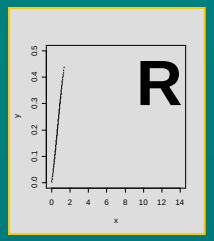


Analyses of continuous II

Stano Pekár

Gamma & Lognormal distributions

■ Gamma and lognormal data arise:

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

• measurements are continuous

- 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

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

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

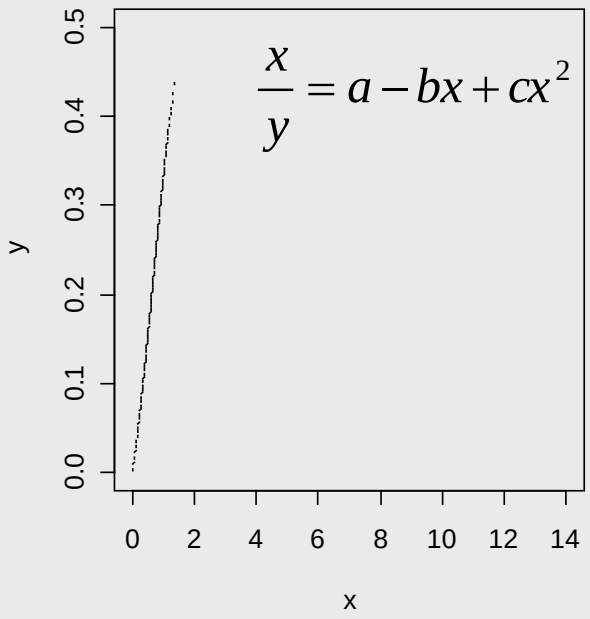
• predicted values: $\exp(Q) = median$

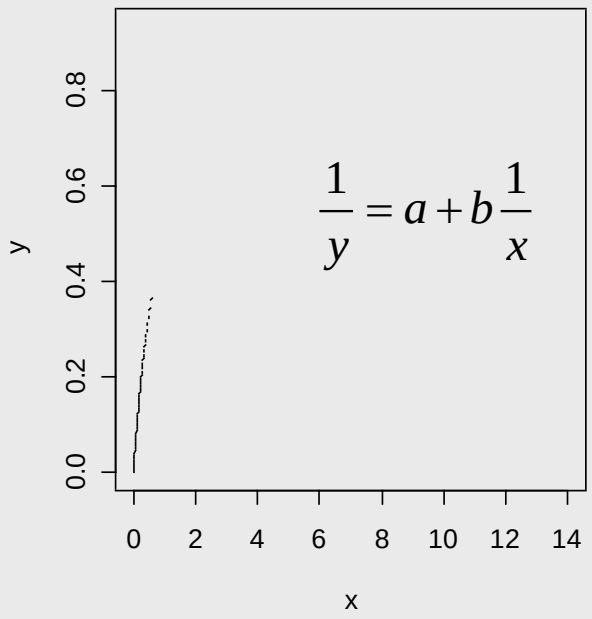
Gamma model

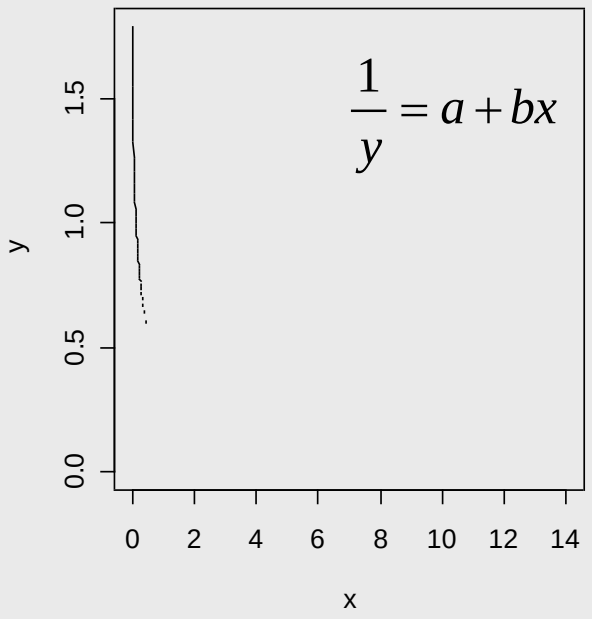
- used to model inverse polynomials moments - 2 parameters (μ, φ)

