1)
m2<-lm(toxin~species*diagnosis1)
anova(m1,m2)
interaction.plot(species,diagnosis1,toxin)
species1<-species
levels(species1)
levels(species1)[2:3]<-"bacterBC"
m3<-lm(toxin~species1*diagnosis1)
anova(m2,m3)
levels(species1)
levels(species1)[c(1,3)]<-"bacterAD"
m4<-lm(toxin~species1*diagnosis1)
anova(m3,m4)
anova(m4)
summary(m4)
anova(m4,m1)
```

```
plot(m4,which=1)
plot(m4,which=2)
both<-paste(species1,diagnosis1)
both<-factor(both)
m5<-lm(toxin~both-1)
summary(m5)
confint(m5)
interaction.plot(species1,diagnosis1,toxin,type="p",
pch=1:2,ylim=c(1,2),ylab="Toxin amount",xlab="Species",legend=F)
legend(1.5,1.9,c("Carc","Non"),pch=1:2)
lines(c(1,1),c(1.85,1.96))
lines(c(1,1),c(1.09,1.2))
lines(c(2,2),c(1.35,1.46))
lines(c(2,2),c(1.07,1.18))
```

1-way ANCOVA

Background

Rate of population increase is a function of temperature in ectotherms, such as mites. A model of the relationship is essential for the control of mite pests.



Design

In the lab, population increase of two pest mite species was studied at 11 temperatures between 10 and 35 °C. The rate of increase was estimated using formula for exponential population growth. For each temperature a single measurement for each species was available.

Hypotheses

Did temperature affect the rate of increase?

Was the rate similar for both species?

What is the model of the relationship?

Variables

GENUS: genA, genB

temp

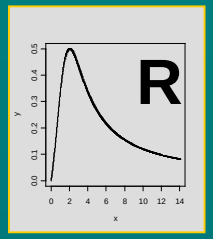
rate

Data

`mite.txt`

Analysis

```
dat<-read.delim("mite.txt"); attach(dat); names(dat)
plot(temp,rate,type="n")
points(temp[genus=="genA"],rate[genus=="genA"])
points(temp[genus=="genB"],rate[genus=="genB"],pch=16)
m1<-lm(rate~poly(temp,3)*genus)
anova(m1)
m2<-lm(rate~poly(temp,3)+genus)
anova(m2)
m3<-lm(rate~temp+I(temp^2)+I(temp^3))
summary(m3)
m4<-lm(rate~temp+I(temp^2))
summary(m4)
plot(temp,rate,xlab="Temperature",ylab="Rate")
x<-seq(from=0,to=40,by=0.1)
lines(x,predict(m4,list(temp=x)))
ci<-predict(m4,list(temp=x),se.fit=T)
names(ci)
ciU<-ci$fit+qt(.975,19)*ci$se.fit
ciL<-ci$fit+qt(.025,19)*ci$se.fit
lines(x,ciL,lty=3)
lines(x,ciU,lty=3)
```



Analyses of continuous \\

Stano Pekár

Gamma & Lognormal distributions

■ Gamma and lognormal data arise:

• precise measurements of small quantities (concentration), weight, time, etc.

• measurements are continuous

- non-negative values and zeros are not allowed
- distribution is skewed to the right

Lognormal model

• logarithmic transformation of measurements will homogenise variance and adjust asymmetry of distribution

