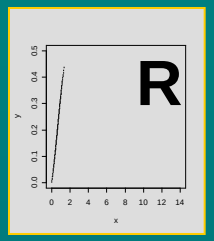


Analysis of Biological Data

Stano Pekár

Content

1) R Environment
   Exploratory Data Analysis
   Regression models

2) The first example
   Systematic components

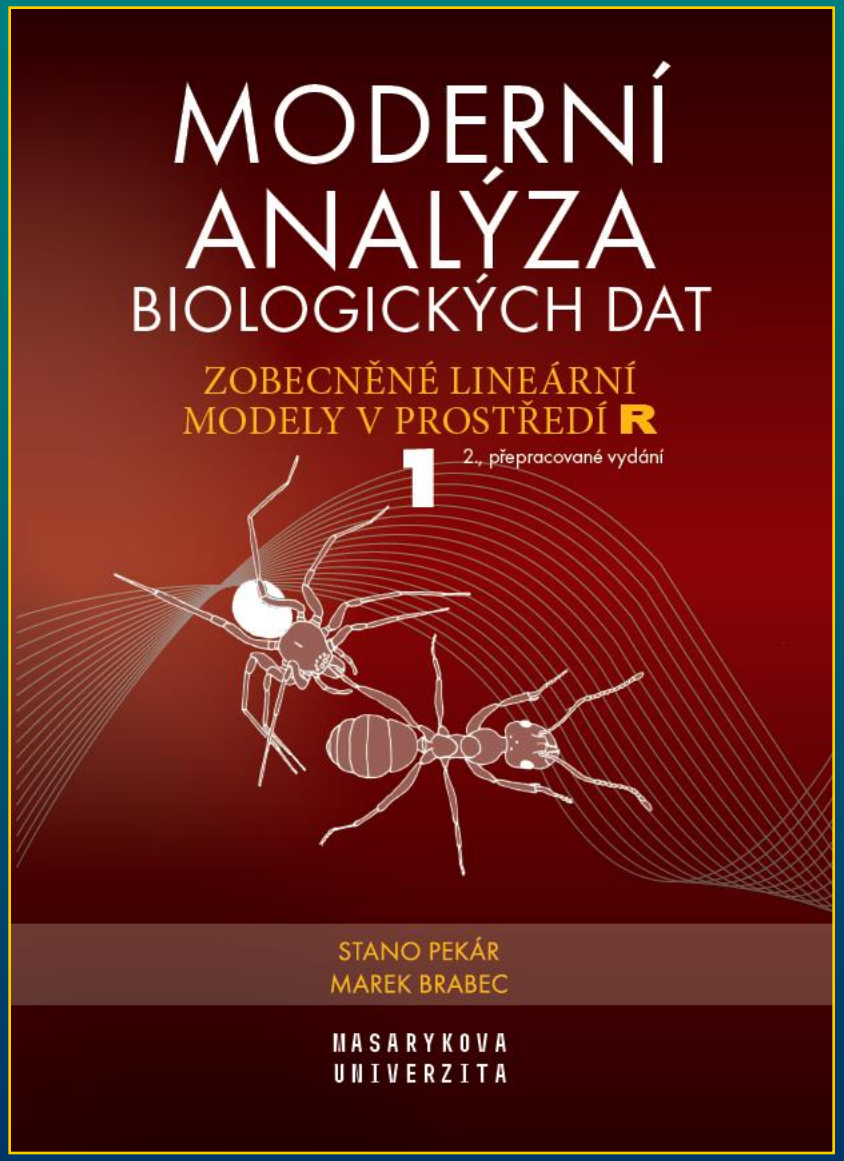
3) Stochastic components
   Analyses of continual measurements

4) Analyses of continual measurements II
   Analyses of counts

5) Analyses of counts II
   Analyses of proportions

Literature


MODERNÍ
ANALÝZA
BIOLOGICKÝCH DAT
ZOBECNĚNÉ LINEÁRNÍ MODELY
V PROSTŘEDÍ R
1
STANO PEKÁR
MAREK BRABEC
Scientia


MODERNÍ
ANALÝZA
BIOLOGICKÝCH DAT
ZOBECNĚNÉ LINEÁRNÍ
MODELY V PROSTŘEDÍ R
1
2., přepracované vydání
STANO PEKÁR
MAREK BRABEC
MASARYKOVA
UNIVERZITA

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

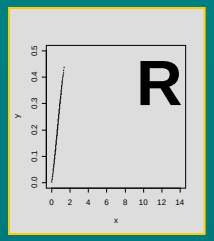
• (dependent) is the variable whose variation we aim to understand, the variable that we measure, it goes on ordinate

- 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

- software available from www.r-project.com

- data from
- https://www.press.muni.cz/edicni-rady-munipress/moderni-analyza-biologickych-dat-1

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
- vectors: numeric, character, logical
- arguments (in parentheses): use their names or without at specified order

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

```
> mean(y)
[1] 6
> var(y)
[1] 11
> y2 <- scale(y); y2
```

```
attr(,"scaled:center")
[1] 6
attr(,"scaled:scale")
[1] 3.316625
> mean(y2)
[1] 0
> var(y2)
     [,1]
[1,]    1
```

Data frames

Created in R:

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

• export: **`write.table`**

Imported:

- from Excel via clipboard

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

• 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

```
> size <- rep(c("small","medium","large"), c(4,3,4)); size
[1] "small"  "small"  "small"  "small"  "medium" "medium" "medium"
[8] "large"  "large"  "large"  "large"
> dat <- data.frame(x, y, size); dat
     x  y   size
1  1.0  1  small
2  1.1  2  small
3  1.2  3  small
```

```
> is.factor(size)
[1] FALSE
> size <- factor(size)
```

```
> levels(size)
[1] "large"  "medium" "small"
> size1 <- relevel(size, ref="medium"); levels(size1)
[1] "medium" "large"  "small"
```

Exploratory Data Analysis

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.

```
> mean(amount)
[1] 0.690625
> median(amount)
[1] 0.705
> mean(amount, trim=0.1)
[1] 0.6864286
```

Variance

- Var(y), σ^2: a theoretical measure of the variability
- 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*.

```
> var(amount)
[1] 0.2162996
> sd(amount)
[1] 0.4650802
> sem <- sd(amount)/sqrt(length(amount)); sem
[1] 0.1162700
```

Confidence Intervals

• of a parameter (mean): if large number of samples is taken from a 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

Example

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

```
> mean(amount) + sem*c(qt(p=0.0255,df=15), qt(p=0.975,df=15))
[1] 0.4428013 0.9384487
```

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 table of means for *SOIL* and *FIELD*, and table of SEM for *FIELD*.

```
> tapply(X=amount, INDEX=list(soil,field), FUN=mean)
       pasture   rape
dry     1.1575 1.1050
moist   0.2375 0.2625
> tapply(amount, soil, function(x) sd(x)/(sqrt(length(x))))
        dry      moist
0.03776986 0.03223795
```

Graphics

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

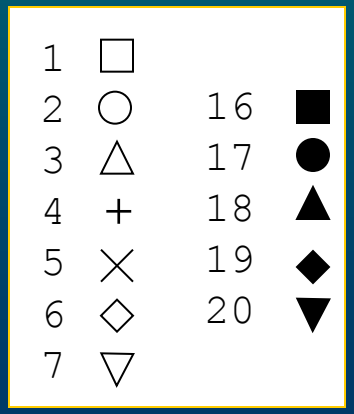
- to get all graphic parameters: **`?par`**
- to split window to subplots: **`par(mfrow)`**
- to add legend .. **`legend`**
- graph window size: **`x11`**

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

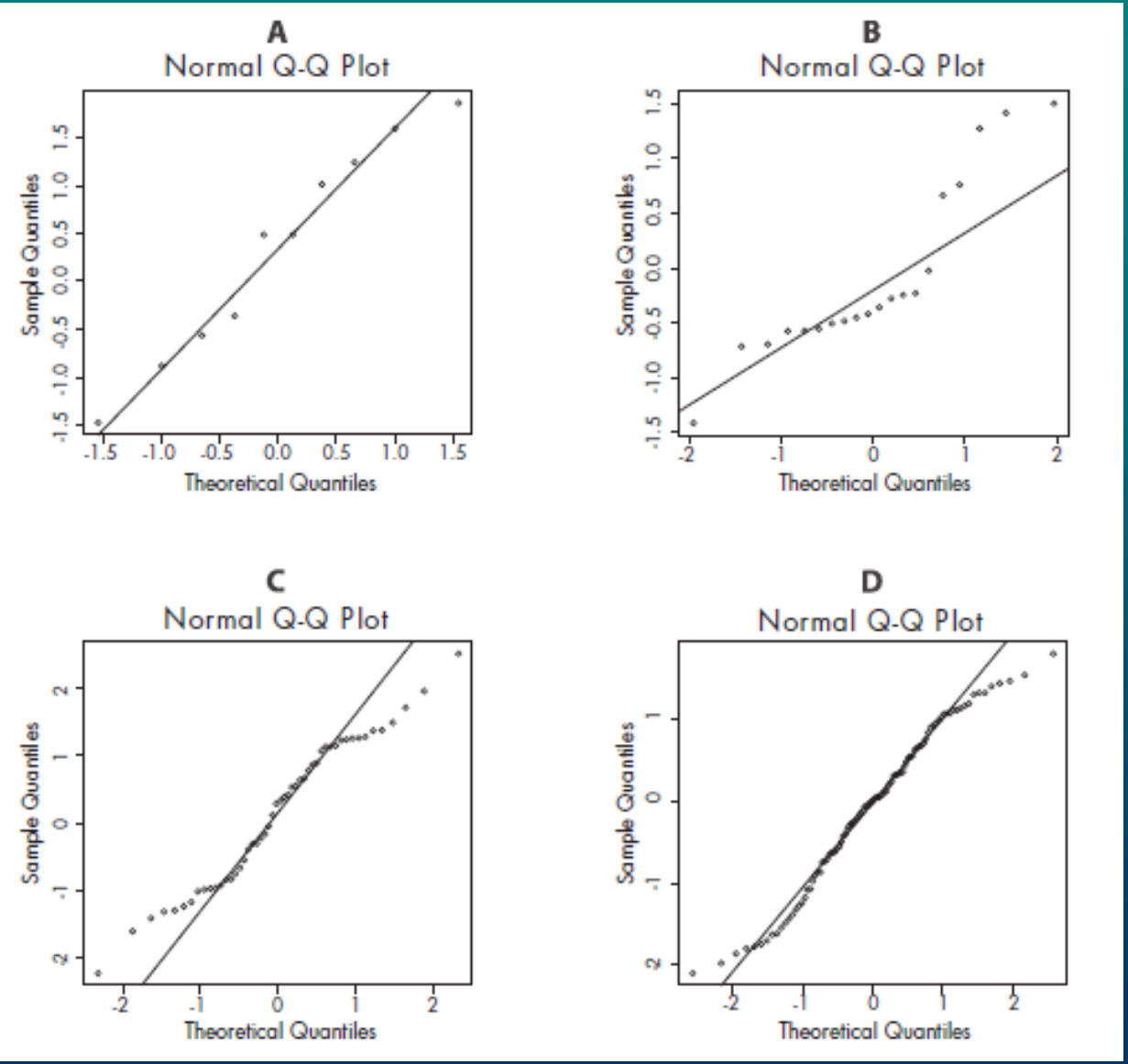
- compare a single variable with normal: **`qqnorm`**
- compare distributions of two variables: **`qqplot`**
- to add diagonal line: **`qqline`**

Example

Make q-q plot of data from normal distribution.

```
> y1 <- rnorm(n=10, mean=0, sd=1)
> y2 <- rnorm(20,0,1)
> y3 <- rnorm(50,0,1)
> y4 <- rnorm(100,0,1)
> qqnorm(y1); qqline(y1)
> qqnorm(y2); qqline(y2)
> qqnorm(y3); qqline(y3)
> qqnorm(y4); qqline(y4)
```

Deviations from normal distribution

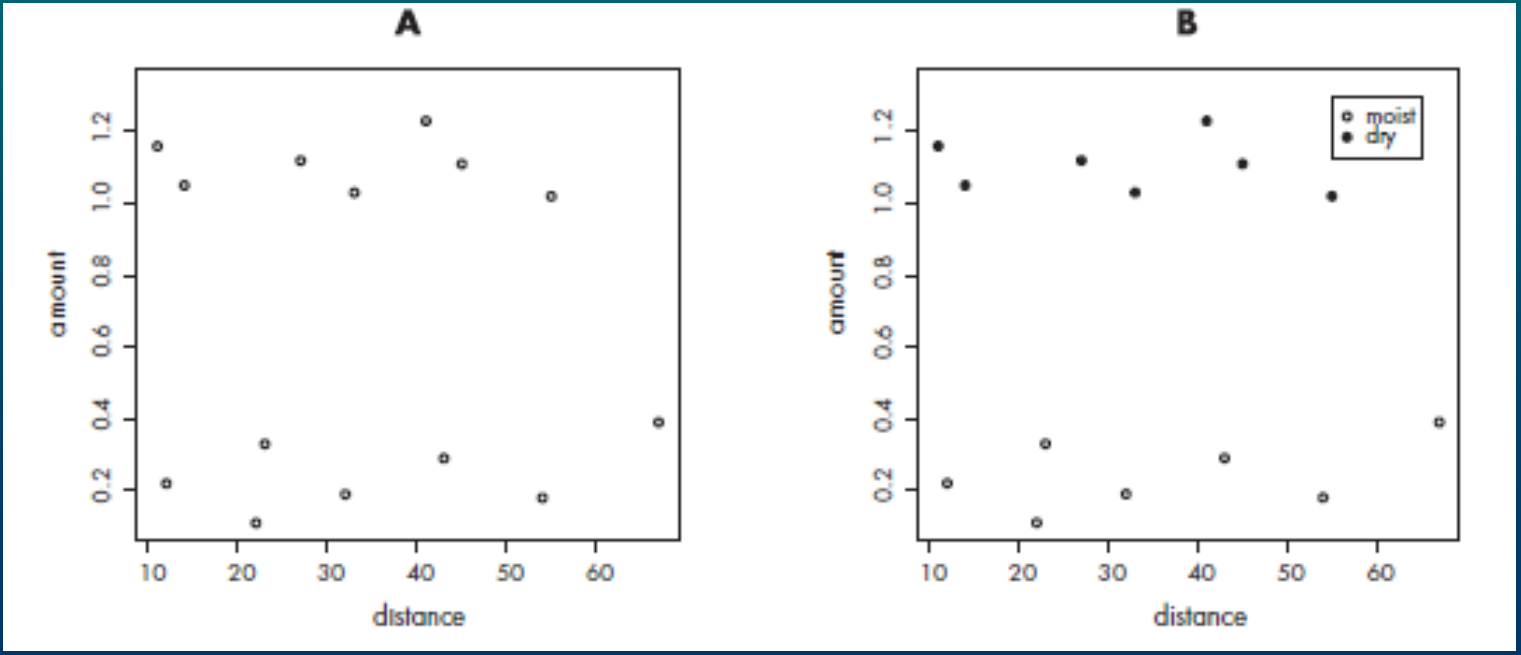


Scatter plots

- 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*.

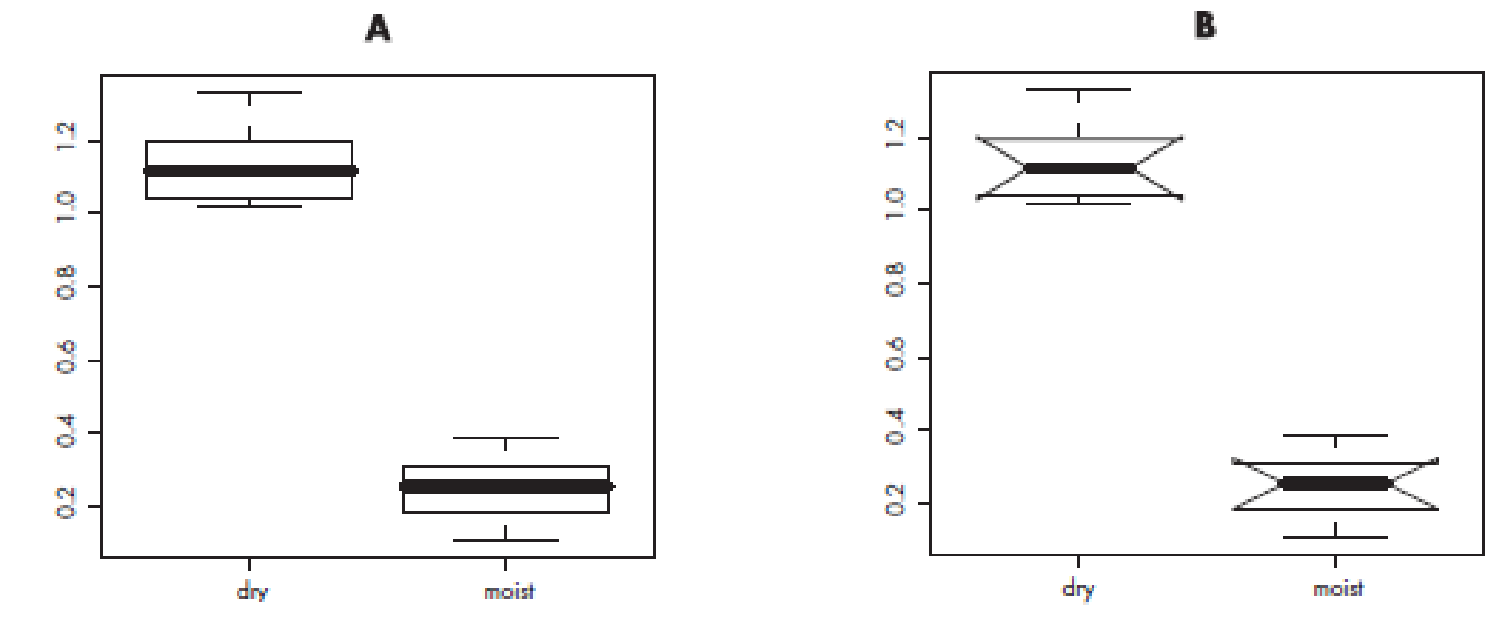


Box plots

- when there are categorical explanatory variables
- argument **`notch`** for boxes with CI_{95} for median