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

- 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`)

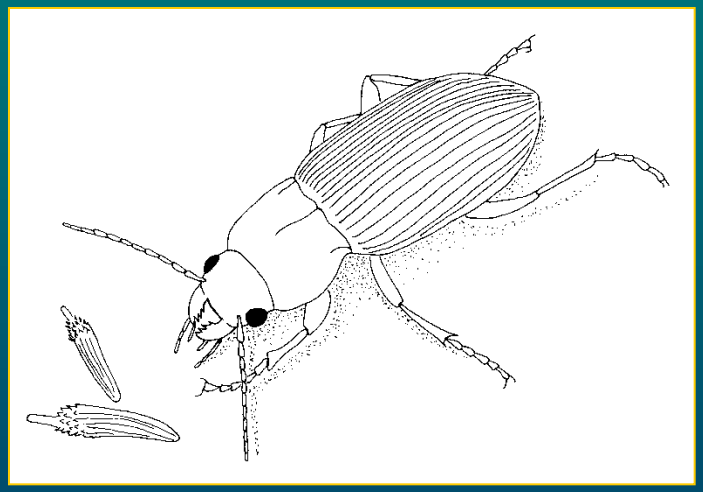
- **identity** (`identity`)

• `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

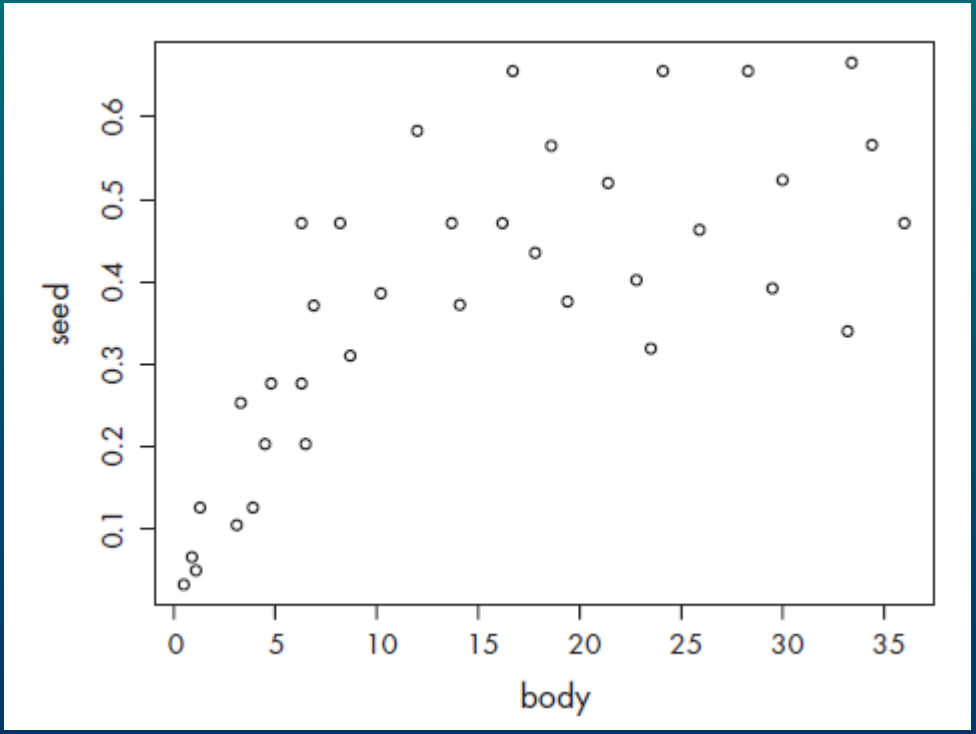
Is size of seeds related to the carabid body size?

What is the shape of the relationship?