• moments - 2 parameters (μ_{tr}, σ_{tr})
- while on log scale variance is independent of mean, on original scale variance is a function of expected mean

$$E(y) = \exp\left(\mu_{tr} + \frac{\sigma_{tr}^2}{2}\right)$$

$$Var(y) = \exp\left(\sigma_{tr}^2 - 1\right)\exp\left(2\mu_{tr} + \sigma_{tr}^2\right)$$

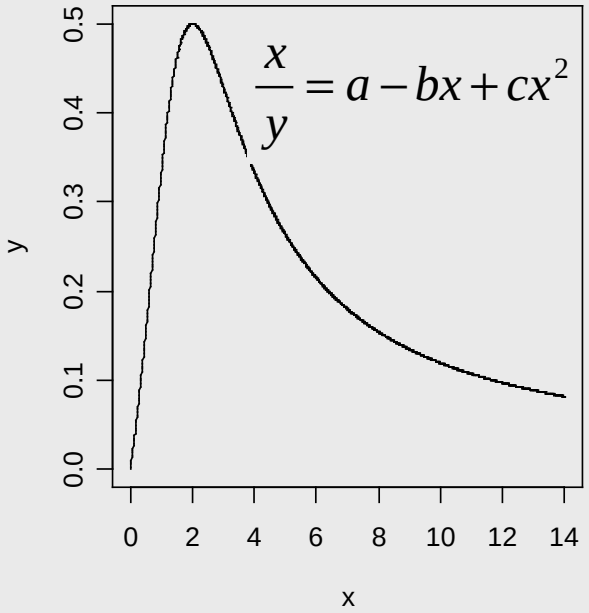
• predicted values: $\exp(Q) = median$

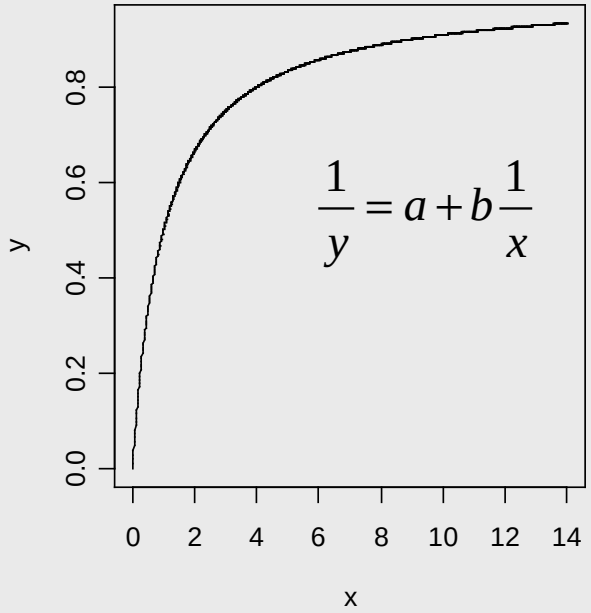
Gamma model

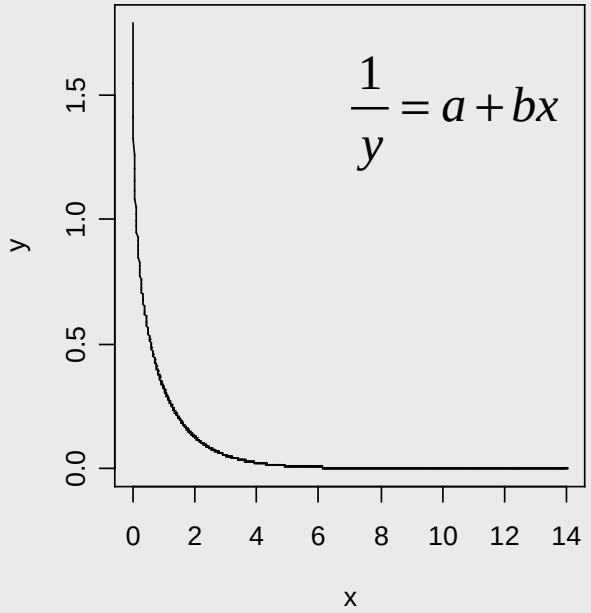
- used to model inverse polynomials moments - 2 parameters (μ, φ)

$$E(y) = \mu \qquad Var(y) = \varphi\mu^2$$

- dispersion parameter (φ) = Var(y) / μ^2







Analytical methods

- **Welch test** (`t.test`) to compare two means with heterogenous variances

- `glm(formula, Gamma(link= ...))`

- links:
- **inverse** (default) $\frac{1}{y}$

- **logarithmic** (`log`)

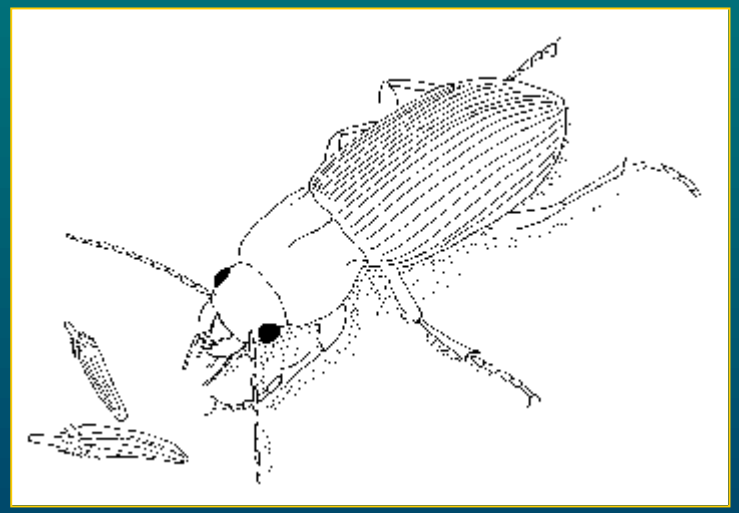
- **identity** (`identity`)

- `lm(log(y)~..)`

Simple Regression

Background

In euryphagous predators the size of prey is positively related to their body size. There is an upper limit due to e.g. morphological constraints.



Design

In the laboratory, acceptance of food was studied in 36 species of granivorous beetles. Each carabid beetle was offered seeds of various sizes [g]. Preferred seed size was recorded. For each beetle body size [mm] was recorded too.

Hypotheses

Is size of seeds related to the carabid body size?

What is the shape of the relationship?

Variables

body

seed

Data

`granivore.txt`

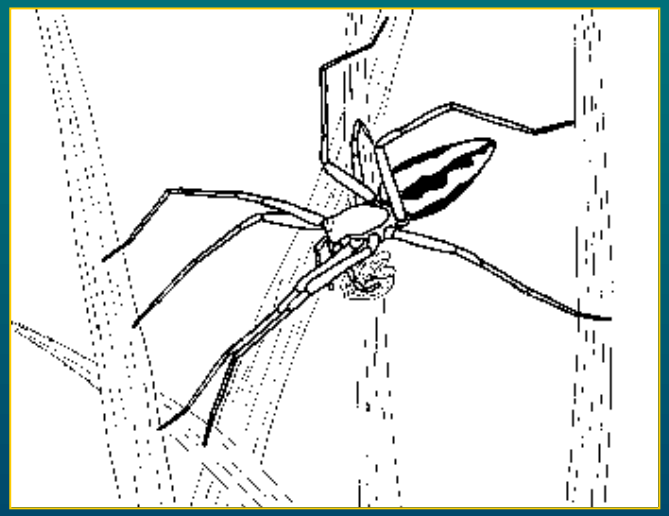
Analysis

```
dat<-read.delim("granivore.txt"); attach(dat); names(dat)
plot(body,seed)
m1<-glm(seed~I(1/body),family=Gamma)
anova(m1,test="F")
m2<-glm(seed~I(1/body)+I(1/body^2),Gamma)
anova(m1,m2,test="F")
plot(m1,which=1)
pr<-resid(m1,type="pearson"); plot(body,pr)
summary(m1)
plot(body,seed,type="n",xlab="Body size",ylab="Seed weight")
x<-seq(from=0,to=40,by=1)
lines(x,predict(m1,list(body=x),type="response"))
ci<-predict(m1,list(body=x),type="link",se.fit=T)
names(ci)
ciU<-ci$fit-qt(0.975,34)*ci$se.fit
ciL<-ci$fit+qt(0.975,34)*ci$se.fit
lines(x,1/ciL,lty=3)
lines(x,1/ciU,lty=3)
m3<-lm(seed~poly(body,2))
summary(m3)
plot(body,seed)
lines(x,predict(m3,list(body=x)))
plot(m3,which=1)
```

2-way ANOVA

Background

In the gift-giving spider a male brings a prey to a female in order to avoid being cannibalised. Several variables can potentially influence how quickly female will accept the gift.