Example

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



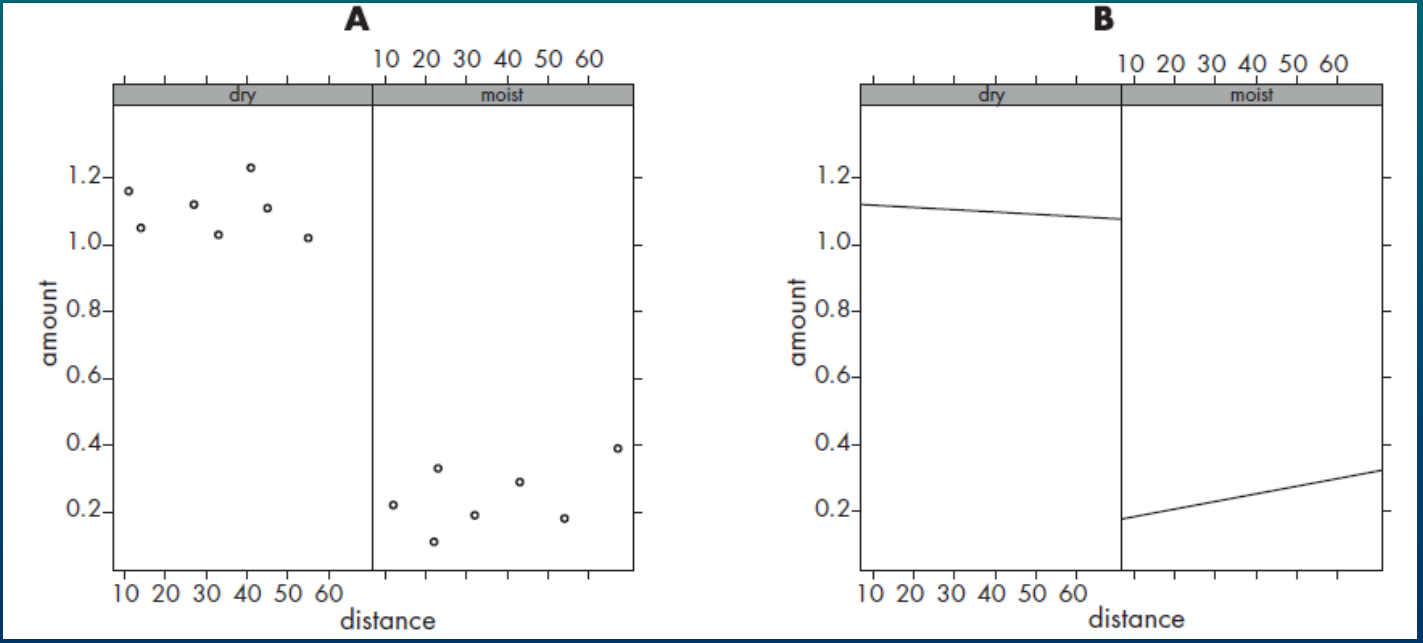
Panel plots

- for data with both categorical and continuous explanatory variables

- **`xyplot`** from library *lattice*

Example

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



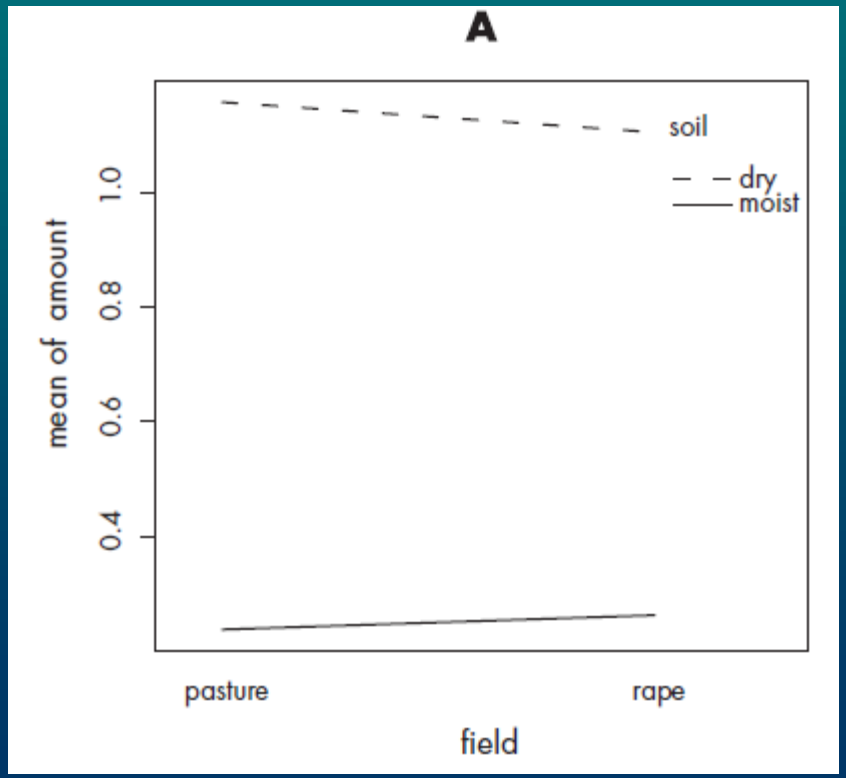
Interaction plot

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

.. `interaction.plot`

Example

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

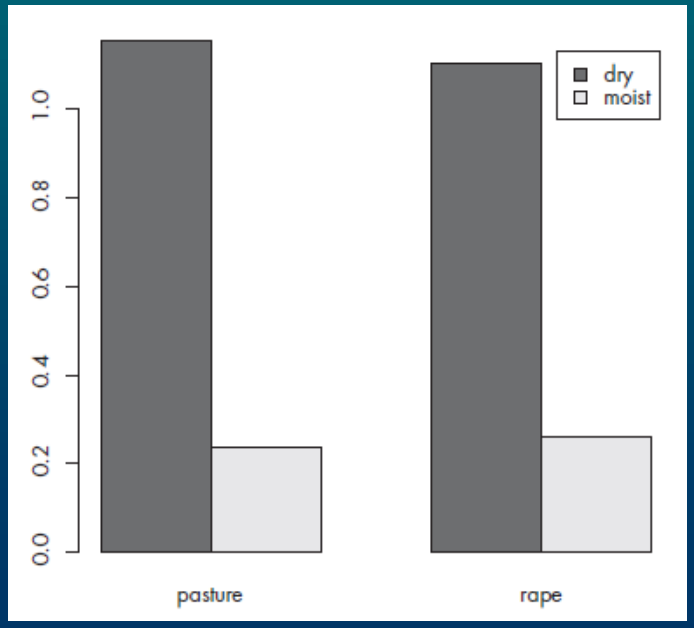


Bar plot

- 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

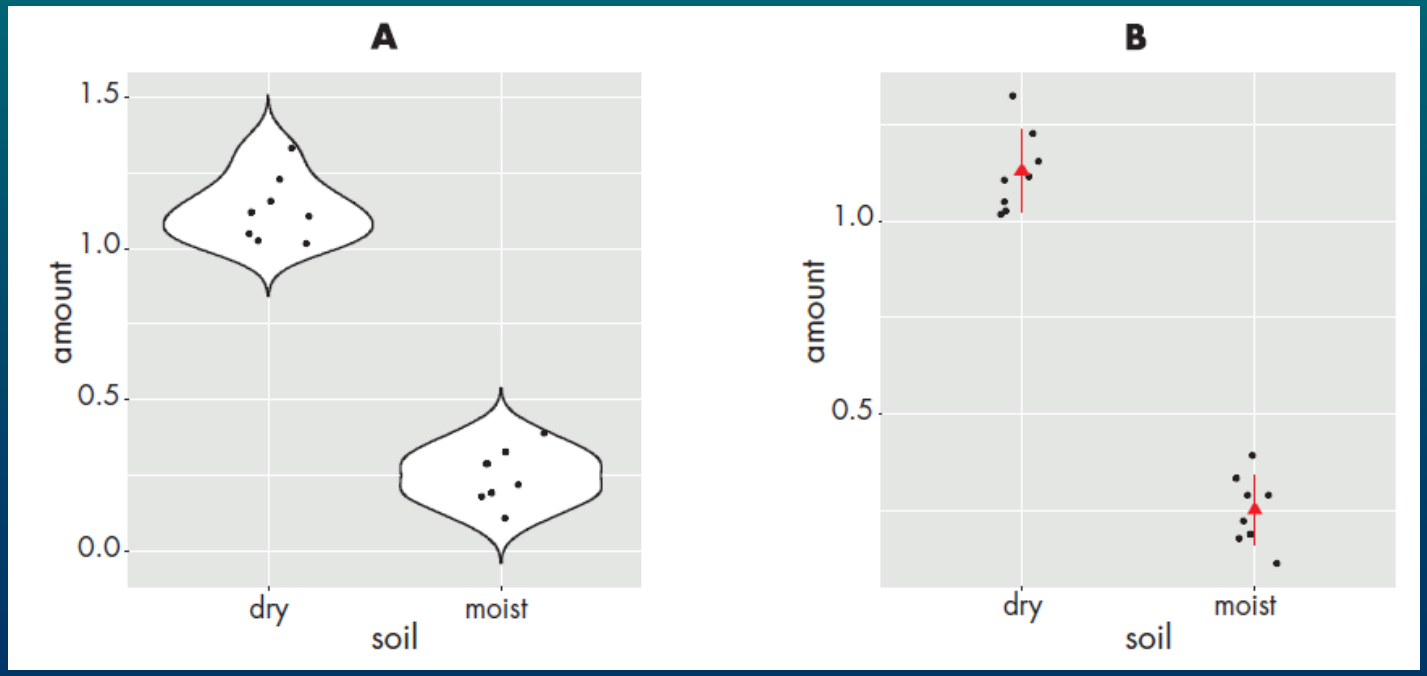
- when data include 2 continuous explanatory variables
- **`wireframe`** (*lattice*) produces 3-dimensional plot

Elegant plots

- use *ggplot2* package
- **`ggplot`**

Example

Make violinplot of *SOIL* and *amount* and dotplot of *SOIL* and *amount* .



Graphs with functions

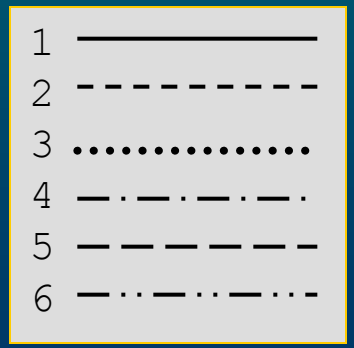
- final plot of estimated models

use *visreg* package

- **`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 $y = \log(x)$

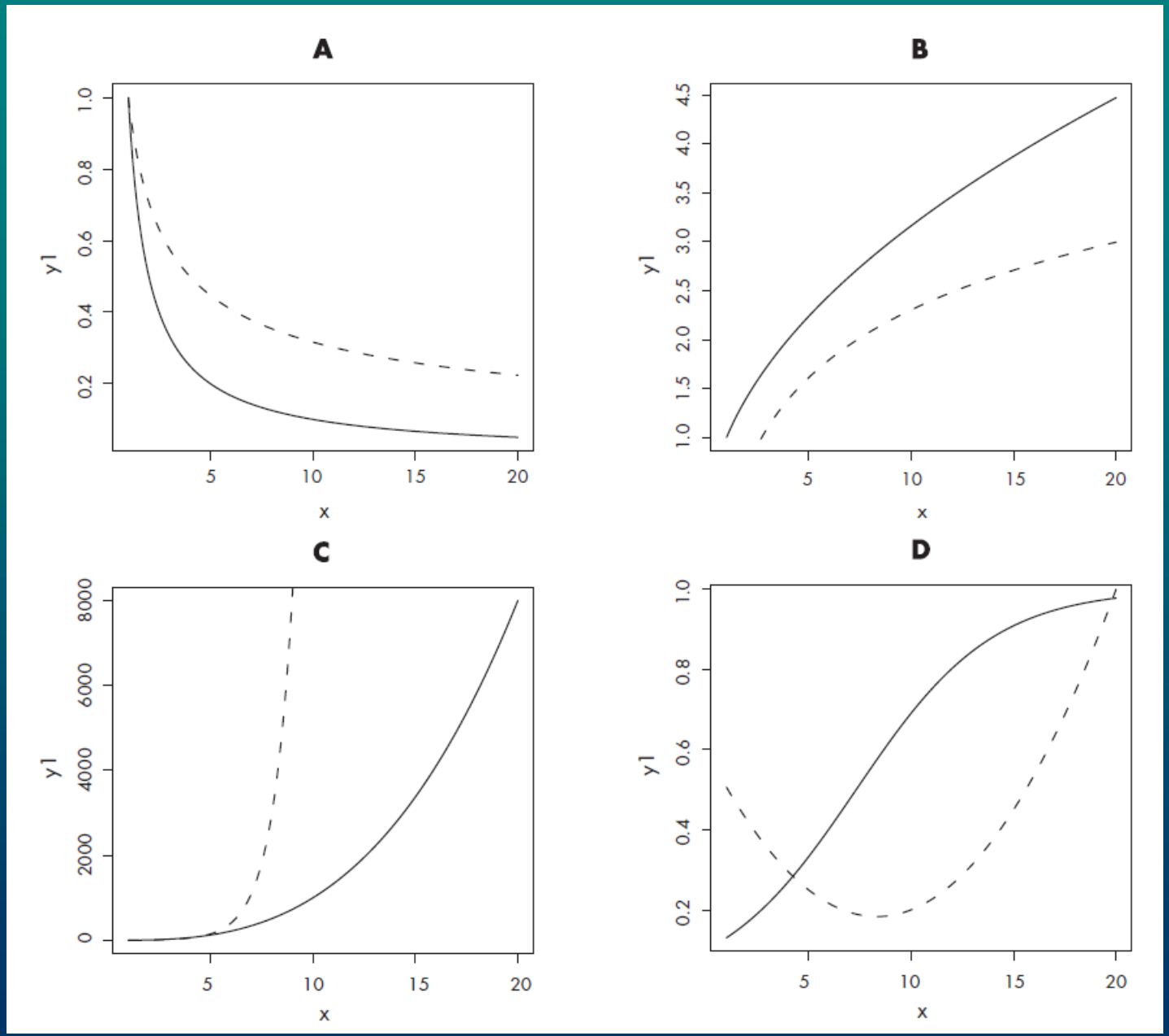
power $y = x^{3}$

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

squareroot $y = \sqrt{x}$

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

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

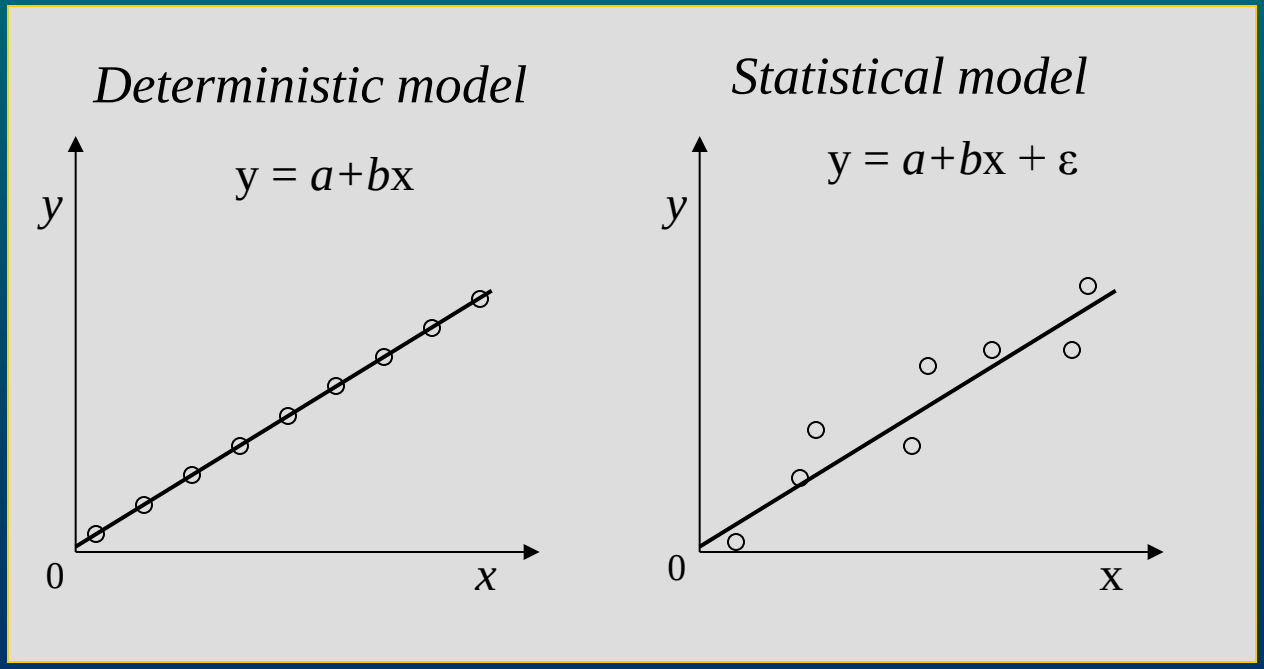

A
B
C
D
y1
x
0.2
0.4
0.6
0.8
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
0
2000
4000
6000
8000
5
10
15
20

Statistical Modelling

Regression model

- includes systematic and stochastic components

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



• assumptions of the stochastic component:

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

= 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

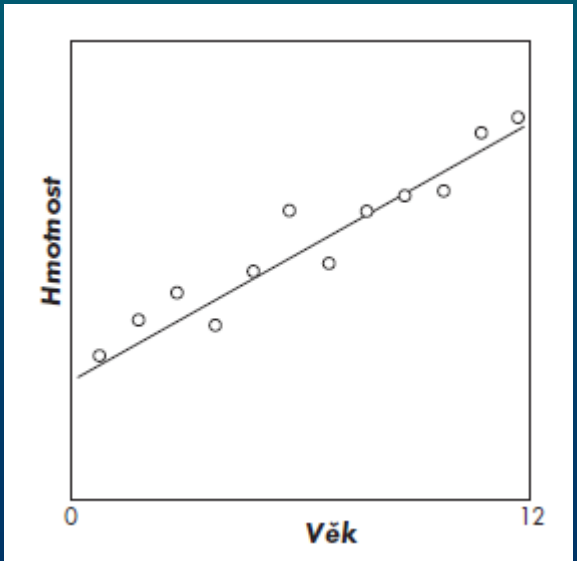
• measure of fit is deviance not sum of squares
- null deviance = SST
- residual deviance = SSE
- ANODEV table = ANOVA table

Rozdělení	Jméno linku	Link funkce	Rozptyl
Gaussovo (normální)	identity	1	μ
Gama	inverse	$\frac{1}{\mu}$	μ^2
Inverzní Gaussovo		$\frac{1}{\mu^2}$	μ^3
Poissonovo	log	$\log(\mu)$	μ
Quasipoissonovo	log	$\log(\mu)$	$\varphi\mu$
Binomické	logit	$\log\frac{\mu}{1-\mu}$	$\frac{\mu(1-\mu)}{n}$
Quasibinomické	logit	$\log\frac{\mu}{1-\mu}$	$\frac{\varphi\mu(1-\mu)}{n}$
Quasi	libovolné v rámci přípustných funkcí	libovolná v rámci přípustných funkcí	odpovídající

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

• **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

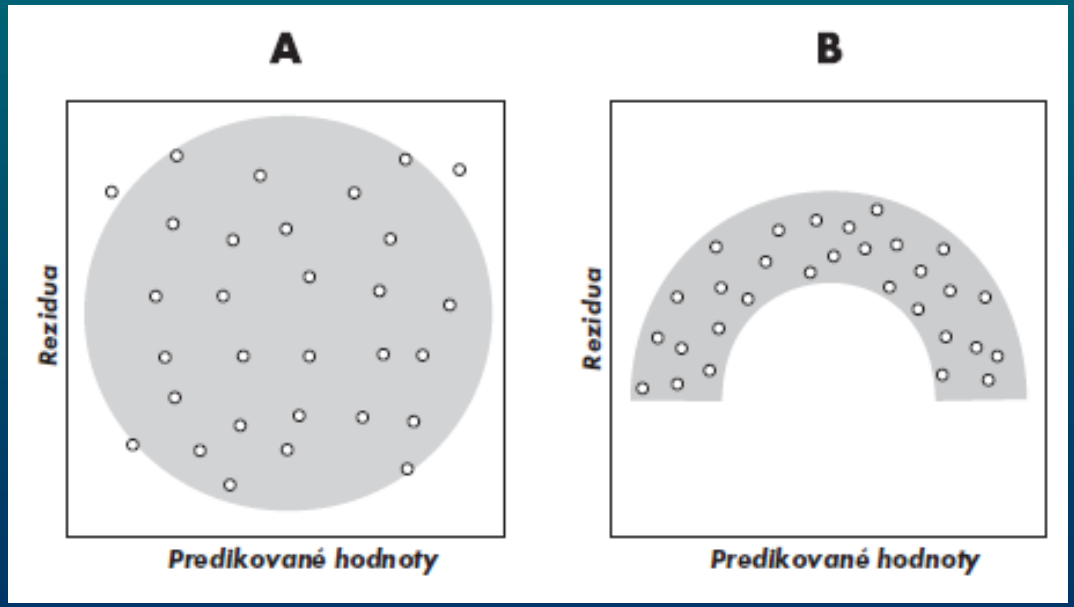
Residuals $\varepsilon_i \sim N(0, \sigma^2)$ $\text{cor}(\varepsilon_i, \varepsilon_{i'}) = 0, i \neq i'$