Variables

body

seed



$$\frac{1}{\mu_i} = \alpha + \beta \frac{1}{body_i},$$

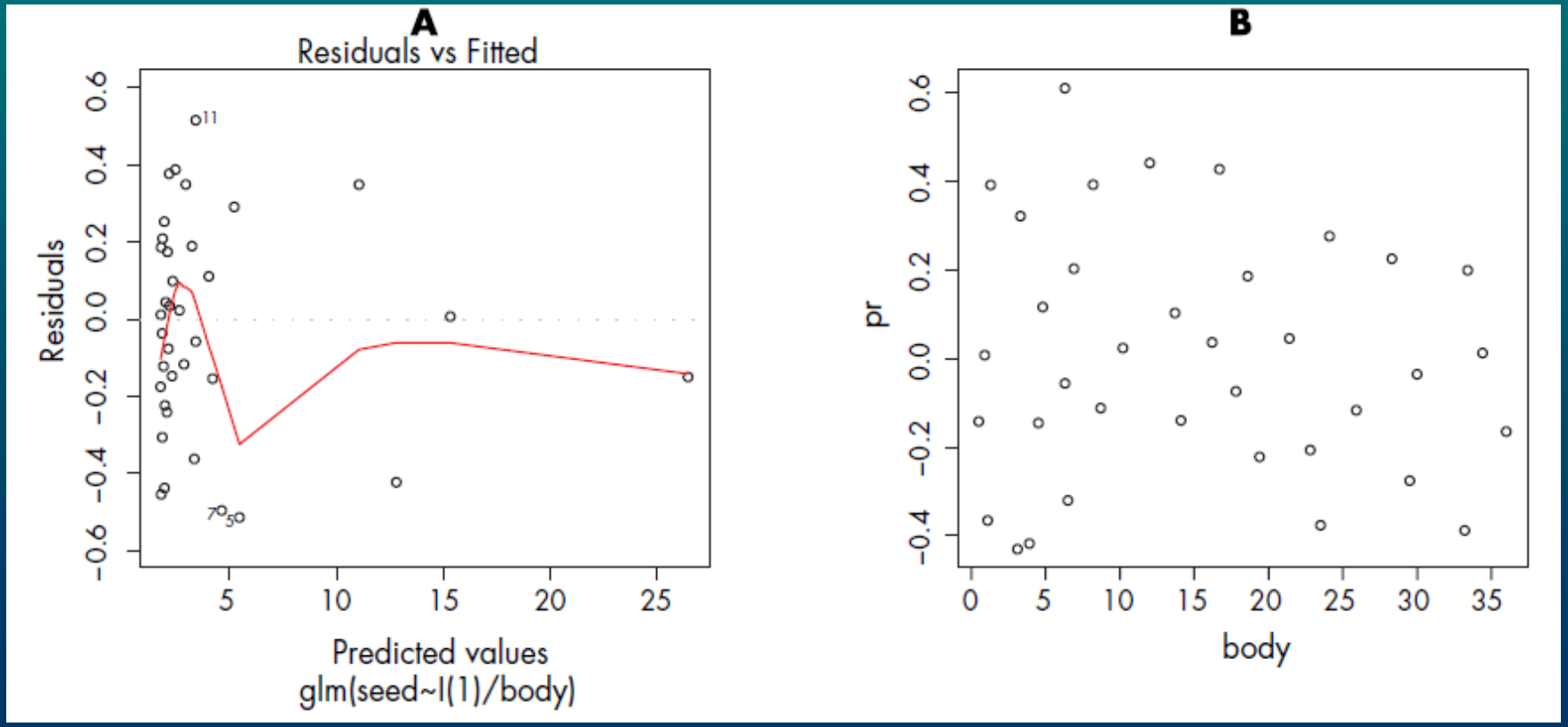
kde $seed_i \sim Gama(\mu_i, \varphi)$, nezávisle pro každého jedince.

```
> m1 <- glm(seed ~ I(1/body), family=Gamma)
> anova(m1, test="F")
Analysis of Deviance Table

Model: Gamma, link: inverse

Response: seed

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                         35    15.3681
I(1/body)  1   8.3662        34     7.0019 42.624 1.787e-07 ***
```

$$\frac{1}{\mu_i} = \alpha + \beta \frac{1}{body_i} + \gamma \frac{1}{body_i^2},$$

$seed_i \sim Gama(\mu_i, \varphi)$, nezávisle pro každého jedince.

```
> m2 <- glm(seed ~ I(1/body) + I(1/body^2), Gamma)
> anova(m1, m2, test="F")
Analysis of Deviance Table

Model 1: seed ~ I(1/body)
Model 2: seed ~ I(1/body) + I(1/body^2)
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        34     7.0019
2        33     7.0016  1   0.0003 0.0013 0.9713
```