Design

In the laboratory, effect of two variables was studied: satiation of female (satiated, starved) and their mating experience (mated, virgin). Time [s] of the gift presentation was recorded. Experiment was fully factorial, for each combination 10 males and females were used.

Hypotheses

Is presentation time affected by any of the two variables?
If it is what is the difference between factor levels?

Variables

MATING: mated, virgin
FEED: satiated, starved
time

Data

`pisaura.txt`

Analysis

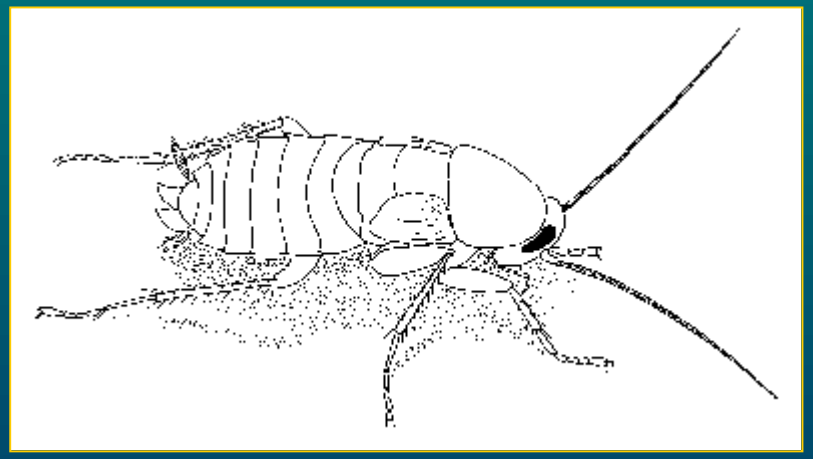
```
dat<-read.delim("pisaura.txt"); attach(dat); names(dat)
interaction.plot(mating,feed,time)
hist(time)
m1<-lm(time~mating*feed)
anova(m1)
m2<-update(m1,~.-mating:feed)
anova(m1,m2)
m3<-update(m2,~.-mating)
anova(m2,m3)
anova(m3)
plot(m3,which=1)
m4<-glm(time~mating*feed,Gamma(link=log))
anova(m4,test="F")
m5<-update(m4,~.-mating:feed)
anova(m5,test="F")
m6<-update(m5,~.-mating)
anova(m6,test="F")
plot(m6,which=1)
summary(m6)
exp(6.8222)
exp(6.8222-1.6982)
```

```
tapply(time,feed,mean)
m7<-lm(log(time)~mating*feed)
anova(m7)
m8<-lm(log(time)~feed)
summary(m8)
tapply(log(time),feed,mean)
m7<-update(m6,~.-1)
exp(confint(m7))
boxplot(918,168,names=c("Satiated","Starved"),
ylab="Presentation time",ylim=c(0,1600))
lines(c(1,1),c(581.03,1574.9))
lines(c(2,2),c(106.3,288.23))
```

2-way ANCOVA

Background

The nutritional quality of the diet affects growth of organisms in a various ways. To find optimal diet for cockroaches the following experiments was performed.



Design

Effect of five diet types (control, lipid1, lipid2, protein1, protein2) was tested on body weight [g] of male and female cockroaches. For each diet 10 females and 7 males were used. Their body weight [g] was recorded before and after the experiment.

Hypotheses

Is weight influenced by the diet type?

If so which diet resulted in largest weight?

Is weight on diets similar for males and females?

Variables

DIET: control, lipid1, lipid2, protein1, protein2

SEX: male, female

start

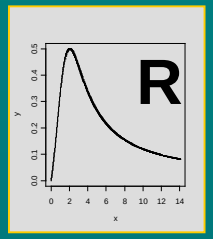
weight

Data

`cockroach.txt`

Analysis

```
dat<-read.delim("cockroach.txt"); attach(dat); names(dat)
m1<-lm(log(weight)~diet*sex*start)
anova(m1)
m7<-lm(log(weight)~diet)
anova(m7)
summary(m7)
diet2<-diet
levels(diet2)[4:5]<-"prot"
levels(diet2)[2:3]<-"lipid"
m9<-lm(log(weight)~diet2)
summary(m9)
plot(m9,which=1)
plot(m9,which=2)
m10<-lm(log(weight)~diet2-1)
exp(coef(m10))
exp(confint(m10))
boxplot(0.948,1.622,2.999,names=c("Control","Lipid","Protein"),
ylim=c(0,3.2),ylab="Weight",xlab="Diet")
lines(c(1,1),c(0.877,1.026))
lines(c(2,2),c(1.535,1.714))
lines(c(3,3),c(2.837,3.17))
```



Analyses of counts I

Stano Pekár

Poisson distribution

■ Poisson data arise when data are:
- counts/frequencies of individuals, species, cells
- events of behaviour, etc.