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

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

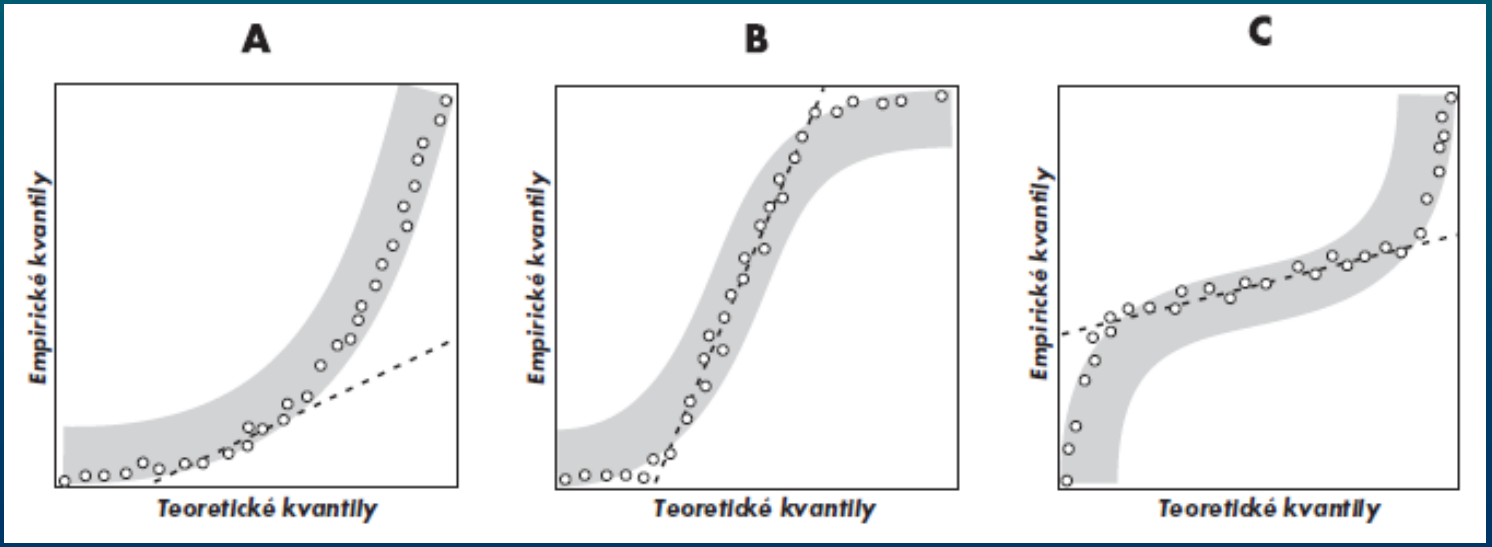
Predictor's adequacy

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



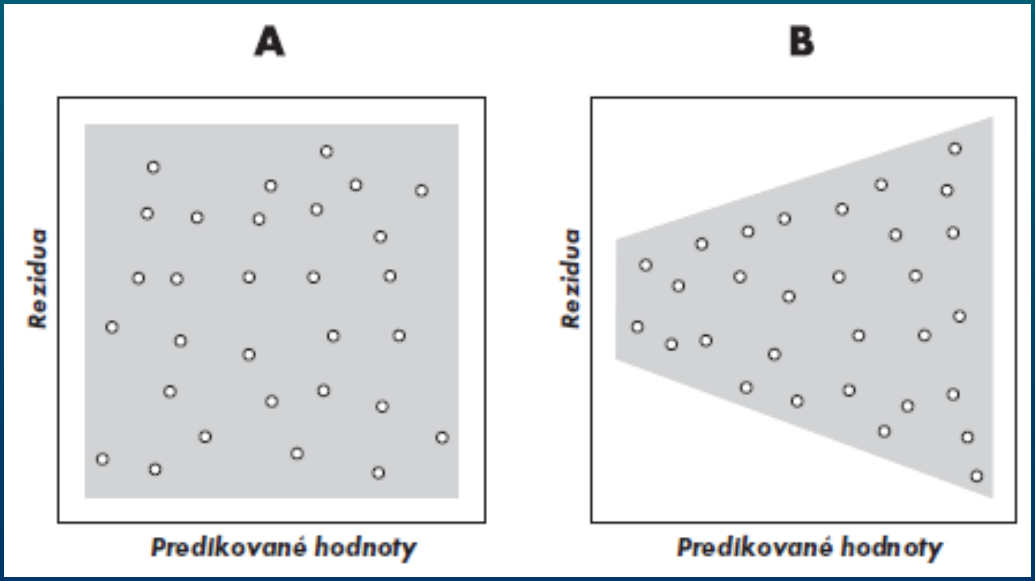
Normality

- 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



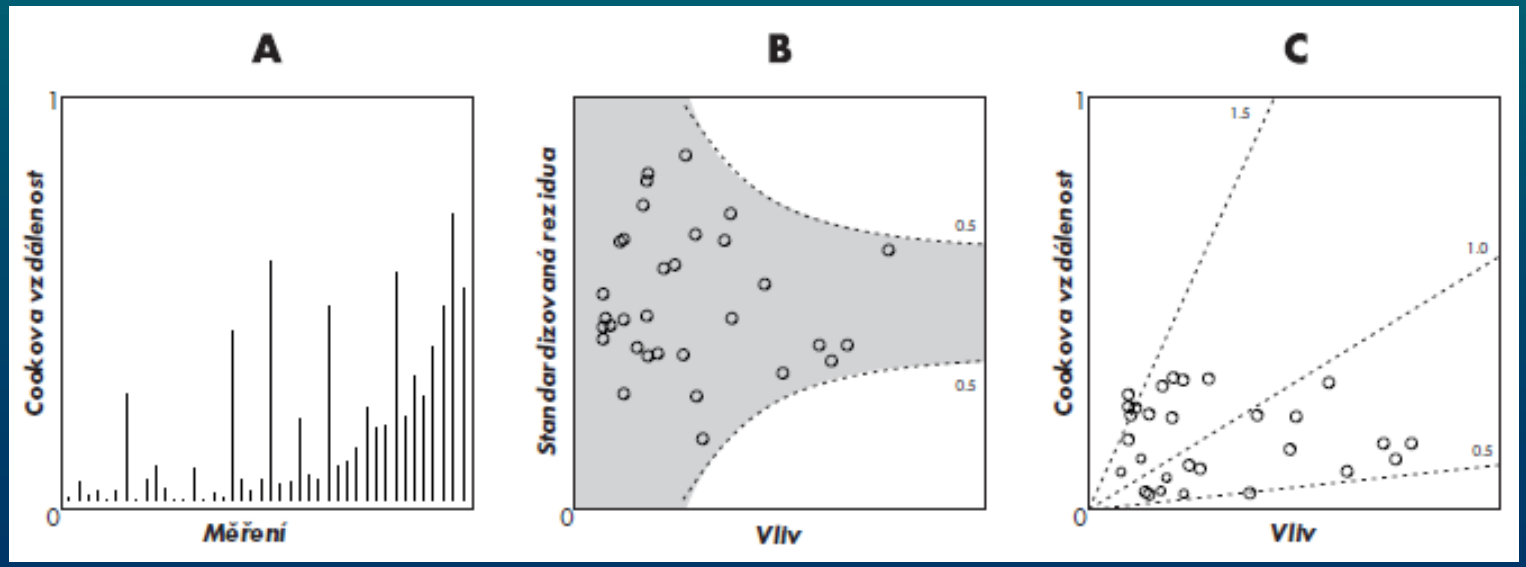
Variance homogeneity

- 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



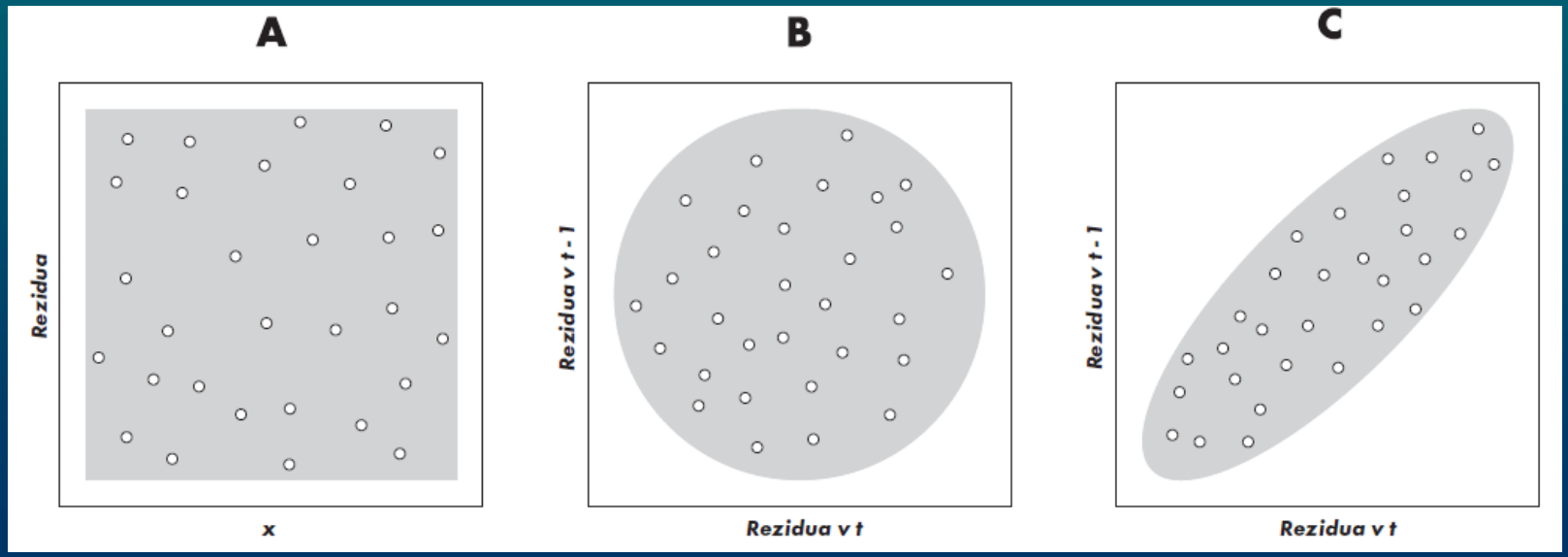
Influence

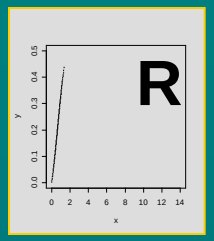
• plot of Cook's distance for each observation shows the influence of individual observations on the model fit
• values of influential observations are close to 1 and higher
• residuals versus leverage
• omit influential observations or transform the explanatory variables (using log, power, reciprocal)



Independence

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





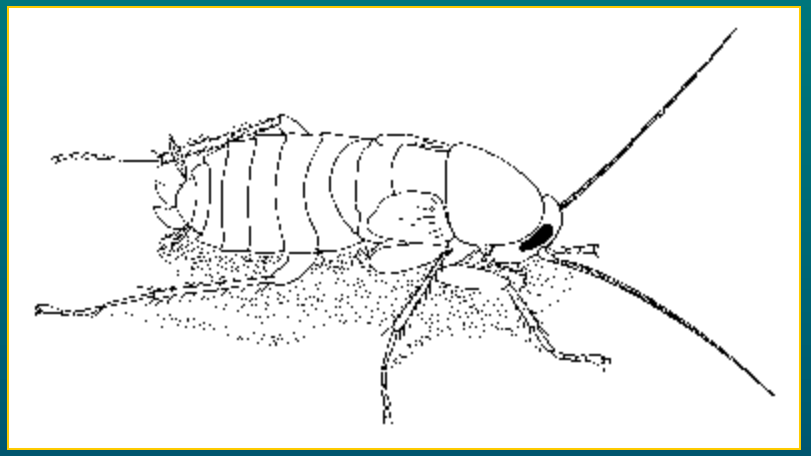
The first trial

Stano Pekár

1-way ANOVA

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 cockroaches. For each diet type there were 17 observations.

Biological hypothesis

Is nutritional quality of the diet affecting size of organisms?

Statistical hypotheses

H0: Weight is similar among diet groups.

HA: Weight is significantly different among diet groups.

Prediction:

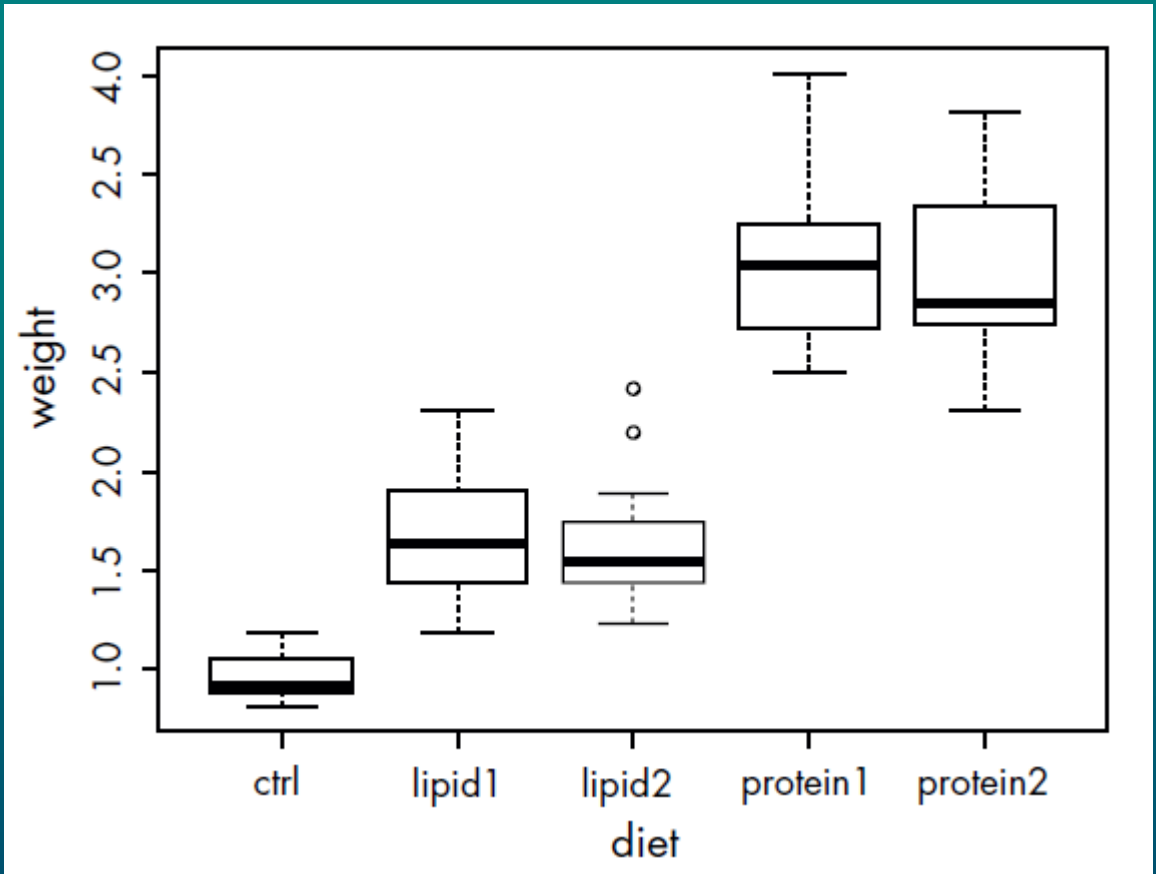
Protein-enriched diet should lead to highest weight.

Variables

DIET: control, lipid1, lipid2, protein1, protein2

weight

EDA:



Model:

$$weight_{ij} = DIET_j + \varepsilon_{ij},$$

kde $\varepsilon_{ij} \sim N(0, \sigma^2)$, nezávisle pro jednotlivá měření.

Analysis:

```
> m1 <- lm(weight~diet)
> anova(m1)
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       4 58.484  14.621  116.55 < 2.2e-16 ***
Residuals 80 10.036   0.125
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
> summary(m1)

Call:
lm(formula = weight ~ diet)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66471 -0.18294 -0.05294  0.16706  0.91706

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.9547     0.0859  11.114  < 2e-16 ***
dietlipid1     0.7282     0.1215   5.994 5.59e-08 ***
dietlipid2     0.6682     0.1215   5.501 4.41e-07 ***
dietprotein1   2.1382     0.1215  17.601  < 2e-16 ***
dietprotein2   2.0100     0.1215  16.545  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3542 on 80 degrees of freedom
Multiple R-Squared: 0.8535,     Adjusted R-squared: 0.8462
F-statistic: 116.6 on 4 and 80 DF,  p-value: < 2.2e-16
```

```
> 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)
         [,1] [,2] [,3] [,4]
ctrl     1.00  0.0  0.0  0.0
lipid1  -0.25 -0.5  0.0 -0.5
lipid2  -0.25 -0.5  0.0  0.5
protein1 -0.25  0.5  0.5  0.0
protein2 -0.25  0.5 -0.5  0.0
```

```
> summary(lm(weight~diet))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.06365    0.03842  53.718   <2e-16 ***
diet1       -1.10894    0.07683 -14.433   <2e-16 ***

diet2        1.37588    0.08590  16.017   <2e-16 ***
diet3        0.12824    0.12148   1.056    0.294
diet4       -0.06000    0.12148  -0.494    0.623
---
```

```
> contrasts(diet) <- 'contr.helmert'
> summary(lm(weight~diet))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.06365    0.03842  53.718  < 2e-16 ***
diet1        0.36412    0.06074   5.994 5.59e-08 ***
diet2        0.10137    0.03507   2.891  0.00495 **
diet3        0.41819    0.02480  16.864  < 2e-16 ***
diet4        0.22526    0.01921  11.727  < 2e-16 ***
---
```

```
> contrasts(diet) <- 'contr.sum'
> summary(lm(weight~diet))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.06365    0.03842  53.718  < 2e-16 ***
diet1       -1.10894    0.07683 -14.433  < 2e-16 ***
diet2       -0.38071    0.07683  -4.955 3.96e-06 ***
diet3       -0.44071    0.07683  -5.736 1.66e-07 ***
diet4        1.02929    0.07683  13.396  < 2e-16 ***
---
```

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

```
> levels(diet1)[4:5] <- "prot"
> levels(diet1)
[1] "ctrl"   "lipid1" "lipid2" "prot"
> contrasts(diet1) <- 'contr.treatment'
> m2 <- lm(weight~diet1)
```

```
> anova(m1, m2)
Analysis of Variance Table

Model 1: weight ~ diet
Model 2: weight ~ diet1
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     80 10.0357
2     81 10.1755 -1   -0.1398 1.1142 0.2943
```

```
> diet2 <- diet1
> levels(diet2)[2:3] <- "lipid"
> m3 <- lm(weight~diet2)
> anova(m2, m3)
Analysis of Variance Table

Model 1: weight ~ diet1
Model 2: weight ~ diet2
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     81 10.1755
2     82 10.2061 -1   -0.0306 0.2436  0.623
```

```
> diet3 <- diet2
> levels(diet3)[2:3] <- "other"
> m4 <- lm(weight~diet3)
> anova(m3, m4)
Analysis of Variance Table

Model 1: weight ~ diet2
Model 2: weight ~ diet3
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     82  10.206
2     83  42.388 -1   -32.182 258.56 < 2.2e-16 ***
---
```

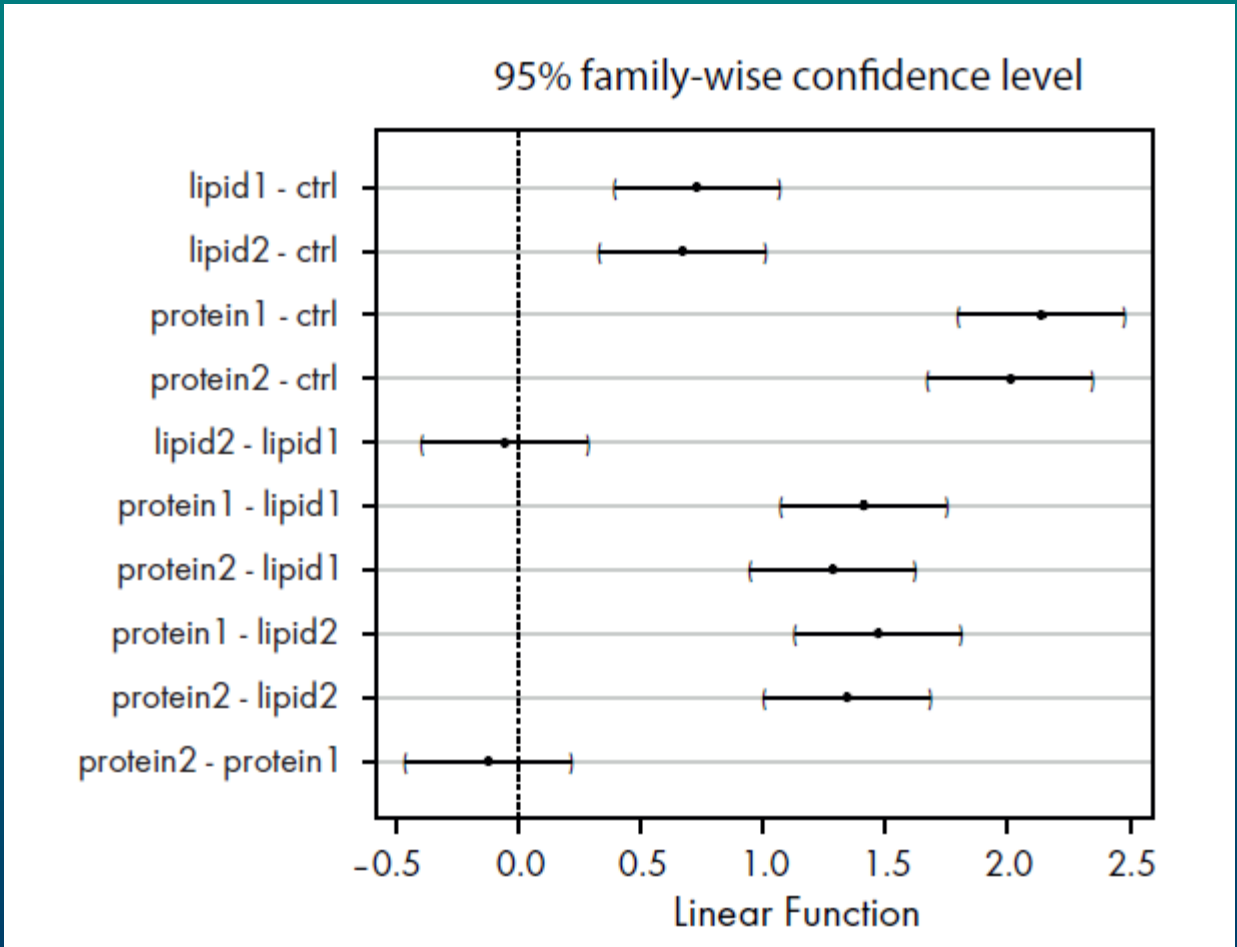
```
> anova(m3,m1)
Analysis of Variance Table

Model 1: weight ~ diet2
Model 2: weight ~ diet
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     82 10.206
2     80 10.036  2   0.17038 0.6791   0.51
```

Multiple comparisons

- *post hoc* tests
- Bonferroni correction – applied to non-orthogonal contrasts
- Dunn test, Scheffe test, Tukey HSD test
- comparison by means of confidence intervals

```
> library(multcomp)
> m4 <- glht(m1, linfct = mcp(diet = "Tukey"))
> summary(m4)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = weight ~ diet)

Linear Hypotheses:
                          Estimate Std. Error t value Pr(>|t|)
lipid1 - ctrl == 0          0.7282     0.1215   5.994   <1e-05 ***
lipid2 - ctrl == 0          0.6682     0.1215   5.501   <1e-05 ***
protein1 - ctrl == 0        2.1382     0.1215  17.601   <1e-05 ***
protein2 - ctrl == 0        2.0100     0.1215  16.545   <1e-05 ***
lipid2 - lipid1 == 0       -0.0600     0.1215  -0.494    0.988
protein1 - lipid1 == 0      1.4100     0.1215  11.606   <1e-05 ***
protein2 - lipid1 == 0      1.2818     0.1215  10.551   <1e-05 ***
protein1 - lipid2 == 0      1.4700     0.1215  12.100   <1e-05 ***
protein2 - lipid2 == 0      1.3418     0.1215  11.045   <1e-05 ***
protein2 - protein1 == 0   -0.1282     0.1215  -1.056    0.828
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
> plot(m4)
```