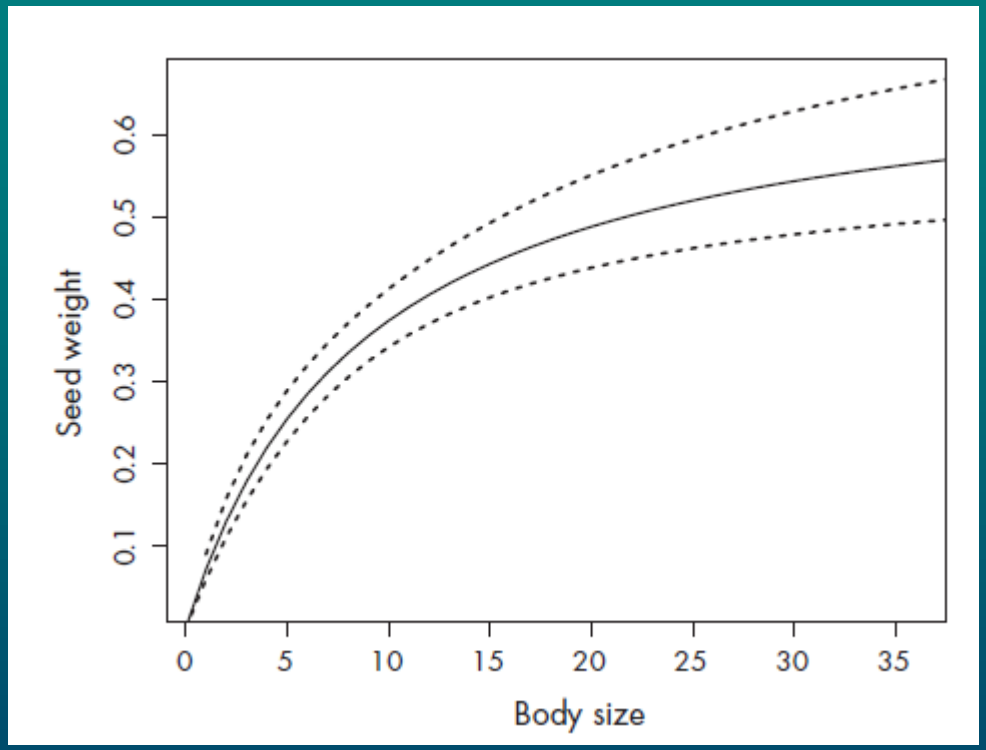


```
> summary(m1)

Call:
glm(formula = seed ~ I(1/body), family = Gamma)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-0.7530027  -0.4237538   0.0008676   0.2527096   0.7024871

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7418     0.3162   5.508 3.76e-06 ***
I(1/body)    11.8626     2.4463   4.849 2.69e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1962785)

    Null deviance: 15.3681  on 35  degrees of freedom
Residual deviance:  7.0019  on 34  degrees of freedom
AIC: -49.676
```