- always positive integers
- counts are often low (including 0)

• we count how many times an event occurred but we do not know how often it did not occur (we do not know *n*)

• moment: $E(y) = \mu = Var(y)$

Analytical methods

- χ^2 **test** (`chisq.test`) to analyse 2-dimension tables
- **Fisher exact test** (`fisher.test`) to analyse 2x2 tables
- **Mantel-Haenszel test** (`mantelhaen.test`) to analyse 3-dimension tables for independence
- **Log-linear analysis** (`loglin`) to study complex frequency tables
- **Contingency tables** (`xtabs`) to study effect of factors
- **Standard regression** (`lm`) can be used after transformation

- squareroot transformation $\sqrt{y}$

- can predict values out of bounds (negative)

- **Poisson GLM** (`glm`) to study effect of both factorial and continuous predictors

Poisson model

- `glm(..., family = poisson(link=...))`

link functions:

- logarithmic (`log`)
- squareroot (`sqrt`)
- identity (`identity`)

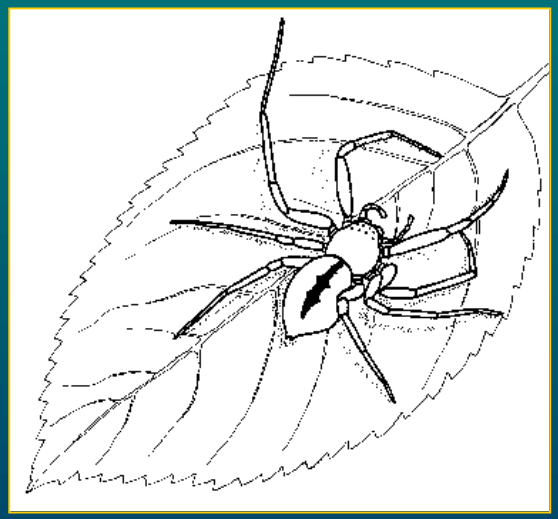
- estimated parameters are on logaritmic scale (-∞, +∞)

- inverse function to log is exp e^Q

1-way ANOVA

Background

Diversity of organisms changes with the age of the habitat. According to the intermediate disturbance hypothesis, the diversity increases and then decreases with age, thus being highest at medium age.



Design

In 15 apple orchards diversity of arachnids was studied on trees. The orchards were of variable age, classified into 3 classes: 0-9, 10-19 and 20-30 years old. Each class was represented by 5 orchards.

Hypotheses

Is diversity related to the age of orchards?

What is the trend of change?

Variables

ORCHARD: young, older, oldest

divers

Data

9,6,8,13,10,
21,14,26,17,29,
15,17,12,10,11

Analysis

```
divers<-c(9,6,8,13,10,21,14,26,17,29,15,17,12,10,11)
orchard<-factor(c(rep("young",5),rep("older",5),rep("oldest",5)))
orchard<-relevel(orchard,ref="young")
plot(orchard,divers)
m1<-glm(divers~orchard,family=poisson)
anova(m1,test="Chi")
summary(m1)
contrasts(orchard)<-"contr.helmert"
m2<-glm(divers~orchard,family=poisson)
summary(m2)
m3<-glm(divers~orchard-1,poisson)
summary(m3)
exp(confint(m3))
barplot(tapply(predict(m1,type="response"),orchard,mean),
ylab="Diversity",ylim=c(0,25))
lines(c(0.7,0.7),c(6.79,12.12))
lines(c(1.9,1.9),c(17.6,25.7))
lines(c(3.1,3.1),c(10.1,16.4))
```

Over- / under-dispersion

• arises when dispersion parameter φ $\varphi = \mathrm{Var}(y)/\mathrm{E}(y) \neq 1$

i.e. the residual deviance is not similar to the residual degrees of freedom

$$E(y) = Var(y) = \mu$$

- overdispersion: variance is larger $\rightarrow \varphi > 1$
- underdispersion: variance is smaller $\rightarrow \varphi < 1$

• causes:
- if the distribution is aggregated
- if counts are not independent
- lack of important variables, etc.
- suspicious data

• solution: use **`quasipoisson`** family

• this will influence SE of parameter estimates
- if $\varphi > 1$ then SE will be larger
- if $\varphi < 1$ then SE will be smaller

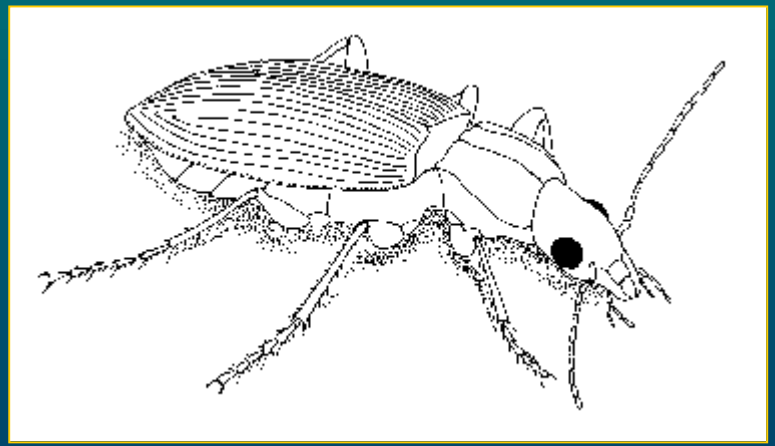
• without correction for overdispersion there would be too many false positive results (in favour of H_A)

• when using **`quasipoisson`** χ^2- and z- tests have to change to F- and t- tests

Multiple Regression

Background

Abundance of carabid beetles in cereals depends on abiotic and biotic factors. If we understand how abiotic factors influence abundance of carabids then we can adapt certain management practices to increase the abundance when needed.