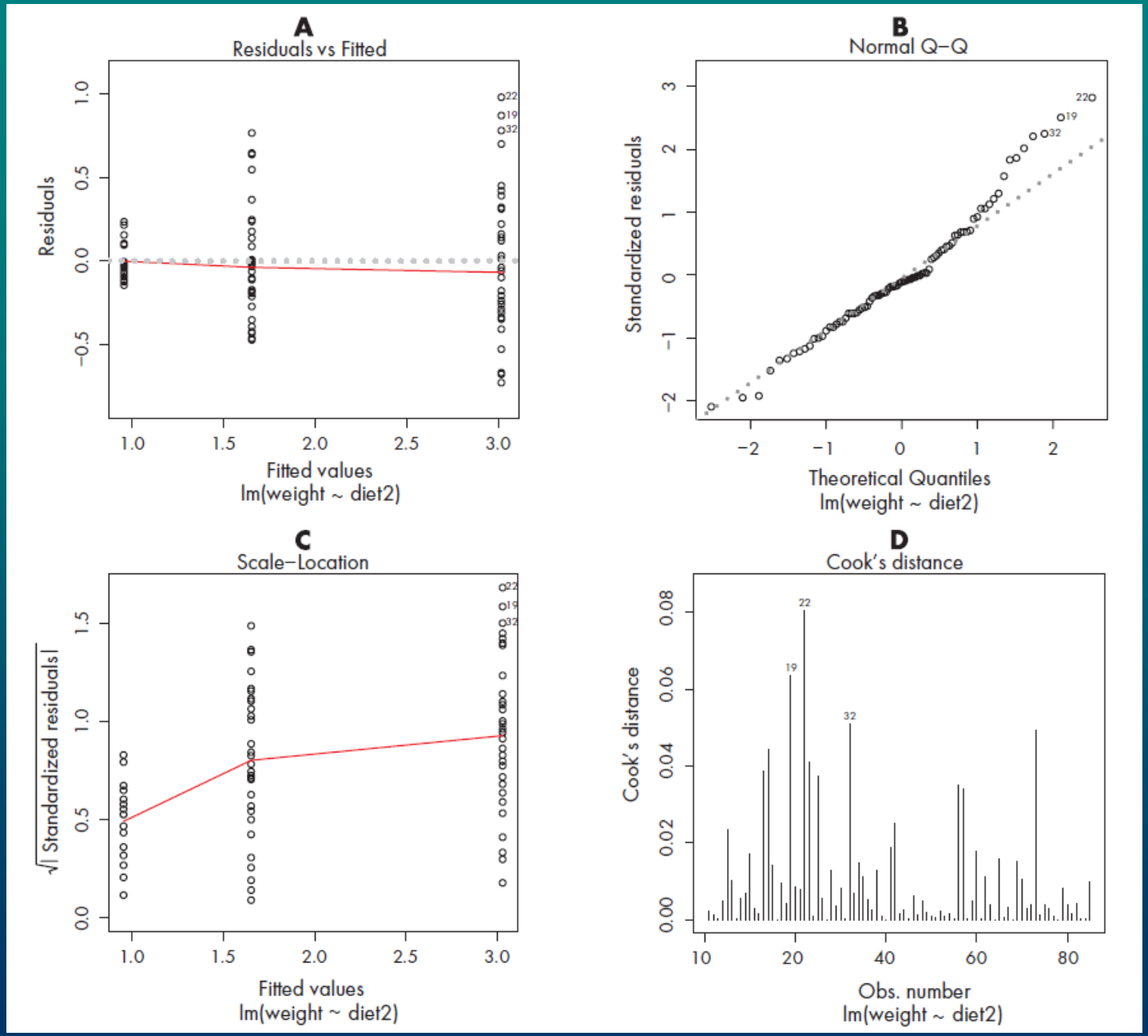

95% family-wise confidence level
lipid1 - ctrl
lipid2 - ctrl
protein1 - ctrl
protein2 - ctrl
lipid2 - lipid1
protein1 - lipid1
protein2 - lipid1
protein1 - lipid2
protein2 - lipid2
protein2 - protein1
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
Linear Function

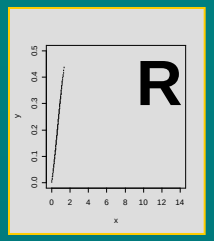
Diagnosis:

We should check as many aspects as possible
- use diagnostic plots
- use formal tests:
- Bartlett test to compare variances
- Shapiro-Wilk test of normality

```
> bartlett.test(weight ~ diet2)

        Bartlett test of homogeneity of variances

data:  weight by diet2
Bartlett's K-squared = 24.2178, df = 2, p-value = 5.51e-06
```

```
> shapiro.test(resid(m3))

        Shapiro-Wilk normality test

data:  resid(m9)
W = 0.9685, p-value = 0.0356
```


A
Residuals vs Fitted
Residuals
Fitted values
lm(weight ~ diet2)
B
Normal Q–Q
Standardized residuals
Theoretical Quantiles
lm(weight ~ diet2)
C
Scale–Location
√| Standardized residuals |
Fitted values
lm(weight ~ diet2)
D
Cook's distance
Cook's distance
Obs. number
lm(weight ~ diet2)



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 variables

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

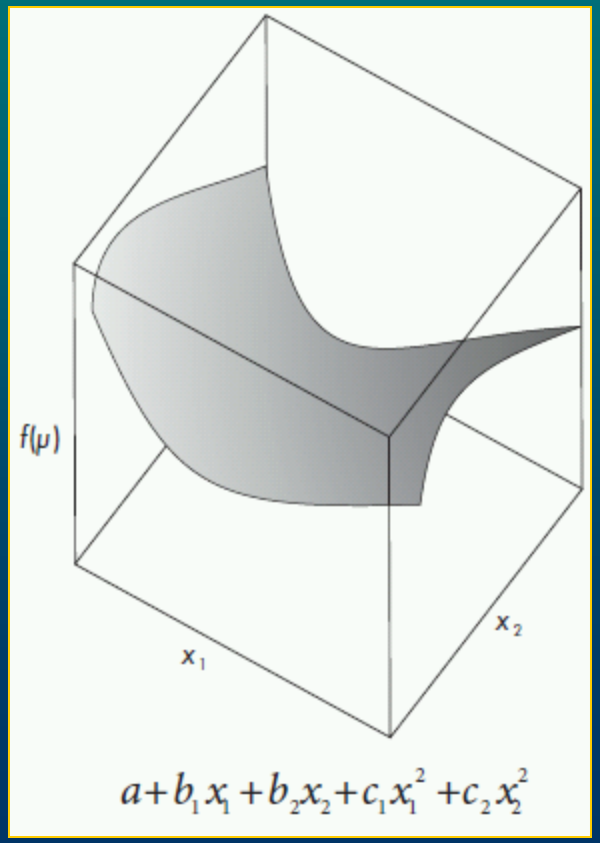
• linear predictor with 2 explanatory variables (x_1, x_2) should include all main effects, all interactions, and quadratic terms

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 x_1^2 + \gamma_2 x_2^2 + \delta x_1 x_2$$

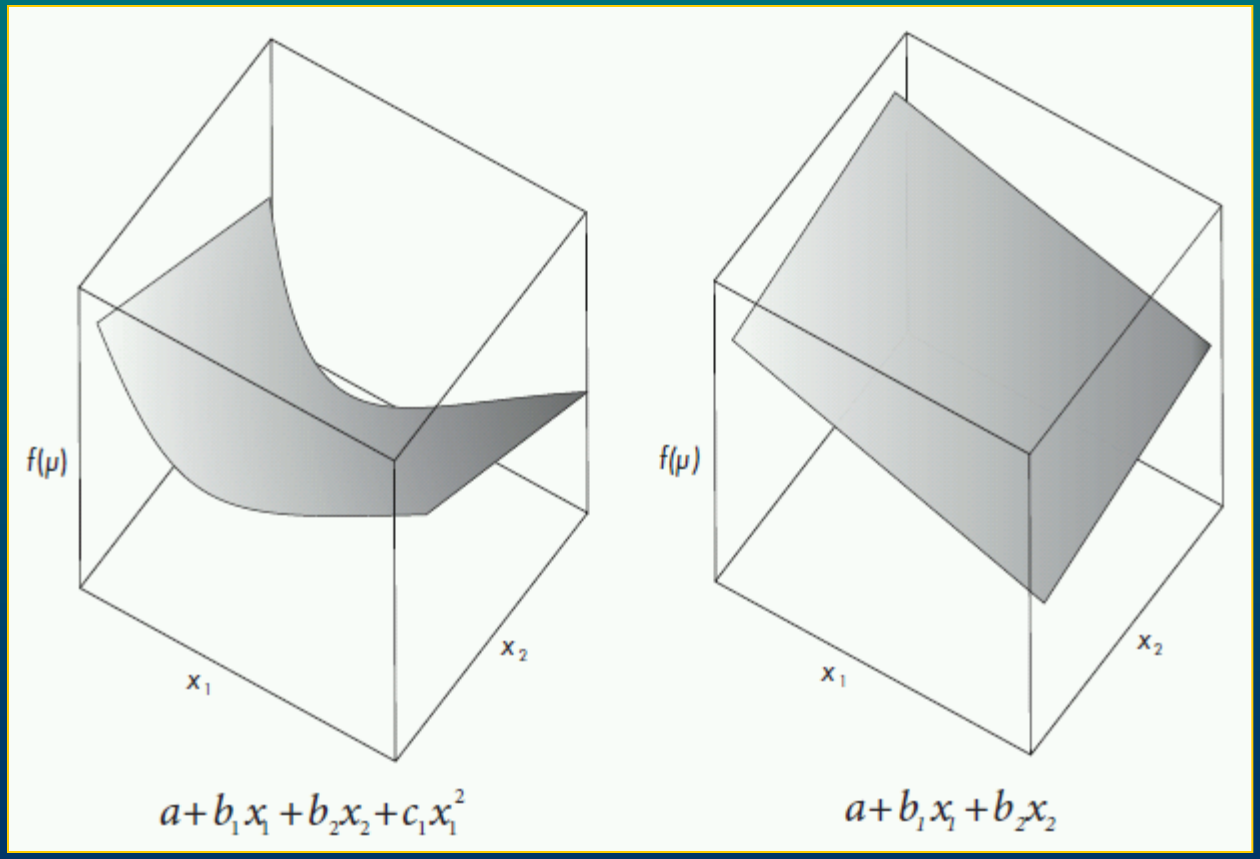
with estimates a, b_1, b_2, c_1,c_2, d

Nested models are:

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

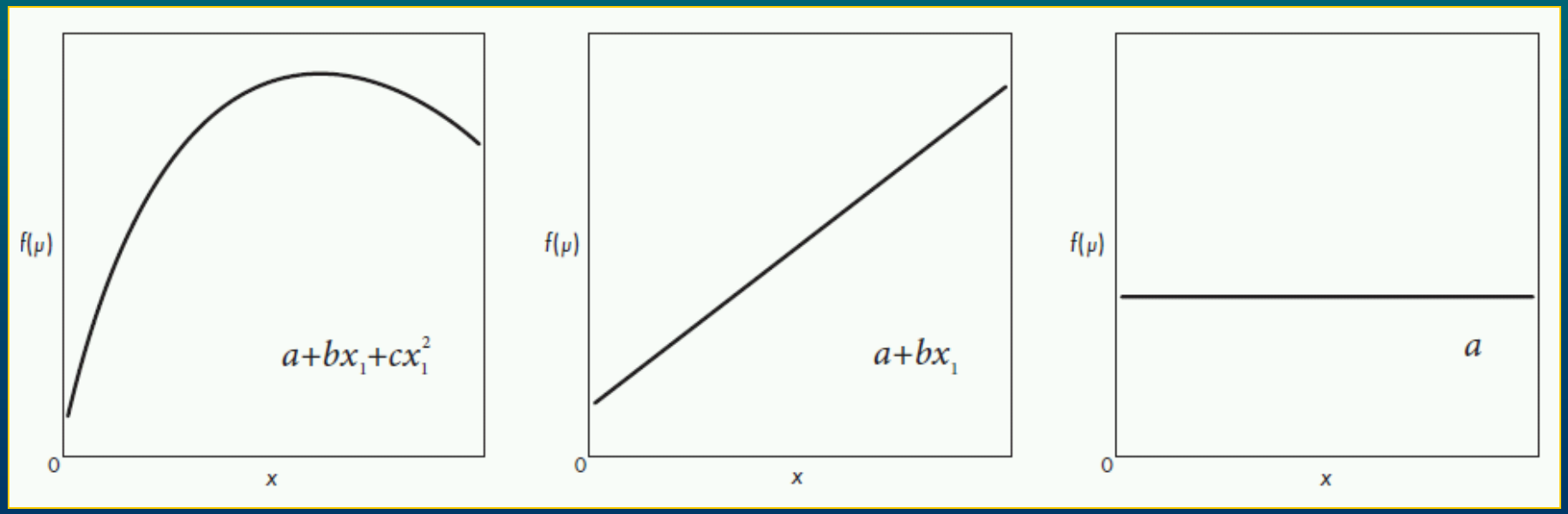


- 4 parameters (a, b_1, b_2, c_1), at least c_1 is significantly different
- 3 parameters (a, b_1, b_2), at least b_1 and b_2 are significantly different



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

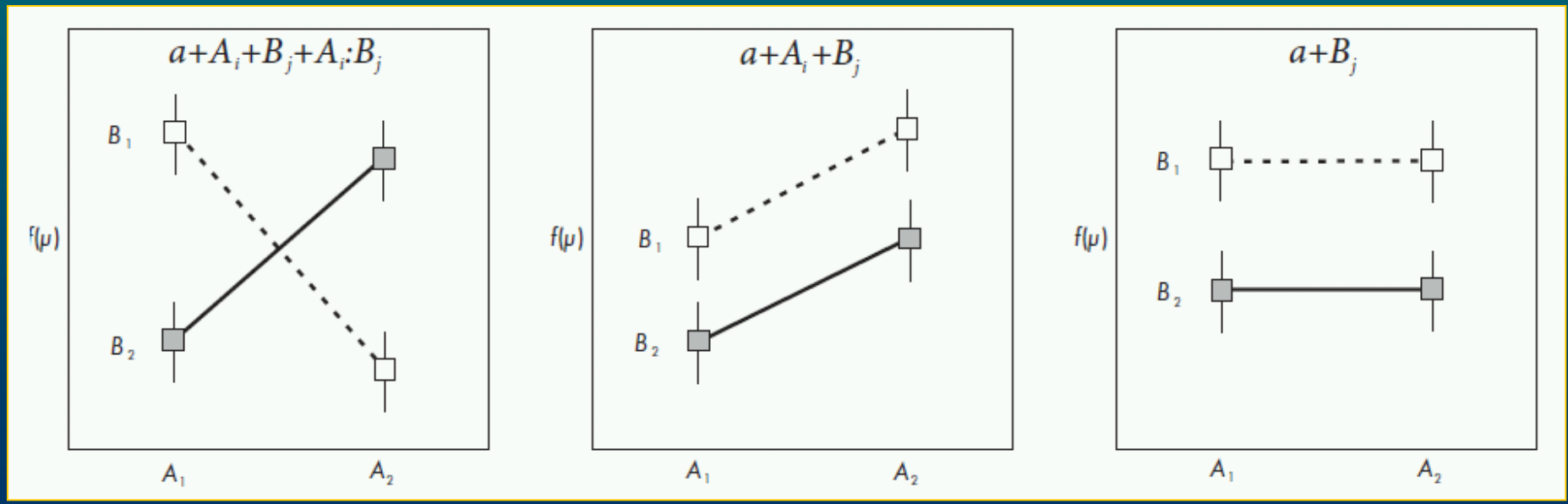
- k-way ANOVA might be with our without interactions

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:

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

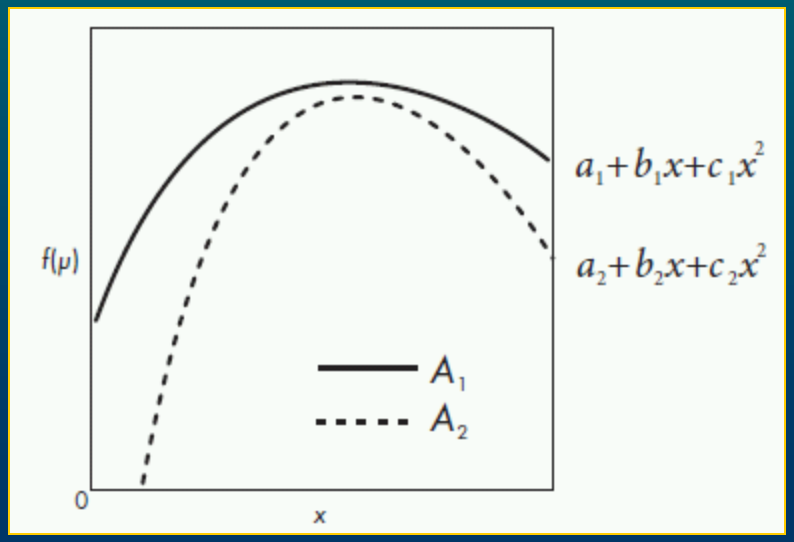
α .. intercept, A_j .. effect of factor, β .. slope, δ .. effect of interaction

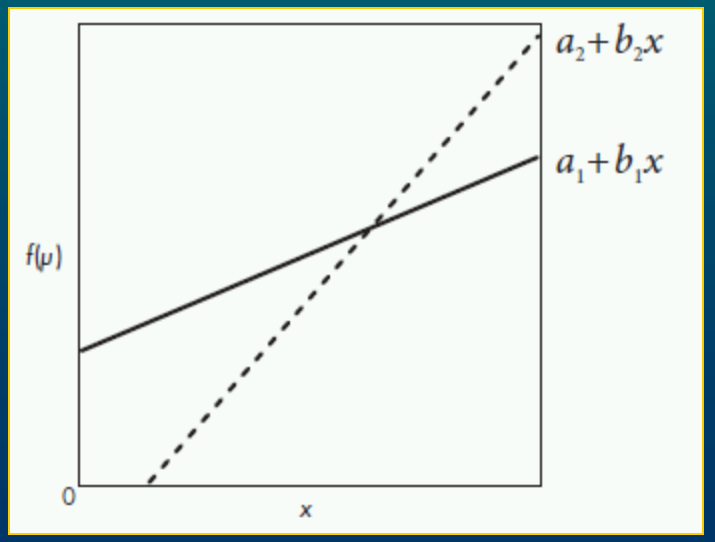
Given 1 categorical variable A with 2 levels (A_1, A_2) and 1 continual x, the linear predictor will be

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

• 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

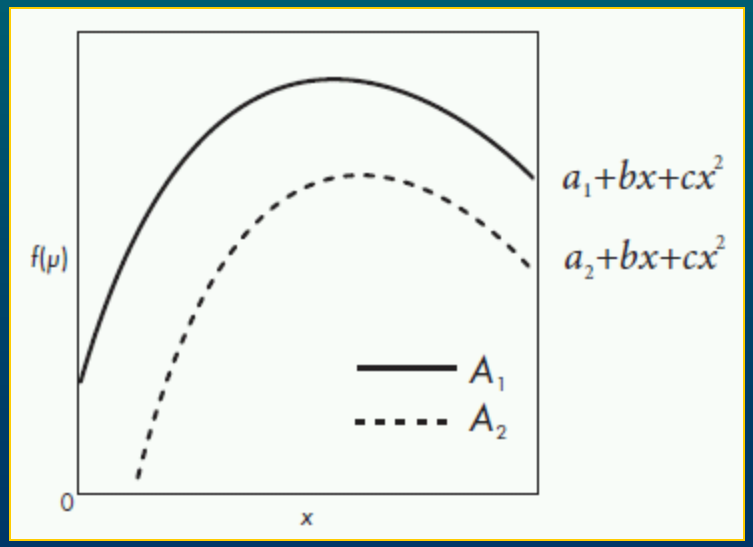


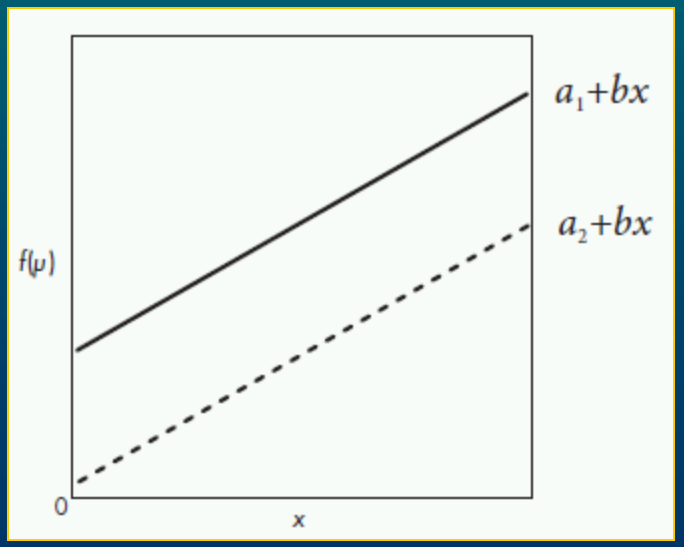


• 4 parameters - 2 intercepts (a_1, a_2–a_1), 1 slope (b), 1 quadratic (c) - interactions A:x^2 and A:x are not significant, but A and quadratic terms are significant

• 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+A:B +A:C+B:C+A:B:C`	

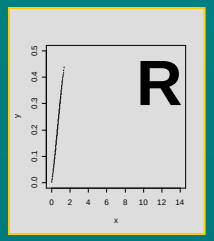
$$f(\mu_{ijk}) = \alpha + A_i + B_j + C_k + A:B_{ij} + A:C_{ik} + B:C_{jk} + A:B:C_{ijk}$$

Model formula	Description
`y ~ (A+B+C)^2`	3-way ANOVA with only three 2-way interactions