Design

In the field, on 21 wheat plots the abundance of carabid beetles was studied by means of pitfall traps. At every site average day temperature [ºC] and average sun activity [W/m^2] was recorded.

Hypotheses

Was abundance of beetles affected by any of the two variables?
If so what is the model of the relationship?

Variables

temp
sun
abun

Data

carabid.txt

Analysis

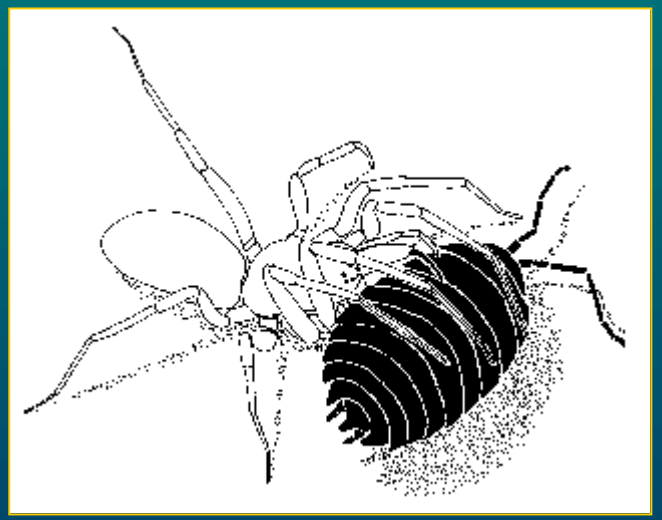
```
dat<-read.delim("carabid.txt"); attach(dat); names(dat)
pairs(abun~temp+sun,panel=panel.smooth)
m1<-glm(abun~temp*sun,family=poisson)
summary(m1)
m2<-update(m1,family=quasipoisson)
anova(m2,test="F")
plot(m2,which=1)
plot(m2,which=4)
pr<-resid(m2,type="pearson")
plot(sun,pr)
plot(temp,pr)
abun[21]
m3<-glm(abun~temp*sun,poisson,subset=-21)
anova(m3,test="Chi")
m4<-update(m3,~.-temp:sun)
anova(m4,test="Chi")
summary(m4)
(75.292-22.836)/75.292
range(sun)
range(temp)
xyz<-expand.grid(sun=seq(900,3500,50),temp=seq(9,30,0.5))
xyz$density<-as.vector(predict(m4,xyz,type="response"))
library(lattice)
wireframe(density~sun+temp,xyz)
```

3-way ANOVA

Background

Some predators use conditional strategies to catch prey. The use of strategy often depends on the characteristics of prey.

	slow		fast	
	small	large	small	large
stratA	19	10	21	12
stratB	4	10	0	8
stratC	0	1	1	2



Design

In the field, it was observed which of three strategies spiders used to capture prey. For each trial, size (two size classes) and movement (slow or fast) of prey was recorded. Altogether 88 trials were observed.

Hypotheses

Is use of strategy influenced by prey size and its movement?

If so which prey is captured by strategy A, B and C?

Variables

PREY: fast, slow

SIZE: large, small

STRATEGY: stratA, stratB, stratC

freq

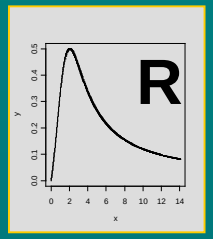
Data

`predator.txt`

Analysis

```
dat<-read.delim("predator.txt"); attach(dat); names(dat)
interaction.plot(strategy,prey,freq)
interaction.plot(strategy,size,freq)
m1<-glm(freq~strategy*size*prey,family=poisson)
summary(m1)
anova(m1,test="Chi")
m2<-update(m1,~.-strategy:size:prey)
anova(m2,test="Chi")
m3<-update(m2,~.-strategy:prey)
anova(m3,test="Chi")
summary(m3)
attacks<-tapply(predict(m3,type="response"),list(size,strategy),mean)
attacks
both<-paste(strategy,size)
m4<-glm(freq~factor(both)-1,poisson)
summary(m4)
exp(confint(m4))
```

```
barplot(attacks,beside=T,ylab="No. of attacks", xlab="Strategy",
legend.text=c("large","small"),ylim=c(0,25))
lines(c(1.5,1.5),c(7,16.3))
lines(c(2.5,2.5),c(14.4,26.9))
lines(c(4.5,4.5),c(5.5,13.8))
lines(c(5.5,5.5),c(0.6,4.6))
lines(c(7.5,7.5),c(0.4,3.9))
lines(c(8.5,8.5),c(0.03,2.2))
```



Analyses of counts II

Stano Pekár

Negative-binomial distribution

■ NB is a parametric alternative to Poisson model with overdispersion

• distribution of y is strongly asymmetric with many zeros

• NB has two parameters, μ and θ

• moments:

$$E(y) = \mu$$

$$Var(y) = \mu + \frac{\mu^2}{\theta}$$

- θ is aggregation parameter $(0,\infty)$

- if $\theta \geq 1$ .. random distribution, $\theta < 1$ .. aggregated distribution

- θ can be estimated from

$$\hat{\theta} = \frac{\bar{y}^2}{s^2 - \bar{y}}$$

NB model

`glm.nb(formula)` from ***MASS*** library

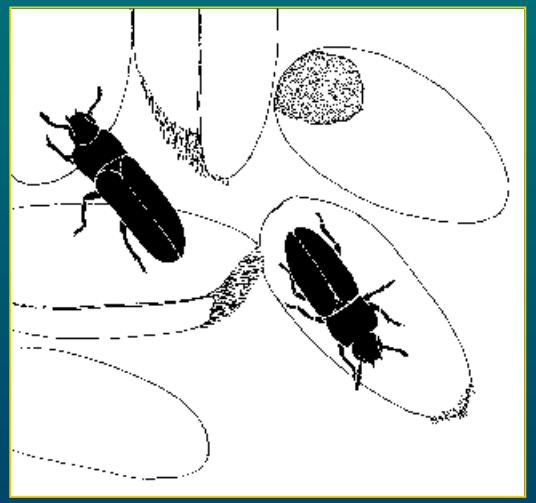
• links:
`log` (default)
`sqrt`
`identity`

• begin with Poisson model, if overdispersion is large switch to
`glm.nb`

1-way ANOVA

Background

Grain beetles are serious pests in grain stores. They may occur not only in the grain but also in crevices of corridors. It is essential to know where they occur before control methods are applied.



Design

Density of grain beetles was surveyed in a grain store by means of sticky traps. Traps were installed in two places: 25 traps in the corridors and 25 traps in the grain. After few days number of beetles was recorded.

Hypotheses

Is density of beetles similar on both places?
If not how different it is?

Variables

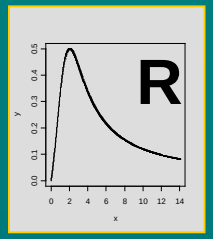
PLACE: floor, grain
density

Data

`beetle.txt`

Analysis

```
dat<-read.delim("beetle.txt"); attach(dat); names(dat)
plot(place,density)
table(density)
tapply(density,place,mean)
m1<-glm(density~place,family=quasipoisson)
anova(m1,test="F")
summary(m1)
plot(m1,which=1)
tapply(density,place,var)/tapply(density,place,mean)
tapply(density,place,function(x) mean(x)^2/(var(x)-mean(x)))
library(MASS)
m2<-glm.nb(density~place)
anova(m2)
summary(m2)
plot(m2,which=1)
exp(confint(m5))
barplot(tapply(predict(m2,type="response"),place,mean),ylab="Density",
ylim=c(0,200))
lines(c(0.7,0.7),c(49.6,197.2))
lines(c(1.9,1.9),c(9.7,38.9))
```



Analyses of proportions

Stano Pekár

Binomial distribution

■ Binomial data arise:

• when we count response to a certain stimulus → **dose-response studies**

• whenever we record whether an event has occurred or not within a known population (n)

• events: death, birth, germination, attack, consumption, reaction, etc.

• there are no classical replications - records are clustered to p or q

• p .. probability of successes, q .. probability of failures

• clustering of responses:

$$p = \frac{100}{200} + \frac{200}{300} = \frac{300}{500} = 0.6$$

$$p = \frac{0.5 + 0.667}{2} = 0.58$$

- distribution is bounded [0 < p < 1]
- variance is not constant, maximal when $p = q = 0.5$
- moments $E(y) = n\pi$ $Var(y) = n\pi(1-\pi)$
- estimated parameters are on logit scale (-∞, +∞)
- logistic model will always asymptote at 0 and 1

$$\log\left(\frac{p}{1-p}\right) = a + bx$$

- predicted values are then always within [0, 1]

- inverse function to logit is ani-logit where Q is a parameter estimate

$$\hat{y} = \frac{1}{1+e^{-Q}}$$

- odds ratio $\frac{p}{1-p} = e^{-Q}$

Analytical methods

- **Exact binomial test** (`binom.test`) to compare a single proportion
- **Proportion test** (`prop.test`) to compare two proportions
- **Contingency tables** (`xtabs`) to study effect of factors
- **Logistic regression** to study effect of continuous predictors
- **Standard regression** (`lm`) can be used after transformation

- angular transformation $\arcsin\sqrt{p}$

- can predict values out of bounds (negative or >1)

- **Binomial GLM** (`glm`) to study effect of both factorial and continuous predictors

Binomial model

- `glm(..., family = binomial(link=...))`

link functions:

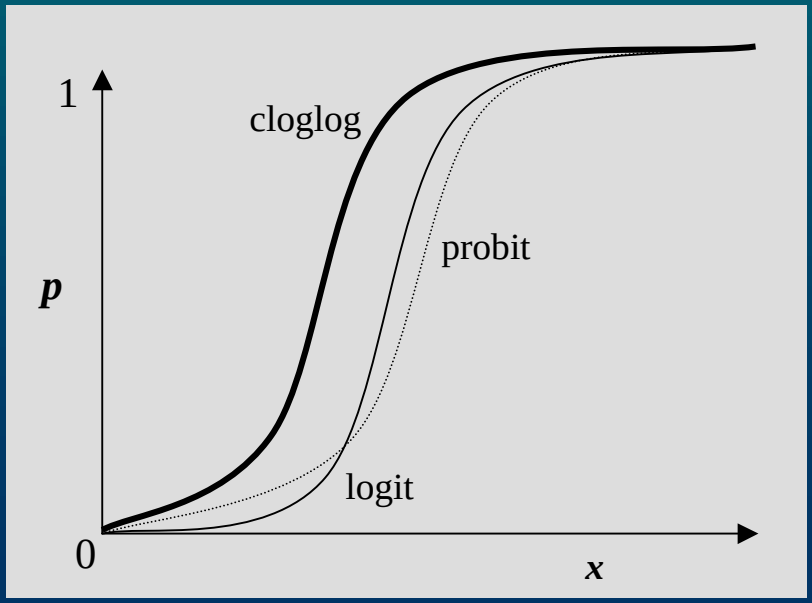
- logit (`logit`)

$$\log\left(\frac{p}{1-p}\right)$$

- probit (`probit`)

- complementary logit (`cloglog`)

$$\log(-\log(1-p))$$



Data format:

• **Binomial distribution ...** individuals within a group are homogenous