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

Model formula	Description
`y ~ x*A`	1-way ANCOVA

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



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

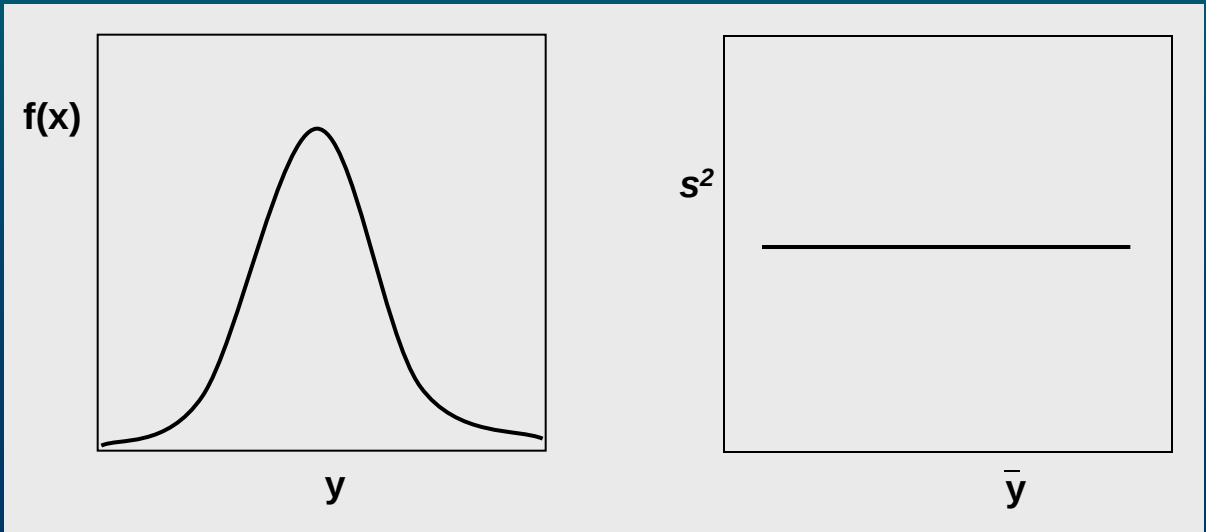
Continuous measurements

- measurements that can be made with infinite precision

Gauss (normal) distribution

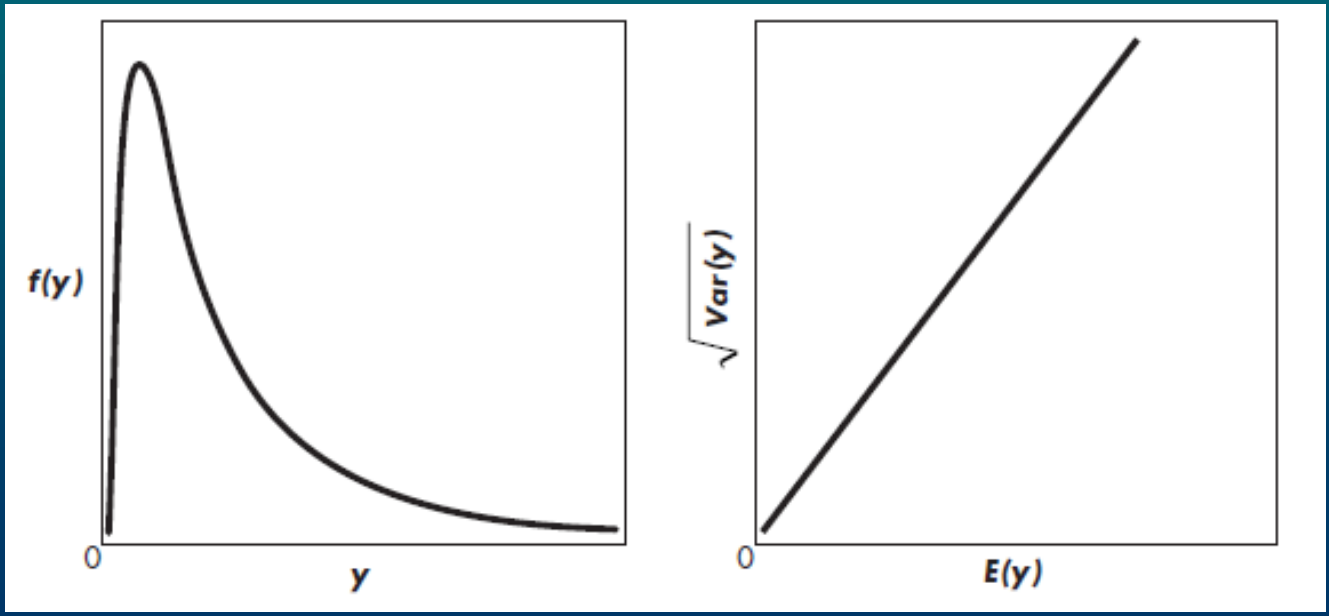
- bell-shaped, symmetric around mean
- mean = median = modus
- parameters: μ, σ^2

- s^2 is independent of mean



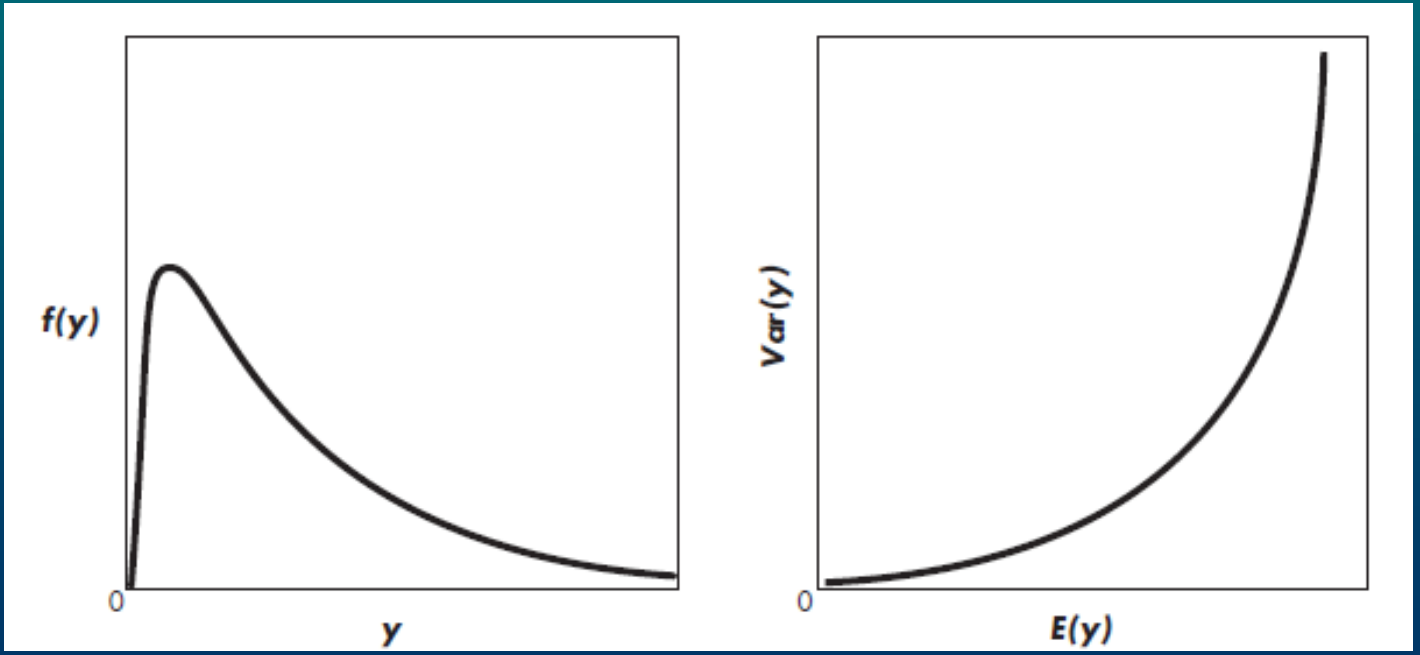
Lognormal distribution

- positive real values
- asymmetric, skewed to the right
- variance increases with mean at quadratic trend
- after logarithmic transformation variances are similar



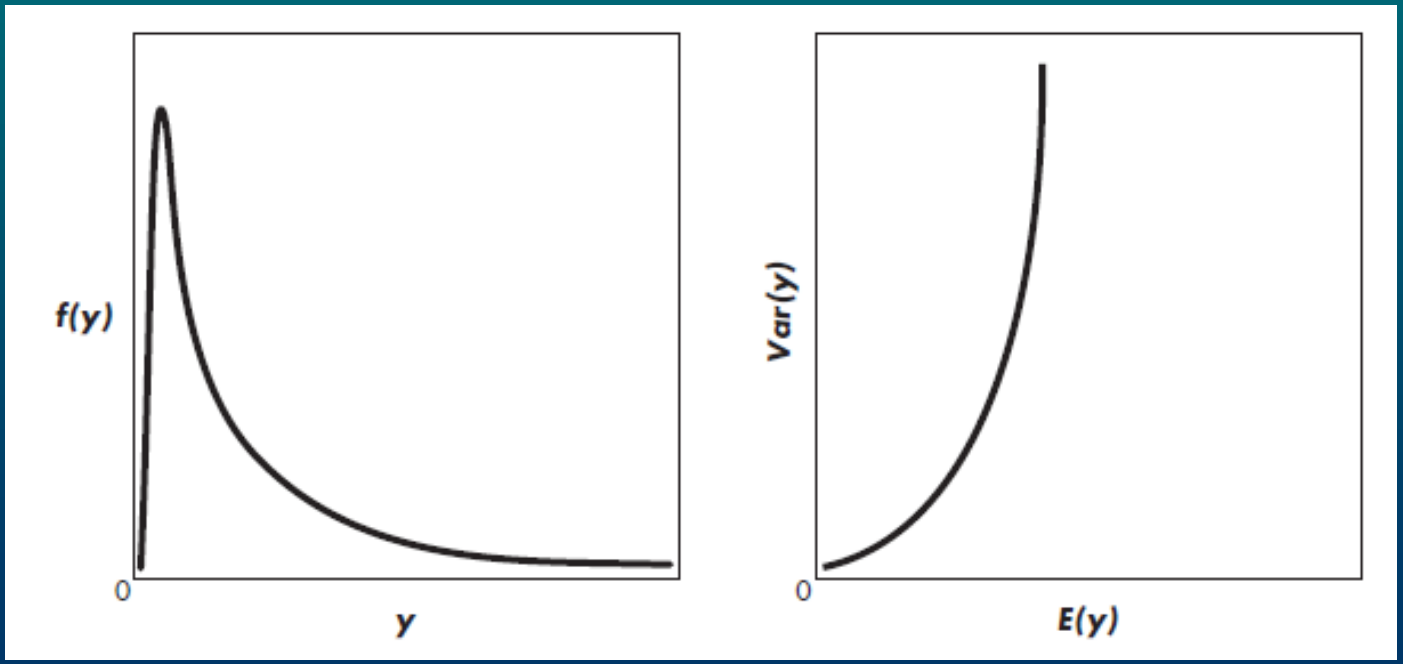
Gamma distribution

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



Inverse Gaussian distributions

- positive real values
- used to model diffusion processes, dispersion in ecology
- variance increases steeply with mean

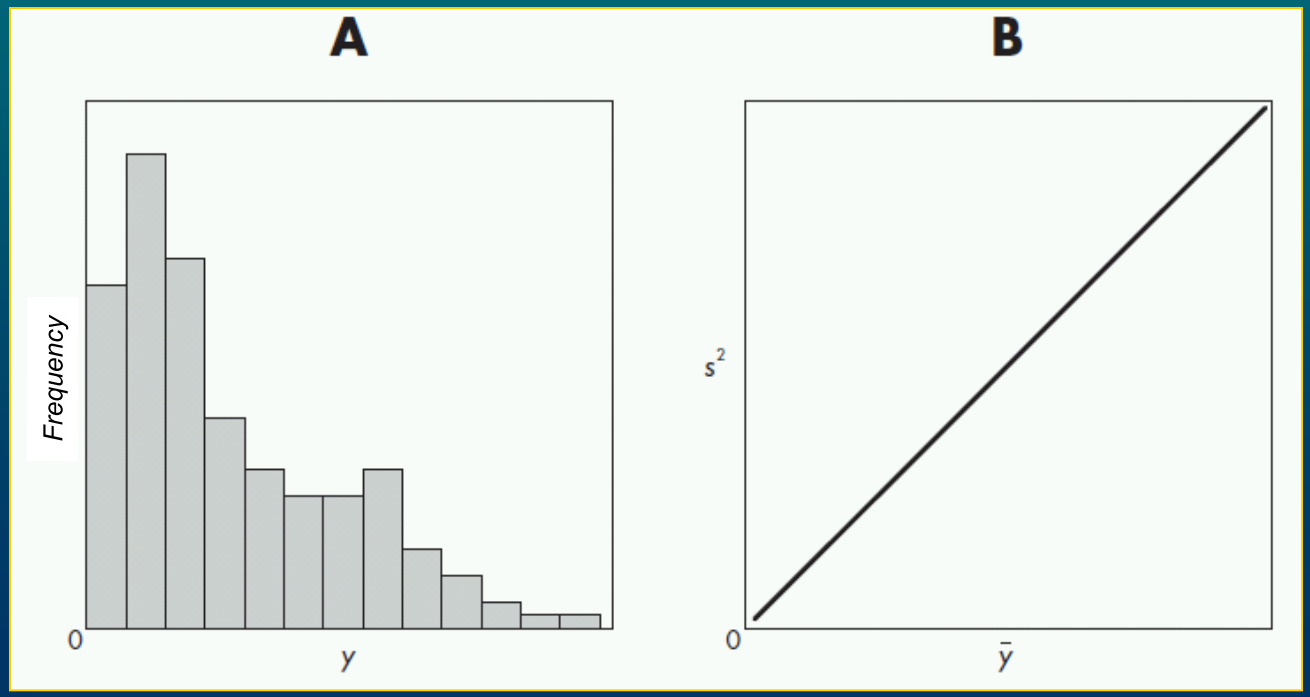


Counts

Poisson distribution

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

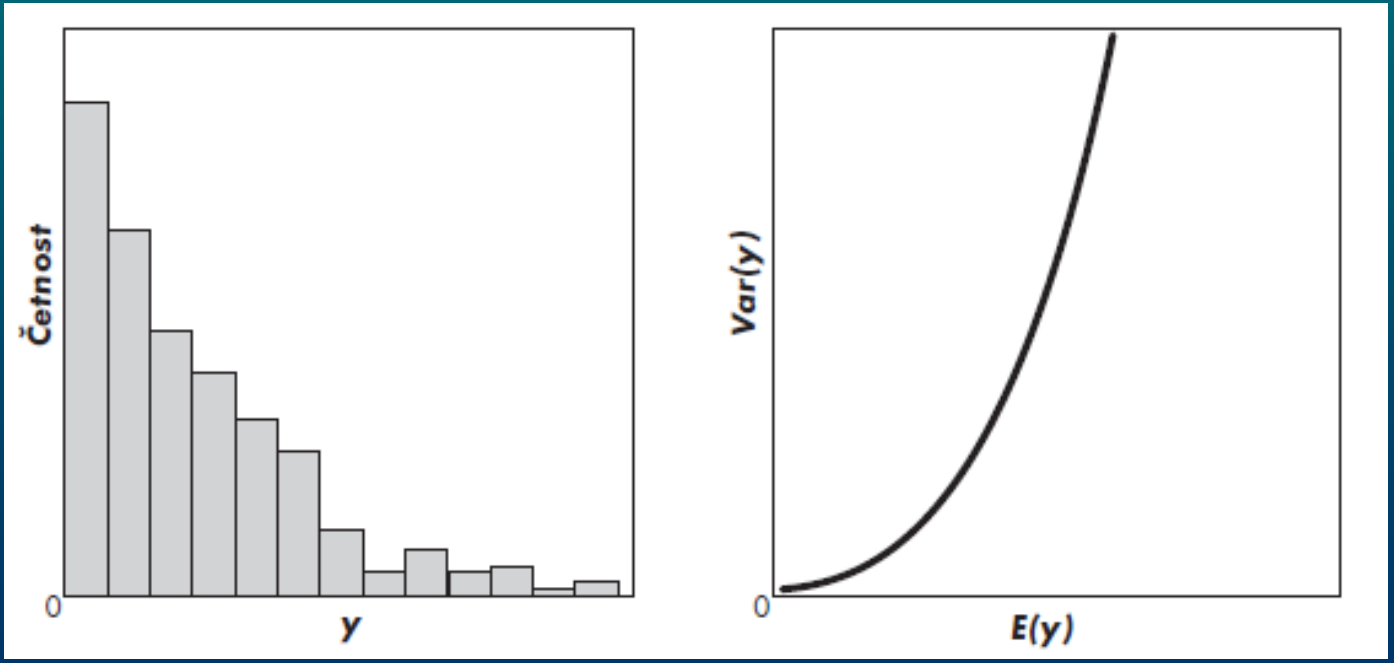
- variance increases with mean



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



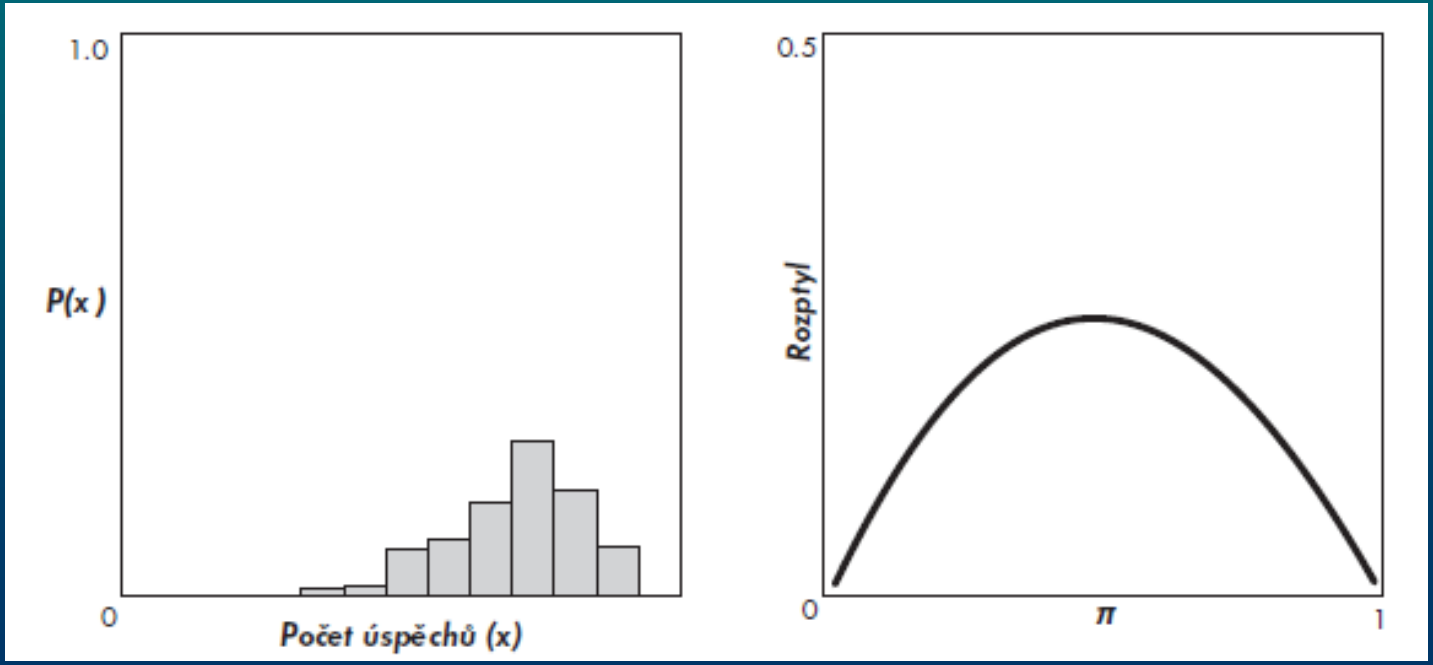
Proportions

- 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

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

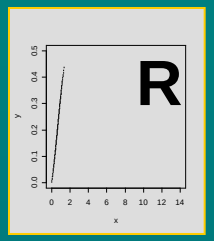
Binomial & Binary distributions

- 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



Quasi “distribution”

- is not any distribution
- specifies expected value and the relationship between expected value and variance
- mixture of available settings



Analyses of continuous I

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

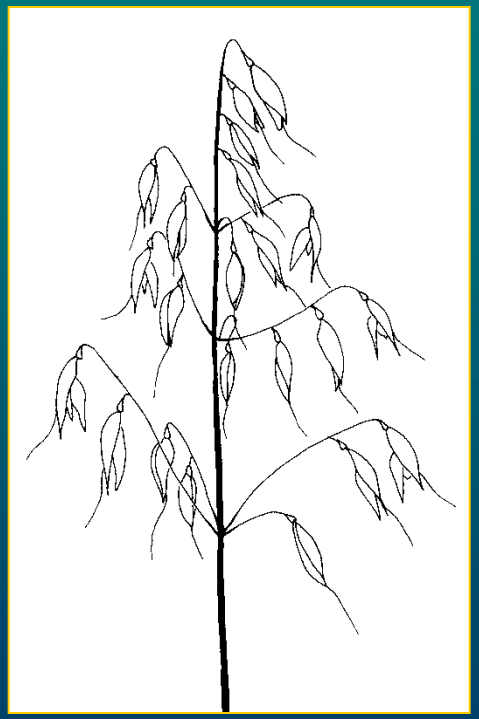
- **t-test** (`t.test`) to compare one or two means
- **Linear model** (`lm`) to study effect of categorical and continuous variables
  - inference is exact, reliable for each *n*
- **GLM** (`glm`) to study effect of categorical and continuous variables
  - Gaussian family (default)
  - link: identity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

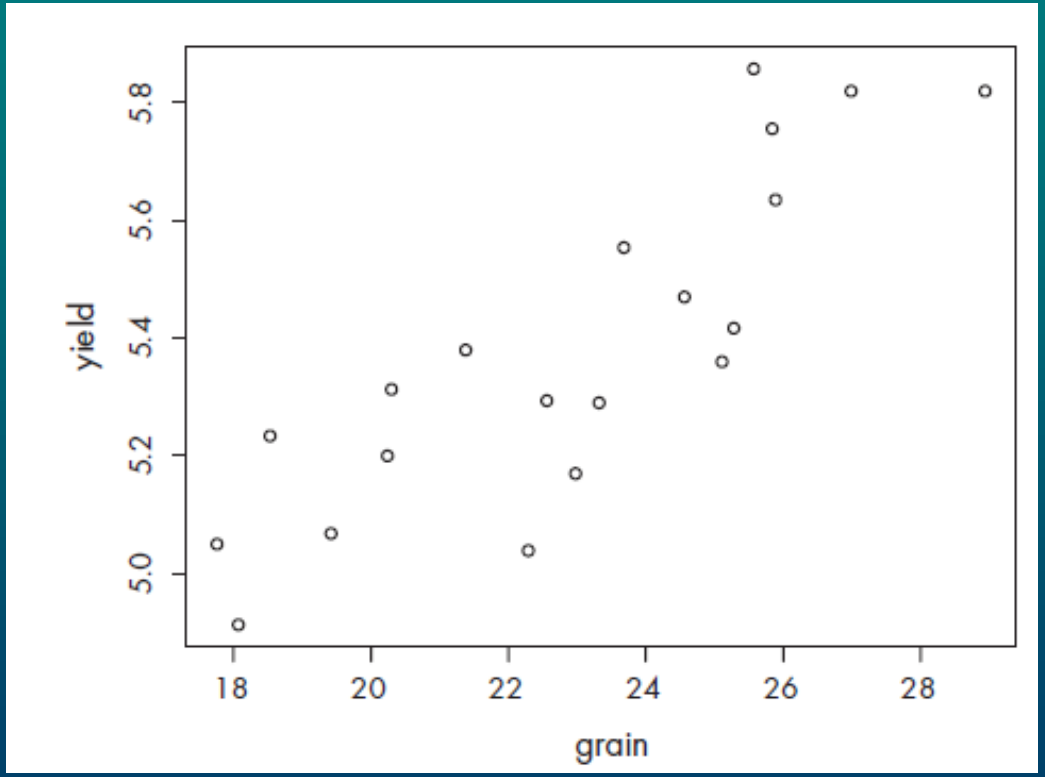
Is number of seeds related to the yield?