- two vectors (y, n-y) or (y, n) of integers

• **Bernoulli (binary) distribution** ... individuals within a group are heterogenous, each characterised by a continuous character
- $n = 1$
- single vector of 0's or 1's

Over- / under-dispersion

• arises when dispersion parameter φ $\varphi = \mathrm{Var}(y)/\mathrm{E}(y) \neq 1$

- overdispersion: variance is larger $\rightarrow \varphi > 1$
- underdispersion: variance is smaller $\rightarrow \varphi < 1$

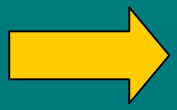
• causes:
- if the model is mispecified
- lacks important explanatory variables
- relative frequency is not constant within a group

• solution: use **`quasibinomial`** family in which variance is

estimated as $Var(y) = n\pi(1-\pi)\varphi$ instead of $Var(y) = n\pi(1-\pi)$

• this will influence SE of parameter estimates

- if $\varphi > 1$ then SE will be larger
- if $\varphi < 1$ then SE will be smaller



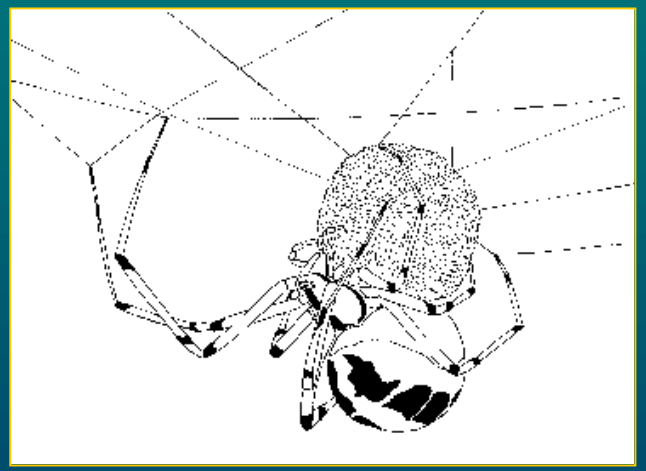
changes *P* values

• when using **`quasibinomial`** χ^2- and z- tests have to change to F- and t- tests

Regression

Background

Production of eggsac is influenced by a number of variables, such as body size, i.e. amount of consumed food. For an experimental study we need to be able to predict probability of production at a range of body sizes.



Design

In the laboratory, production of eggsacs was studied in a spider with a variable body size [mm]. As the body size was measured with the precision of 0.5 mm, all 160 individuals were classified into size classes each containing 15 to 30 specimens. Females that produced eggsac were recorded.

Hypotheses

• Is eggsac production related to the body size?

• If it is what is the shape of the relationship?

• What is the model that can be used to predict eggsac production for spider sizes of 3–12 mm?

Variables:

body

n

eggs

Data

`spider.txt`

Analysis

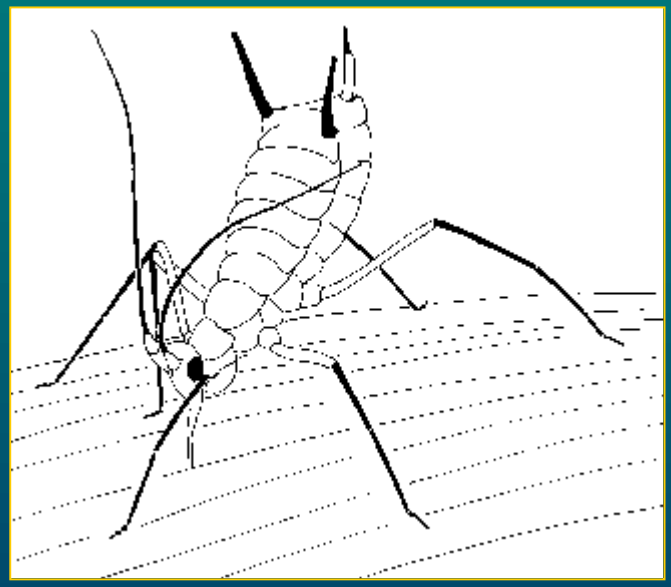
```
dat<-read.delim("spider.txt"); attach(dat); names(dat)
p<-eggs/n
plot(body,p)
tr<-asin(sqrt(p))
m1<-lm(tr~body+I(body^2),weights=n)
summary(m1)
m2<-update(m1,~.-I(body^2))
summary(m2)
x<-seq(0,12,by=0.1)
plot(body,tr)
lines(x,predict(m1,list(body=x)))
abline(m2,lty=2)
legend(3,1.5,c("m1","m2"),lty=1:2)
plot(body,p,xlim=c(3,12),ylim=c(0,1))
lines(x,sin(predict(m1,list(body=x)))^2)
lines(x,sin(predict(m2,list(body=x)))^2,lty=2)
legend(5,0.4,c("m1","m2"),lty=1:2)
y<-cbind(eggs,n-eggs)
m3<-glm(y~body+I(body^2),family=binomial)
summary(m3)
m4<-update(m3,~.-I(body^2))
```

```
plot(body,p,xlim=c(3,12),ylim=c(0,1))
lines(x,predict(m3,list(body=x),type="response"))
lines(x,predict(m4,list(body=x),type="response"),lty=2)
legend(5,0.7,c("m3","m4"),lty=1:2)
summary(m4)
m5<-update(m4,family=quasibinomial)
summary(m5)
anova(m5,test="F")
```

1-way ANCOVA

Background

Synthetic insecticides often have a species-specific efficiency. The recommended doses or concentrations then have to adjusted.