What is the predictive model of this relationship?

Variables

grain

yield



$$yield_i = \alpha + \beta grain_i + \varepsilon_i ,$$

kde $\varepsilon_i \sim N(0, \sigma^2)$, nezávisle pro jednotlivé plochy.

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 non-monotonous trend
• use either `poly(x,2)` or `x + I(x^2)`

$$yield_i = \alpha + \beta grain_i + \gamma grain_i^2 + \varepsilon_i ,$$

kde $\varepsilon_i \sim N(0, \sigma^2)$, nezávisle pro jednotlivé plochy.

```
> m1 <- lm(yield ~ poly(grain,2))
> summary(m1)

Call:
lm(formula = yield ~ poly(grain, 2))

Residuals:
      Min        1Q    Median        3Q       Max
-0.261562 -0.121112 -0.003686  0.142558  0.281990

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.38205    0.03417 157.520  < 2e-16 ***
poly(grain, 2)1  1.04875    0.15280   6.864 2.75e-06 ***
poly(grain, 2)2  0.15416    0.15280   1.009    0.327
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1528 on 17 degrees of freedom
Multiple R-Squared: 0.739,      Adjusted R-squared: 0.7083
F-statistic: 24.06 on 2 and 17 DF,  p-value: 1.101e-05
```

```
> summary(lm(yield ~ grain + I(grain^2)))

…

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.427559   1.795165   3.023  0.00766 **
grain       -0.083870   0.159018  -0.527  0.60472
I(grain^2)   0.003507   0.003476   1.009  0.32718
```

```
> m2 <- lm(yield ~ grain)
> summary(m2)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.63509    0.25694   14.15 3.42e-11 ***
grain        0.07617    0.01110    6.86 2.03e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1529 on 18 degrees of freedom
Multiple R-Squared: 0.7233,     Adjusted R-squared: 0.708
F-statistic: 47.06 on 1 and 18 DF,  p-value: 2.033e-06
```

$$yield_i = \beta grain_i + \varepsilon_i,$$

kde $\varepsilon_i \sim N(0, \sigma^2)$, nezávisle pro jednotlivé plochy.

```
> m3 <- update(m2, ~.-1)
> summary(m3)
...
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
grain 0.231855   0.005006   46.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.518 on 19 degrees of freedom
Multiple R-Squared: 0.9912,     Adjusted R-squared: 0.9908
F-statistic:  2146 on 1 and 19 DF,  p-value: < 2.2e-16
```

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`**)

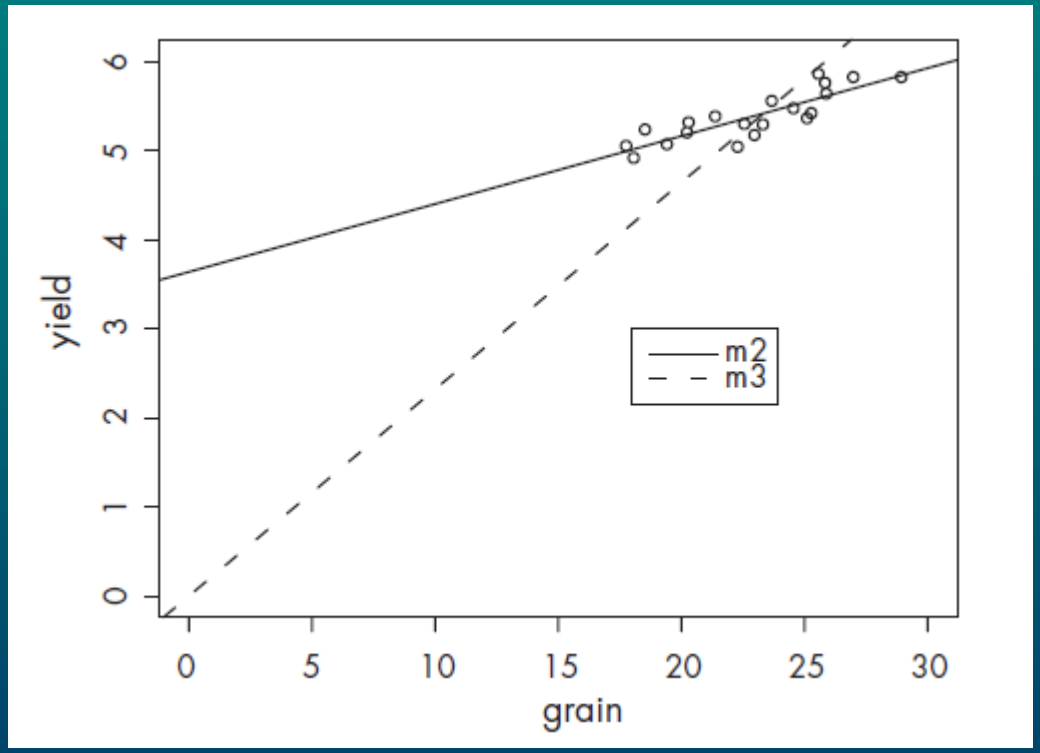
• **Akaike Information Criterion** (AIC):

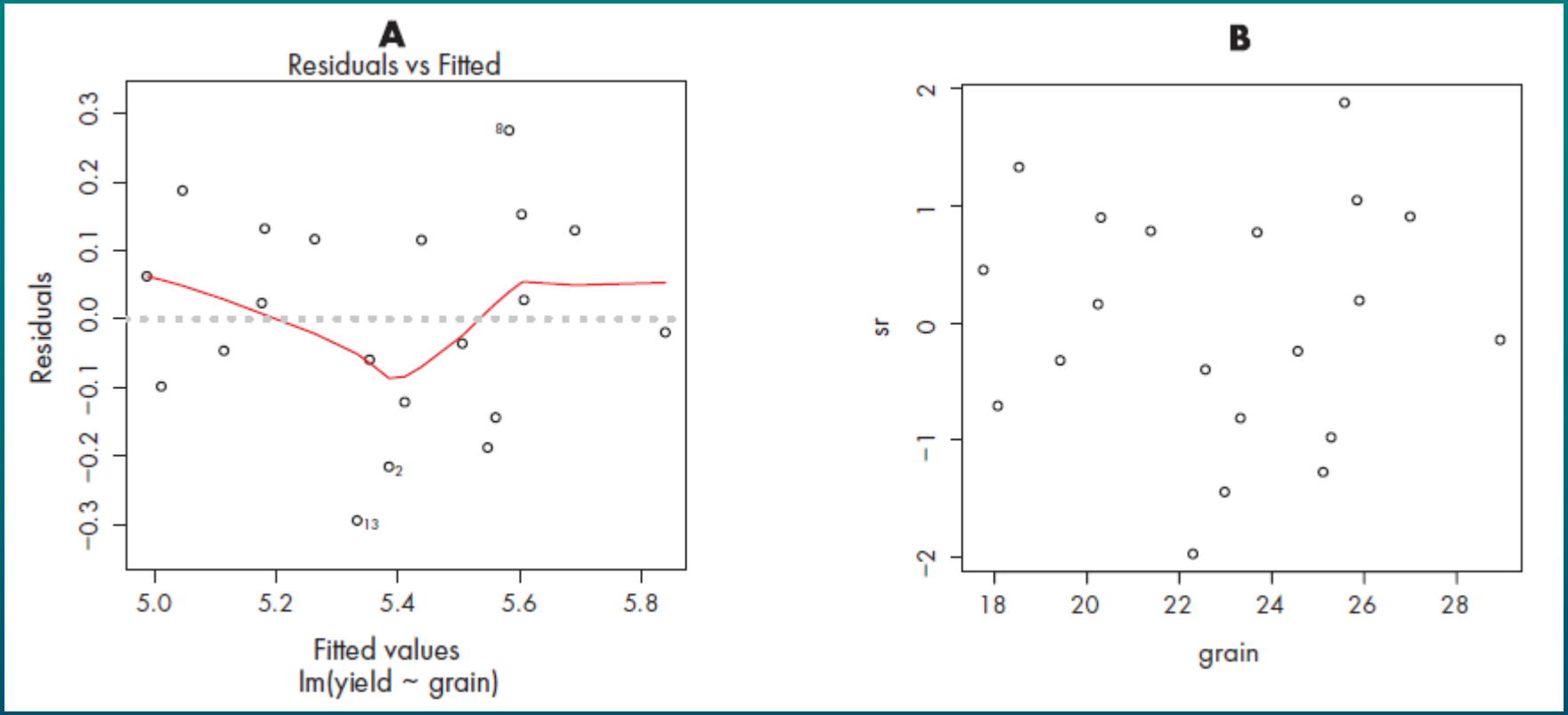
$$\mathrm{AIC} = -2LogLik + 2p$$

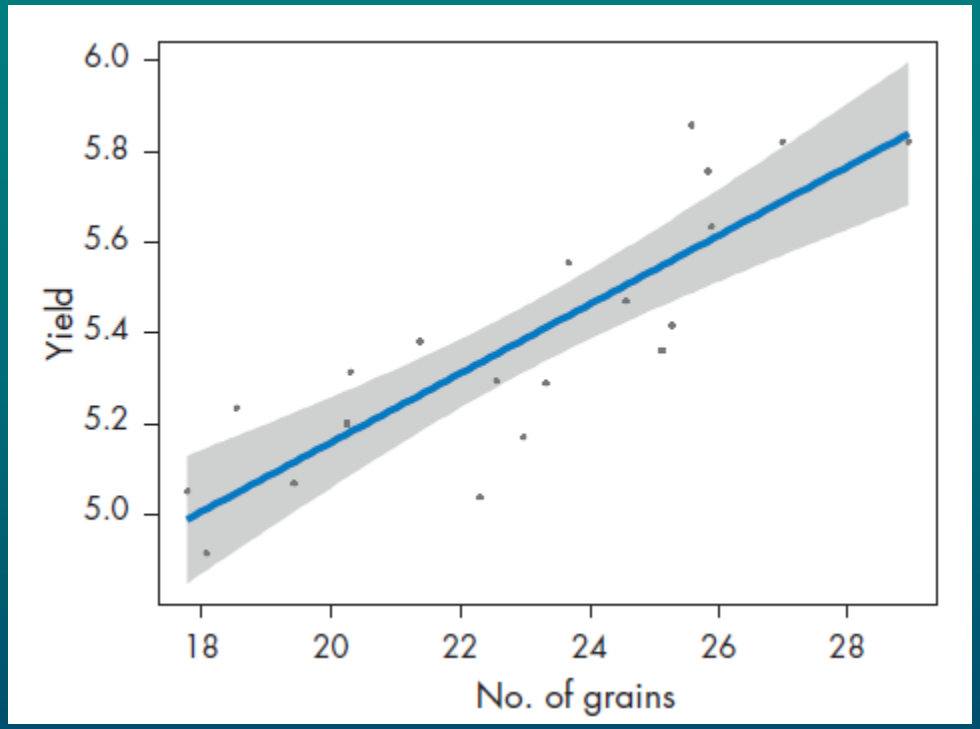
- the more there are parameters in the model the better fit but worse explanatory power of the model

- the lower AIC the better model

```
> AIC(m2, m3)
    df       AIC
m2   3 -14.47461
m3   2  33.42234
```




A
Residuals vs Fitted
Residuals
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
8
2
13
5.0
5.2
5.4
5.6
5.8
Fitted values
lm(yield ~ grain)
B
sr
2
1
0
−1
−2
18
20
22
24
26
28
grain

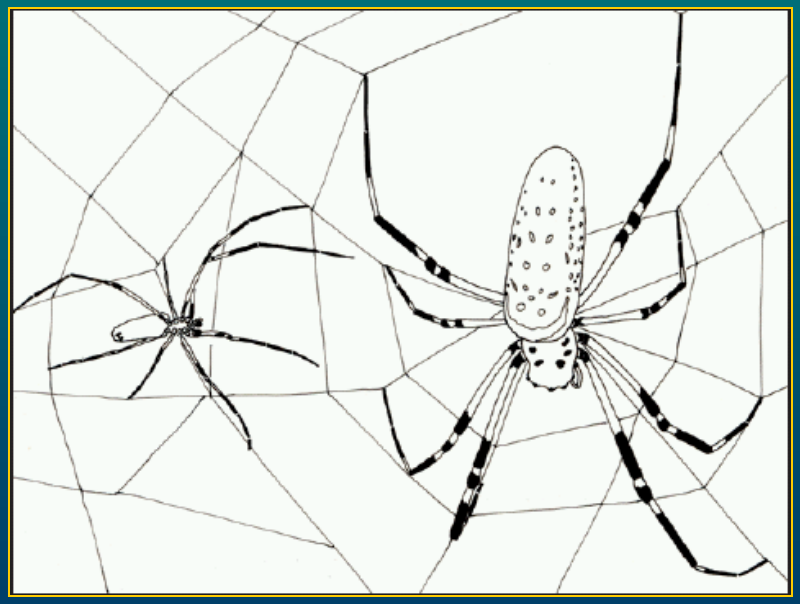

6.0
5.8
5.6
5.4
5.2
5.0
Yield
18
20
22
24
26
28
No. of grains

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

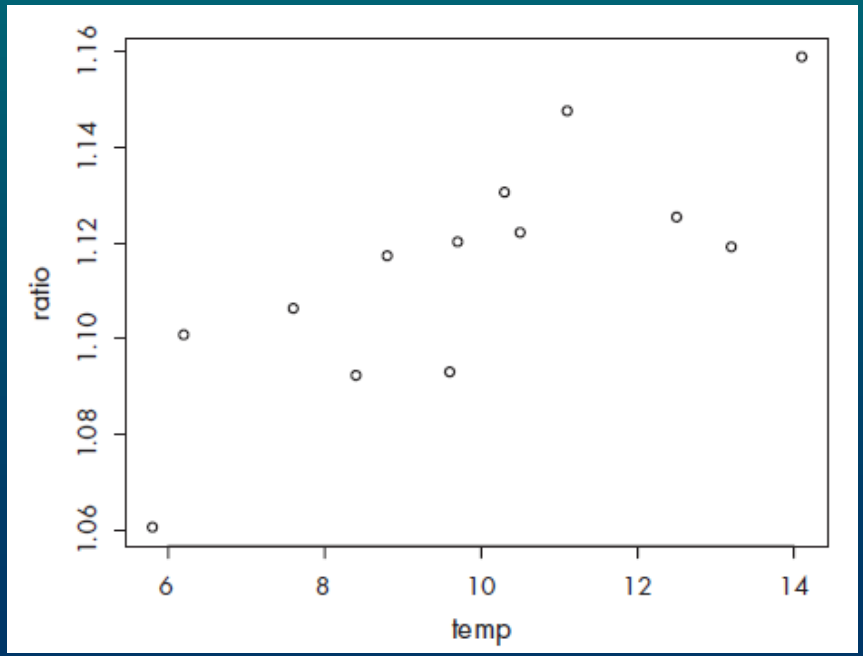
Males and females of *Zodarion* spiders were sampled on 13 sites with a different temperature [ºC]. Of the average size of males and females a size ratio was calculated for each site. The number of individuals varied between sites (2 to 62 specimens).

Hypotheses

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

Variables

temp
number
ratio



$$ratio_i = \alpha + \beta temp_i + \gamma temp_i^2 + \varepsilon_i,$$

kde $\varepsilon_i \sim N(0, \sigma^2)$, nezávisle pro různé lokality.

```
> m1 <- lm(ratio ~ poly(temp,2))
> summary(m1)
...
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.114961   0.004597 242.538  < 2e-16 ***
poly(temp, 2)1  0.069484   0.016575   4.192  0.00185 **
poly(temp, 2)2 -0.010728   0.016575  -0.647  0.53206
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01657 on 10 degrees of freedom
Multiple R-Squared: 0.6428,     Adjusted R-squared: 0.5713
F-statistic: 8.996 on 2 and 10 DF,  p-value: 0.005818
```

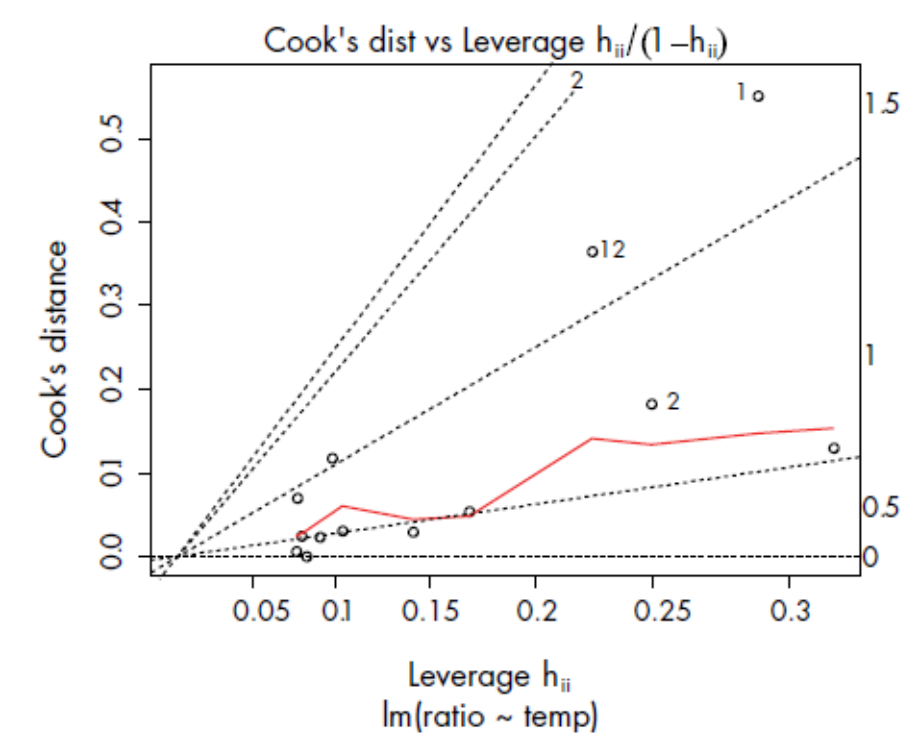
```
> m2 <- lm(ratio ~ temp)
```

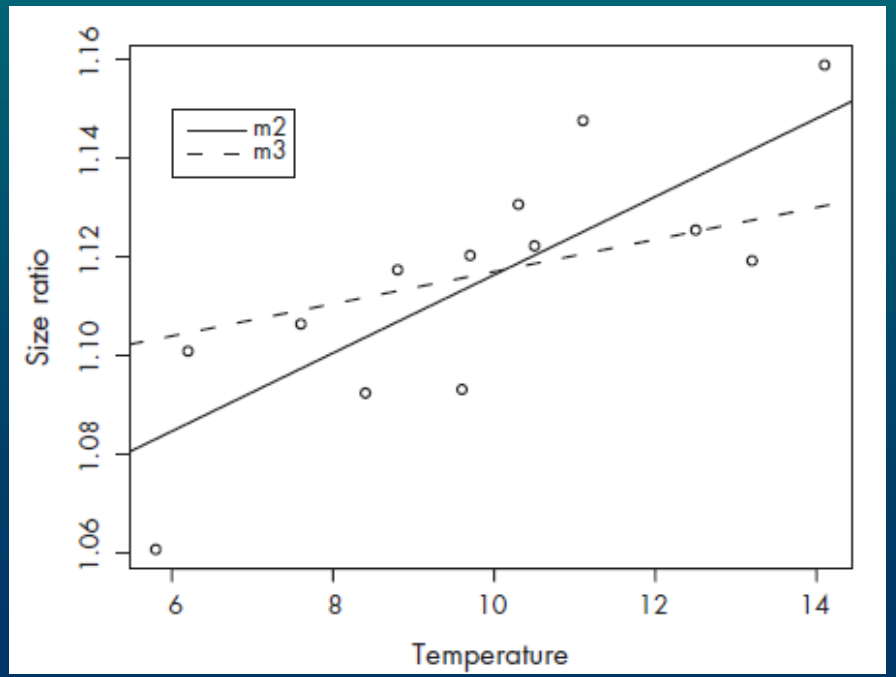
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.036897   0.018667  55.547 7.94e-15 ***
temp        0.007941   0.001843   4.307  0.00124 **
---
```

$$ratio_i = \alpha + \beta temp_i + \varepsilon_i,$$

$$\varepsilon_i \sim N\left(0, \frac{\sigma^2}{number}\right), \text{ nezávisle pro různé lokality.}$$

```
> m3 <- update(m2, weights=number)
> summary(m3)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.084297   0.015481  70.038 6.24e-16 ***
temp        0.003265   0.001510   2.162   0.0535 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04727 on 11 degrees of freedom
Multiple R-Squared: 0.2982,     Adjusted R-squared: 0.2345
F-statistic: 4.675 on 1 and 11 DF,  p-value: 0.0535
```

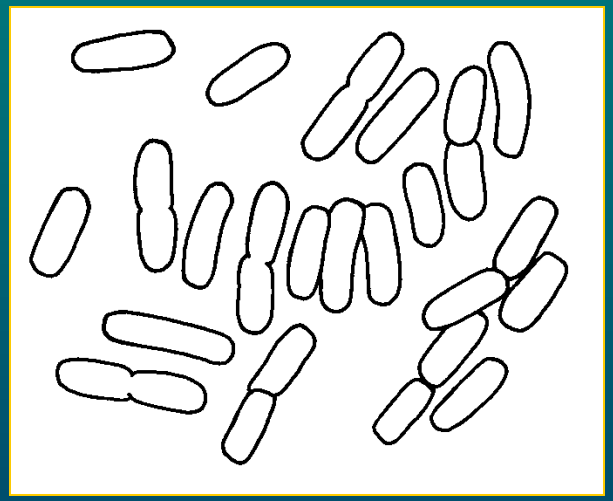

Cook's dist vs Leverage $h_{ii}/(1-h_{ii})$
Cook's distance
Leverage h_{ii}
lm(ratio ~ temp)
0.0
0.1
0.2
0.3
0.4
0.5
0.05
0.1
0.15
0.2
0.25
0.3
0.5
1
1.5
1
2
12
2


Size ratio
Temperature
m2
m3
1.06
1.08
1.10
1.12
1.14
1.16
6
8
10
12
14

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?

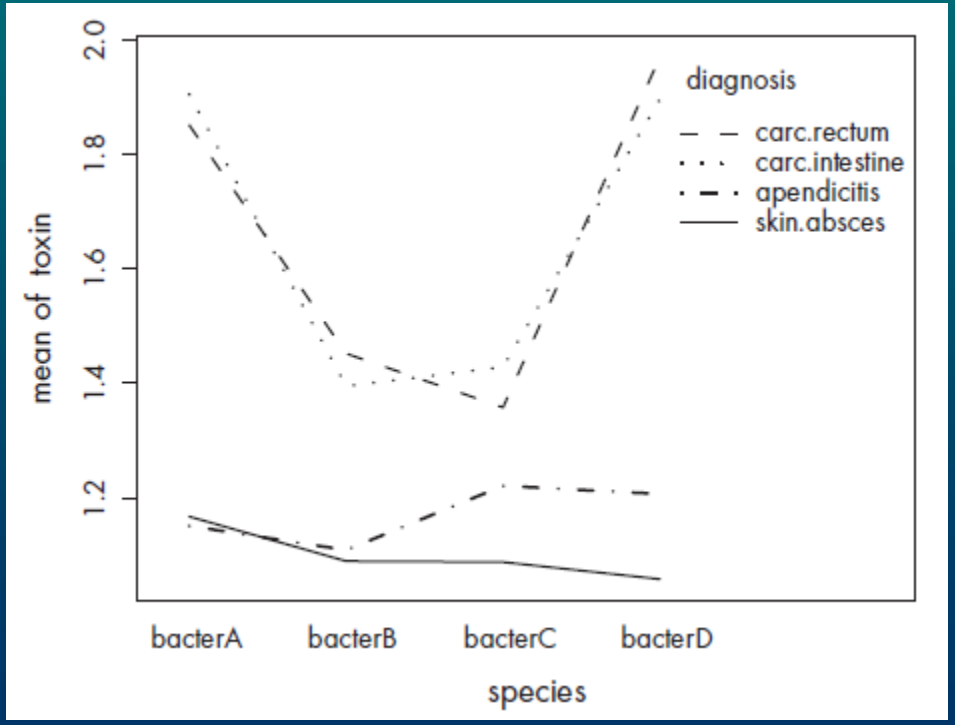
Which species can be used as an indicator?

Variables

SPECIES: bacterA, bacterB, bacterC, bacterD

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

toxin



$$toxin_{ijk} = \alpha + SPECIES_j + DIAGNOSIS_k + SPECIES{:}DIAGNOSIS_{jk} + \varepsilon_{ijk},$$

kde $\varepsilon_{ijk} \sim N(0, \sigma^2)$, nezávisle pro jednotlivé pacienty.

```
> m1 <- lm(toxin ~ species*diagnosis)
> anova(m1)
Analysis of Variance Table

Response: toxin
                  Df Sum Sq Mean Sq  F value    Pr(>F)
species            3 1.3364  0.4455  28.0325 1.077e-11 ***
diagnosis          3 5.4775  1.8258 114.8965 < 2.2e-16 ***
species:diagnosis  9 1.2704  0.1412   8.8827 1.528e-08 ***
Residuals         64 1.0170  0.0159
```

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

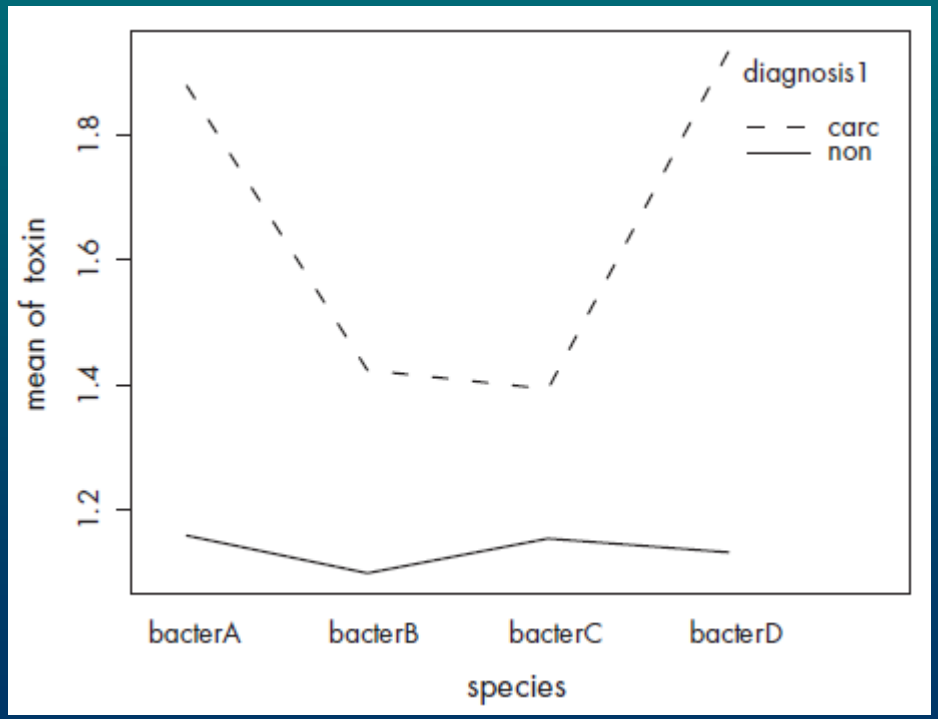
Orthogonality
• independent variables are orthogonal - effects are straightforward
• correlated variables are unorthogonal - effects are complicated
- when there are missing values or unequal number of observations *per* treatment

```
> summary(m1)
...
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                1.15100    0.05638  20.417  < 2e-16
speciesbacterB                            -0.04220    0.07973  -0.529 0.598427
speciesbacterC                             0.07000    0.07973   0.878 0.383233
speciesbacterD                             0.05580    0.07973   0.700 0.486536
diagnosiscarc.intestine                    0.75400    0.07973   9.457 9.07e-14
diagnosiscarc.rectum                       0.70040    0.07973   8.785 1.34e-12
diagnosisskin.absces                       0.01640    0.07973   0.206 0.837678
speciesbacterB:diagnosiscarc.intestine    -0.46760    0.11275  -4.147 0.000101
speciesbacterC:diagnosiscarc.intestine    -0.54620    0.11275  -4.844 8.42e-06
speciesbacterD:diagnosiscarc.intestine    -0.06420    0.11275  -0.569 0.571083
speciesbacterB:diagnosiscarc.rectum       -0.35700    0.11275  -3.166 0.002366
speciesbacterC:diagnosiscarc.rectum       -0.56340    0.11275  -4.997 4.78e-06
speciesbacterD:diagnosiscarc.rectum        0.06300    0.11275   0.559 0.578282
speciesbacterB:diagnosisskin.absces       -0.03580    0.11275  -0.318 0.751889
speciesbacterC:diagnosisskin.absces       -0.14960    0.11275  -1.327 0.189287
speciesbacterD:diagnosisskin.absces       -0.16520    0.11275  -1.465 0.147771
```

```
> tapply(predict(m1), list(species,diagnosis), mean)
        apendicitis carc.intestine carc.rectum skin.absces
bacterA      1.1510         1.9050      1.8514       1.1674
bacterB      1.1088         1.3952      1.4522       1.0894
bacterC      1.2210         1.4288      1.3580       1.0878
bacterD      1.2068         1.8966      1.9702       1.0580
```

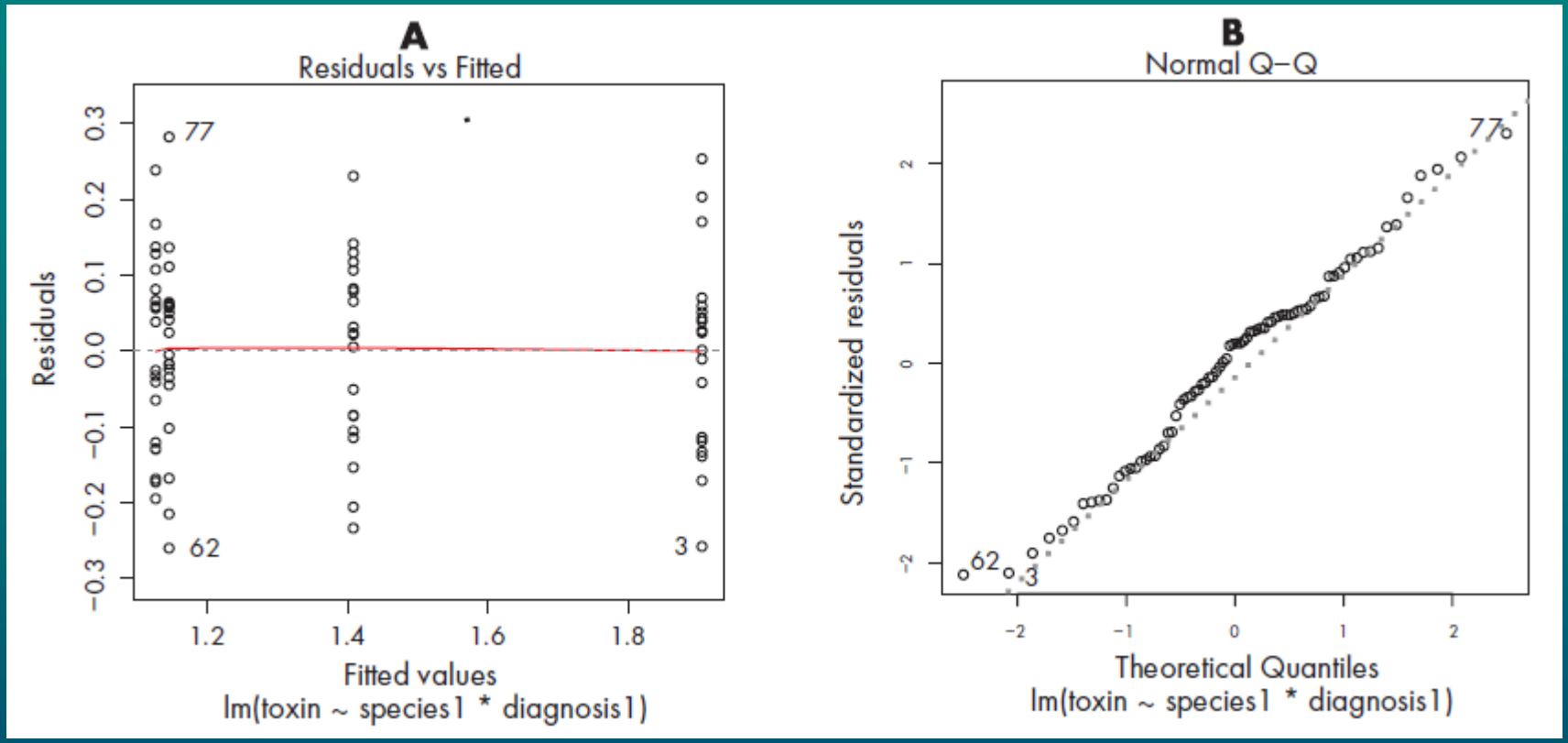
```
> diagnosis1 <- c(rep("carc",40), rep("non",40))
> diagnosis1 <- factor(diagnosis1)
```

```
> m2 <- lm(toxin ~ species*diagnosis1)
> anova(m1, m2)
Analysis of Variance Table

Model 1: toxin ~ species * diagnosis
Model 2: toxin ~ species * diagnosis1
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     64  1.0170
2     72  1.1597 -8   -0.1427 1.1225 0.3605
```



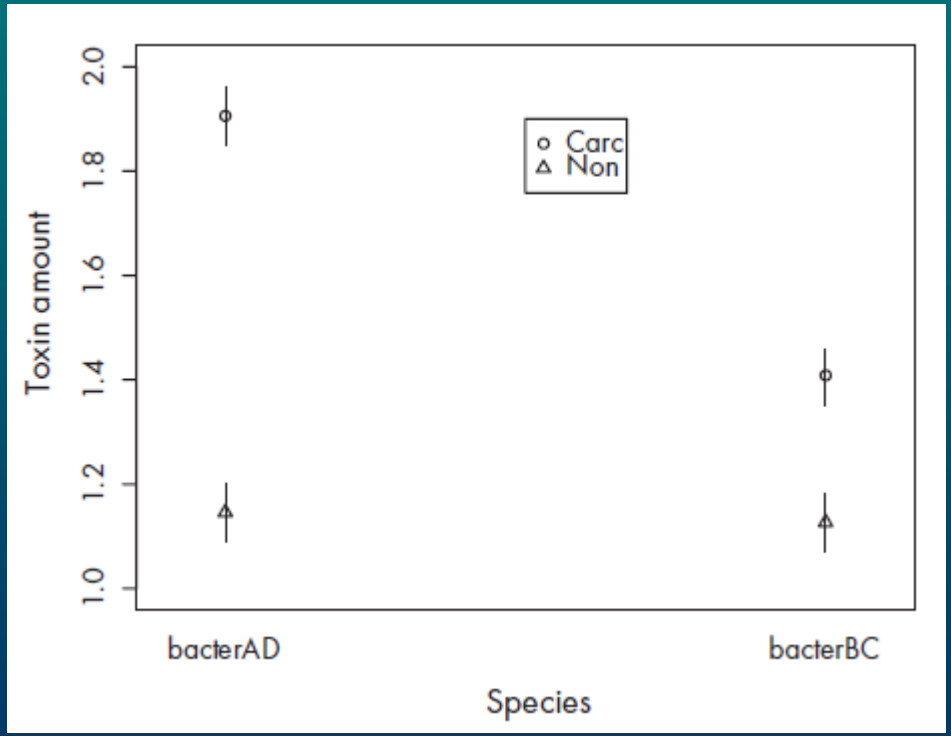
```
> species1 <- species
> levels(species1)
[1] "bacterA" "bacterB" "bacterC" "bacterD"
> levels(species1)[2:3] <- "bacterBC"
> m3 <- lm(toxin ~ species1*diagnosis1)
> anova(m2, m3)
Analysis of Variance Table

Model 1: toxin ~ species * diagnosis1
Model 2: toxin ~ species1 * diagnosis1
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     72  1.15974
2     74  1.17962 -2  -0.01988 0.6171 0.5423
> levels(species1)
[1] "bacterA"  "bacterBC" "bacterD"
> levels(species1)[c(1,3)] <- "bacterAD"
> m4 <- lm(toxin ~ species1*diagnosis1)
> anova(m3, m4)
Analysis of Variance Table

Model 1: toxin ~ species1 * diagnosis1
Model 2: toxin ~ species1 * diagnosis1
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     74  1.17962
2     76  1.19845 -2  -0.01883 0.5905 0.5566
```

```
> anova(m4)
Analysis of Variance Table

Response: toxin
                   Df Sum Sq Mean Sq F value    Pr(>F)
species1            1 1.3328  1.3328  84.522 5.690e-14 ***
diagnosis1          1 5.4267  5.4267 344.139 < 2.2e-16 ***
species1:diagnosis1 1 1.1434  1.1434  72.508 1.134e-12 ***
Residuals          76 1.1984  0.0158
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(m4)
...
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.90580    0.02808  67.872  < 2e-16 ***
species1bacterBC              -0.49725    0.03971 -12.522  < 2e-16 ***
diagnosis1non                 -0.76000    0.03971 -19.139  < 2e-16 ***
species1bacterBC:diagnosis1non  0.47820    0.05616   8.515 1.13e-12 ***
```