Design

In the laboratory an effect of an insecticide on the mortality of two aphid species was studied. The insecticide was applied at 6 concentrations [ppm]. Each concentration was tested on 30 individuals of both aphid species.

Hypotheses

- Is mortality affected by the concentration?
- Was the efficiency similar for both species?
- What is the LC_{50} (i.e. 50% lethal concentration) for both species?

Variables:

SPECIES: A, B

conc

n

dead

Data

aphid.txt

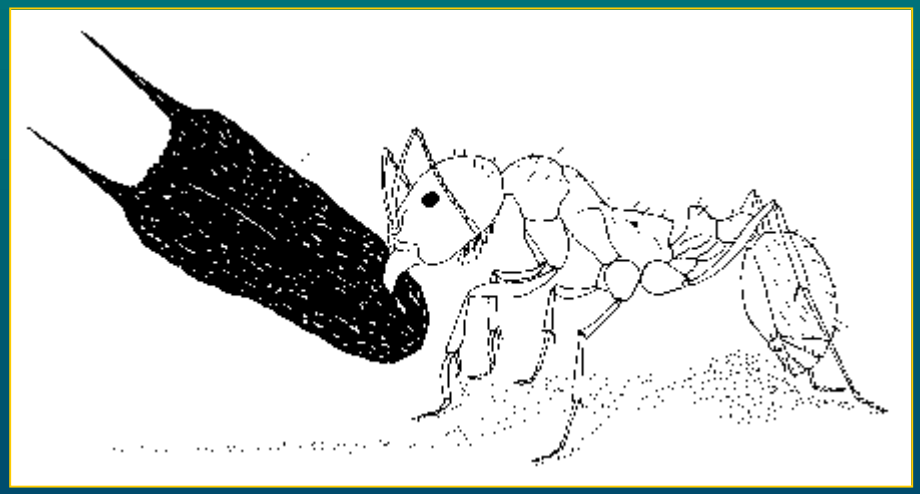
Analysis

```
dat<-read.delim("aphid.txt"); attach(dat); names(dat)
p<-dead/n
plot(conc,p,type="n")
text(conc,p,labels=as.character(species))
y<-cbind(dead,n-dead)
m1<-glm(y~log(conc)*species,binomial)
anova(m1,test="Chi")
m2<-update(m1,~.-log(conc):species)
anova(m2,test="Chi")
summary(m2)
plot(m2,which=1)
pr<-resid(m2,type="pearson"); plot(log(conc),pr)
plot(log(conc),p,type="n",xlab="Log(Concentration)",ylab="Mortality")
x<-seq(-3,2,0.1)
A<-1/(1+exp(-1.3825-1.2328*x)); lines(x,A)
B<-1/(1+exp(-1.3825+2.2117-1.2328*x)); lines(x,B,lty=2)
legend(1,0.3,c("A","B"),lty=1:2)
m3<-glm(y~species+log(conc)-1,binomial)
summary(m3)
library(MASS)
dose.p(m3,cf=c(1,3),p=0.5)
dose.p(m3,cf=c(2,3),p=0.5)
```

1-way Binary ANCOVA

Background

Granivorous ants collect various seeds and bring them into nest. Sympatrically occurring species may show trophic niche partitioning related to the size of collected seeds.



Design

Seed preference of two ant species was studied in the laboratory. Each of 25 ants of both species was offered seeds of variable size expressed as its weight [mg]. Response of ants was classified as “yes” or “no” if it took or refused to take a seed, respectively.

Hypotheses

- Is acceptance related to the seed size?
- Did both species have similar preference for seed sizes?
- If not what is the threshold size of seeds for both species?

(The threshold size is defined as a size that is accepted with higher than 90% probability)

Variables:

SPECIES: specA, specB

seed

take

Data

`ant.txt`

Analysis

```
dat<-read.delim("ant.txt"); attach(dat); names(dat)
library(lattice)
xyplot(take~seed|species)
m1<-glm(take~seed*species,family=binomial)
summary(m1)
anova(m1,test="Chi")
m2<-glm(take~log(seed)*species,binomial)
AIC(m1,m2)
plot(seed,take,type="n",xlab="Seed weight",ylab="Transported")
x<-seq(0,3,0.01)
A<-1/(1+exp(-4.012+8.364*x)); lines(x,A)
B<-1/(1+exp(-4.012+10.957+(8.364-19.147)*x));lines(x,B,lty=2)
legend(1.5,0.8,c("specA","specB"),lty=1:2)
(log(0.9/0.1)-4.012)/-8.346
(log(0.9/0.1)-4.012+10.957)/(-8.346+19.147)
```