```
> both <- paste(species1, diagnosis1)
> both <- factor(both)
> m5 <- lm(toxin ~ both - 1)
> summary(m5)
...
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
bothbacterAD carc   1.90580    0.02808   67.87   <2e-16 ***
bothbacterAD non    1.14580    0.02808   40.81   <2e-16 ***
bothbacterBC carc   1.40855    0.02808   50.16   <2e-16 ***
bothbacterBC non    1.12675    0.02808   40.13   <2e-16 ***
```

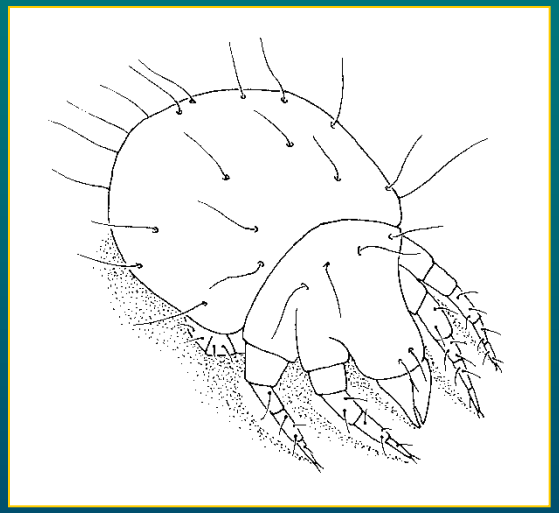
```
> confint(m5)
                          2.5%    97.5%
bothbacterAD carc 1.849875 1.961725
bothbacterAD non  1.089875 1.201725
bothbacterBC carc 1.352625 1.464475
bothbacterBC non  1.070825 1.182675
```



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?

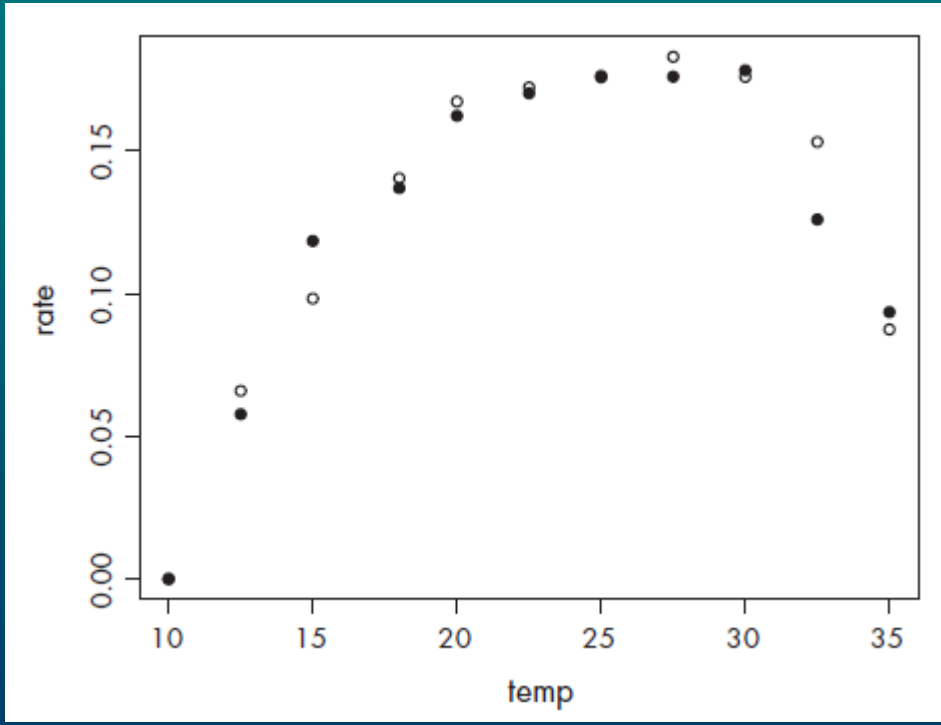
What is the model of the relationship?

Variables

GENUS: genA, genB

temp

rate



$$rate_{ij} = \alpha + GENUS_j + \beta temp_i + \gamma temp_i^2 + \tau \mathrm{temp}_i^3 + \delta_j temp_i + \omega_j temp_i^2 + \eta_j temp_i^3 + \varepsilon_{ij},$$

kde $\varepsilon_{ij} \sim N(0, \sigma^2)$, nezávisle pro jednotlivé populace. (8-12)

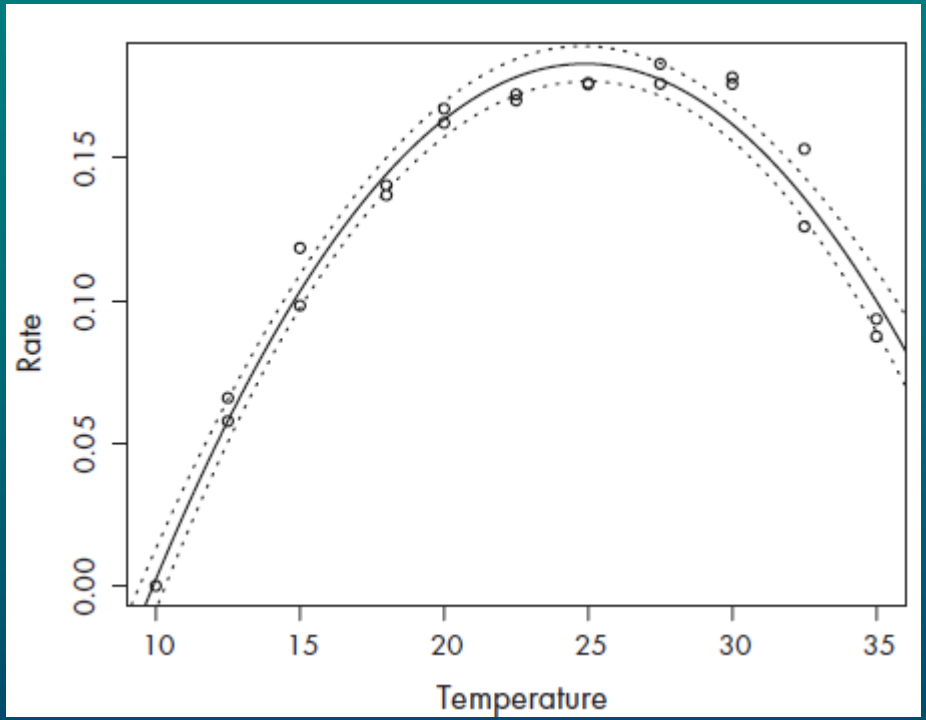
```
> m1 <- lm(rate ~ poly(temp,3)*genus)
> anova(m1)
Analysis of Variance Table

Response: rate
                     Df   Sum Sq  Mean Sq  F value    Pr(>F)
poly(temp, 3)         3 0.065953 0.021984 210.0675 7.125e-12 ***
genus                 1 0.000028 0.000028   0.2644    0.6152
poly(temp, 3):genus   3 0.000108 0.000036   0.3454    0.7930
Residuals            14 0.001465 0.000105
---
```

```
> m2 <- lm(rate ~ poly(temp,3)+genus)
> anova(m1, m2)
```

```
  Res.Df         RSS Df   Sum of Sq      F Pr(>F)
1     14  0.00146516
2     17  0.00157360 -3 -0.00010844 0.3454  0.793
> anova(m2)
Analysis of Variance Table

Response: rate
              Df   Sum Sq  Mean Sq  F value    Pr(>F)
poly(temp, 3)  3 0.065953 0.021984 237.5038 4.509e-14 ***
genus          1 0.000028 0.000028   0.2989    0.5917
Residuals     17 0.001574 0.000093
---
```

```
> m3 <- lm(rate ~ temp + I(temp^2) + I(temp^3))
> summary(m3)

...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.675e-01  5.138e-02  -5.205 5.97e-05 ***
temp         3.191e-02  7.855e-03   4.063  0.00073 ***
I(temp^2)   -3.986e-04  3.704e-04  -1.076  0.29608
I(temp^3)   -6.178e-06  5.464e-06  -1.131  0.27309
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009432 on 18 degrees of freedom
Multiple R-Squared: 0.9763,     Adjusted R-squared: 0.9723
F-statistic: 247.1 on 3 and 18 DF,  p-value: 8.234e-15
```

```
> m4 <- lm(rate ~ temp + I(temp^2))
> summary(m4)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.222e-01  1.734e-02  -18.59 1.20e-13 ***
temp         4.060e-02  1.662e-03   24.42 8.20e-16 ***
I(temp^2)   -8.154e-04  3.649e-05  -22.35 4.20e-15 ***
---
```


Rate
0.00
0.05
0.10
0.15
10
15
20
25
30
35
Temperature

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

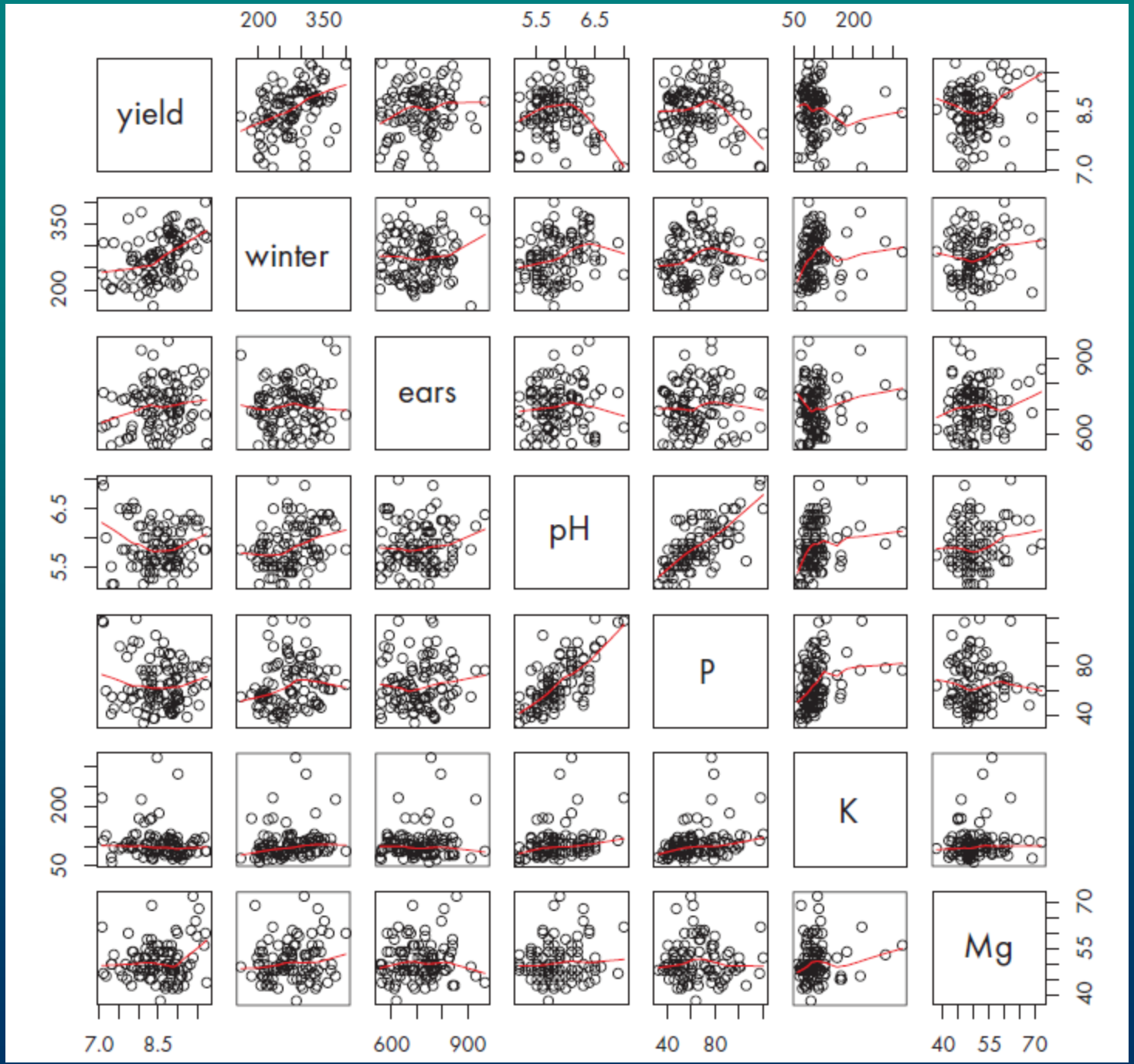
P

K

Mg

yield

```
> pairs(yield ~ winter + ears + pH + P + K + Mg, panel=panel.smooth)
```


yield
winter
ears
pH
P
K
Mg
200
350
5.5
6.5
50
200
7.0
8.5
600
900
40
80
40
55
70

Collinearity

- When two or more variables show correlation
- PCA can reduce dimensionability of variables – use PCA scores instead

```
> pca <- princomp(~ pH + P,cor=T)
> summary(pca)
Importance of components:
                          Comp.1    Comp.2
Standard deviation     1.3049566 0.5450579
Proportion of Variance 0.8514559 0.1485441
Cumulative Proportion  0.8514559 1.0000000
```

```
> PpH <- pca$scores[,1]
```

$$yield_i = \alpha + \beta_1 winter_i + \beta_2 ears_i + \beta_3 PpH_i + \beta_4 K_i + \beta_5 Mg_i + \varepsilon_i ,$$

$$\varepsilon_i \sim N(0, \sigma^2), \text{ nezávisle pro různé parcely.}$$

```
> m1 <- lm(yield ~ (winter+ears+PpH+K+Mg)^2 + I(winter^2) + I(ears^2) +
+ I(PpH^2) + I(K^2) + I(Mg^2))
> anova(m1)

Analysis of Variance Table

Response: yield
             Df  Sum Sq Mean Sq F value    Pr(>F)
winter        1  6.5802  6.5802 30.7489 3.764e-07 ***
ears          1  0.7288  0.7288  3.4059  0.068712 .
PpH           1  2.1751  2.1751 10.1643  0.002053 **
K             1  0.8568  0.8568  4.0039  0.048830 *
Mg            1  0.3331  0.3331  1.5568  0.215821
I(winter^2)   1  0.0765  0.0765  0.3576  0.551530
I(ears^2)     1  0.2315  0.2315  1.0818  0.301467
I(PpH^2)      1  5.1354  5.1354 23.9977 5.029e-06 ***
I(K^2)        1  0.5878  0.5878  2.7470  0.101404
I(Mg^2)       1  0.6129  0.6129  2.8643  0.094507 .
winter:ears   1  0.1428  0.1428  0.6672  0.416483
winter:PpH    1  0.1404  0.1404  0.6561  0.420386
winter:K      1  0.1144  0.1144  0.5344  0.466933
winter:Mg     1  0.1899  0.1899  0.8874  0.349062
ears:PpH      1  0.1256  0.1256  0.5871  0.445817
ears:K        1  0.0176  0.0176  0.0823  0.774937
ears:Mg       1  0.1679  0.1679  0.7847  0.378402
PpH:K         1  0.1648  0.1648  0.7702  0.382831
PpH:Mg        1  0.1922  0.1922  0.8982  0.346156
K:Mg          1  0.1277  0.1277  0.5965  0.442209
Residuals    79 16.9057  0.2140
```

```
> m2 <- step(m1)
...
> anova(m2)
Analysis of Variance Table

Response: yield
           Df  Sum Sq Mean Sq F value    Pr(>F)
winter      1  6.5802  6.5802 33.3952 1.018e-07 ***
ears        1  0.7288  0.7288  3.6990  0.057538 .
K           1  1.7444  1.7444  8.8531  0.003737 **
Mg          1  0.2961  0.2961  1.5026  0.223400
I(PpH^2)    1  5.9517  5.9517 30.2056 3.446e-07 ***
ears:Mg     1  0.6489  0.6489  3.2935  0.072815 .
K:Mg        1  1.5297  1.5297  7.7634  0.006475 **
Residuals  92 18.1276  0.1970
---
```

```
> summary(m2, corr=T)
```

```
Correlation of Coefficients:
          (Intercept) winter ears  K     Mg    I(PpH^2) ears:Mg
winter    -0.05
ears      -0.91       -0.02
K         -0.38        0.00  -0.02
Mg        -0.99       -0.02   0.91  0.36
I(PpH^2)  -0.01        0.07  -0.15  0.31 -0.02
ears:Mg    0.90        0.02  -0.99  0.02 -0.92  0.16
K:Mg       0.38       -0.02   0.02 -1.00 -0.36 -0.33    -0.03
```

```
> 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)
Analysis of Variance Table

Response: yield
           Df  Sum Sq Mean Sq F value    Pr(>F)
w1          1  6.5802  6.5802 32.3949 2.542e-07 ***
e1          1  0.7288  0.7288  3.5882  0.062208 .
pH1         1  2.1735  2.1735 10.7005  0.001646 **
P1          1  0.0964  0.0964  0.4748  0.492988
K1          1  0.9223  0.9223  4.5407  0.036513 *
Mg1         1  0.4403  0.4403  2.1675  0.145310
I(w1^2)     1  0.0232  0.0232  0.1140  0.736647
I(e1^2)     1  0.2189  0.2189  1.0776  0.302707
I(pH1^2)    1  5.2872  5.2872 26.0295 2.629e-06 ***
I(P1^2)     1  0.0390  0.0390  0.1919  0.662635
I(K1^2)     1  0.6269  0.6269  3.0861  0.083213 .
I(Mg1^2)    1  0.7946  0.7946  3.9120  0.051770 .
w1:e1       1  0.1109  0.1109  0.5461  0.462318
```

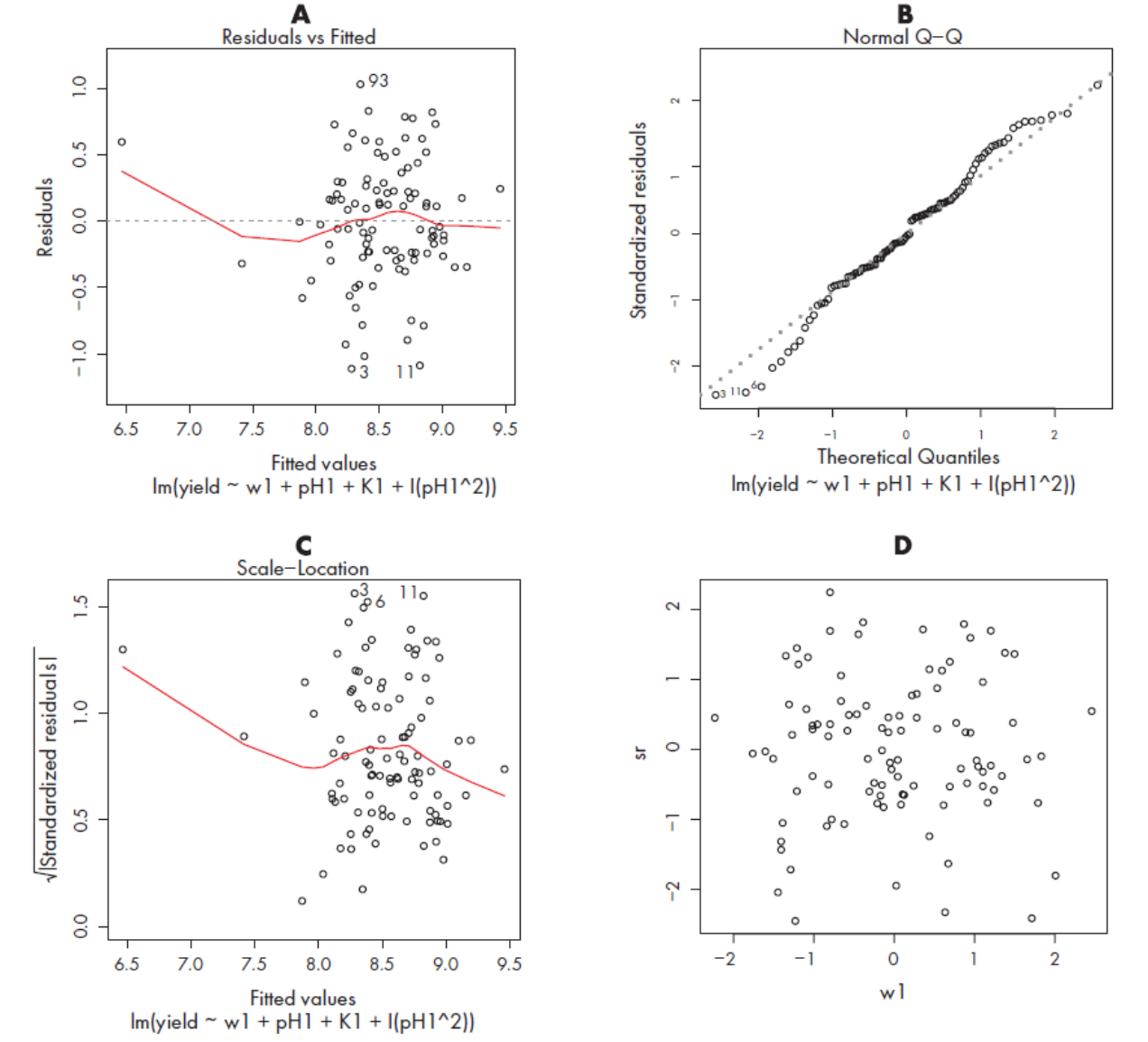
```
w1:pH1       1  0.0001  0.0001  0.0005  0.981871
w1:P1        1  0.3435  0.3435  1.6911  0.197610
w1:K1        1  0.2448  0.2448  1.2051  0.275966
w1:Mg1       1  0.1437  0.1437  0.7072  0.403144
e1:pH1       1  0.1750  0.1750  0.8616  0.356386
e1:P1        1  0.3084  0.3084  1.5182  0.221900
e1:K1        1  0.0010  0.0010  0.0048  0.945214
e1:Mg1       1  0.0699  0.0699  0.3442  0.559238
pH1:P1       1  0.0076  0.0076  0.0372  0.847532
pH1:K1       1  0.0695  0.0695  0.3421  0.560434
pH1:Mg1      1  0.4321  0.4321  2.1272  0.149055
P1:K1        1  0.6919  0.6919  3.4061  0.069067 .
P1:Mg1       1  0.0921  0.0921  0.4535  0.502830
K1:Mg1       1  0.3609  0.3609  1.7766  0.186774
Residuals   72 14.6249  0.2031
---
```

```
> m4 <- update(m3, ~.-w1:pH1); anova(m4)
> m5 <- update(m4, ~.-e1:K1); anova(m5)
...
> anova(m26)
Analysis of Variance Table

Response: yield
           Df  Sum Sq Mean Sq F value    Pr(>F)
w1          1  6.5802  6.5802 30.4123 3.001e-07 ***
pH1         1  1.9487  1.9487  9.0066  0.003436 **
K1          1  0.8711  0.8711  4.0262  0.047642 *
I(pH1^2)    1  5.6527  5.6527 26.1256 1.653e-06 ***
Residuals  95 20.5547  0.2164
---
```

$$a + b_1 \frac{winter - \bar{y}_{winter}}{s_{winter}} + b_2 \frac{pH - \bar{y}_{pH}}{s_{pH}} + c_1 \left(\frac{pH - \bar{y}_{pH}}{s_{pH}} \right)^2 + b_3 \frac{K - \bar{y}_K}{s_K}$$

```
> mean(winter); sd(winter)
[1] 275.64
[1] 50.94392
> mean(pH); sd(pH)
[1] 5.852
[1] 0.3812473
> mean(K); sd(K)
[1] 106.66
[1] 40.39657
```

```
> summary(m26, corr=T)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.71416    0.05917 147.278  < 2e-16 ***
w1           0.28494    0.04941   5.766 1.00e-07 ***
pH1         -0.01134    0.05490  -0.206   0.8368
K1          -0.09666    0.04885  -1.979   0.0508 .
I(pH1^2)    -0.18880    0.03694  -5.111 1.65e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4652 on 95 degrees of freedom
Multiple R-Squared: 0.4227,     Adjusted R-squared: 0.3984
F-statistic: 17.39 on 4 and 95 DF,  p-value: 9.75e-11

Correlation of Coefficients:
          (Intercept) w1    pH1   K1
w1        -0.06
pH1        0.25       -0.28
K1         0.00       -0.10 -0.22
I(pH1^2) -0.62        0.10 -0.40 -0.01
```



$$8.714 + 0.285\frac{winter - 275.6}{50.94} - 0.011\frac{pH - 5.852}{0.381} - 0.189\left(\frac{pH - 5.852}{0.381}\right)^2 - 0.097\frac{K - 106.7}{40.39}.$$