Coefficient of determination: $(15.3681 - 7.0019) / 15.3681 = 0.54.$

Asymptote: $1/1.7418 = 0.574,$

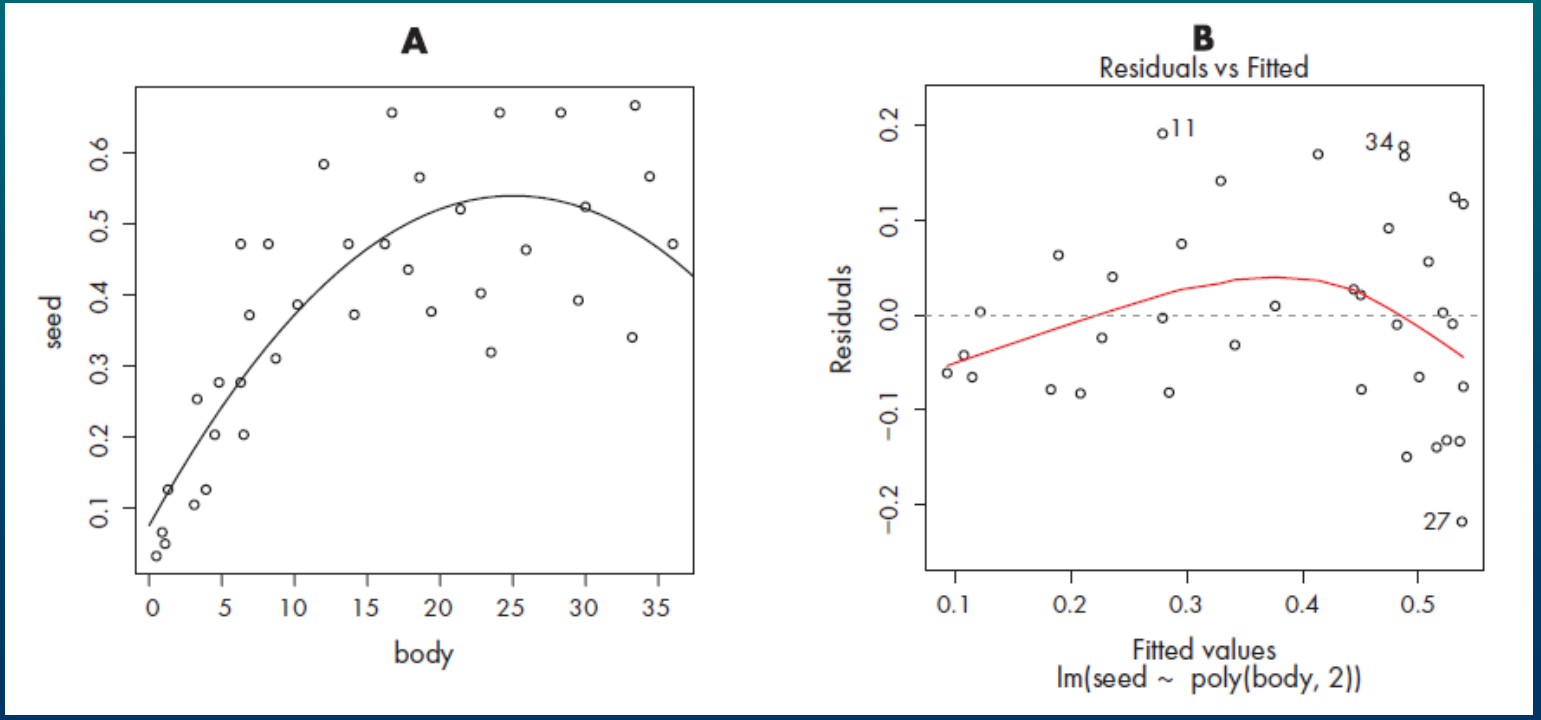
$$\frac{body}{1.742body + 11.86}$$



$$seed_i = \alpha + \beta body_i + \gamma body_i^2 + \varepsilon_i ,$$

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

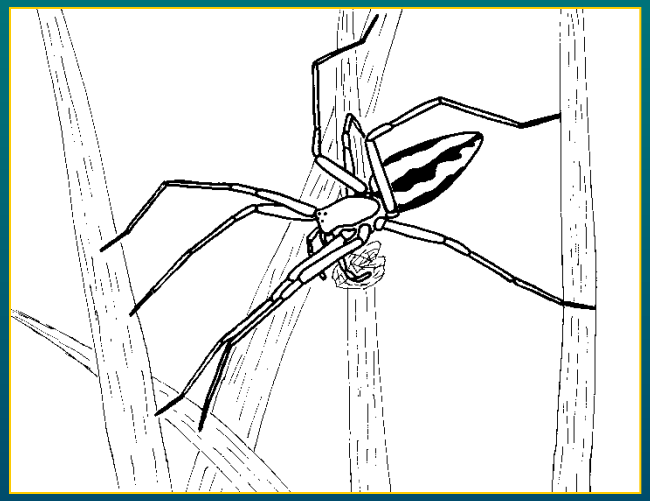
```
> m3 <- lm(seed ~ poly(body,2))
> summary(m3)
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.30842    0.02318  13.305 8.17e-15 ***
poly(body, 2)1  0.55682    0.13908   4.004 0.000333 ***
poly(body, 2)2 -0.41591    0.13908  -2.990 0.005235 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1391 on 33 degrees of freedom
Multiple R-Squared: 0.4307,     Adjusted R-squared: 0.3962
F-statistic: 12.49 on 2 and 33 DF,  p-value: 9.173e-05
```



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?

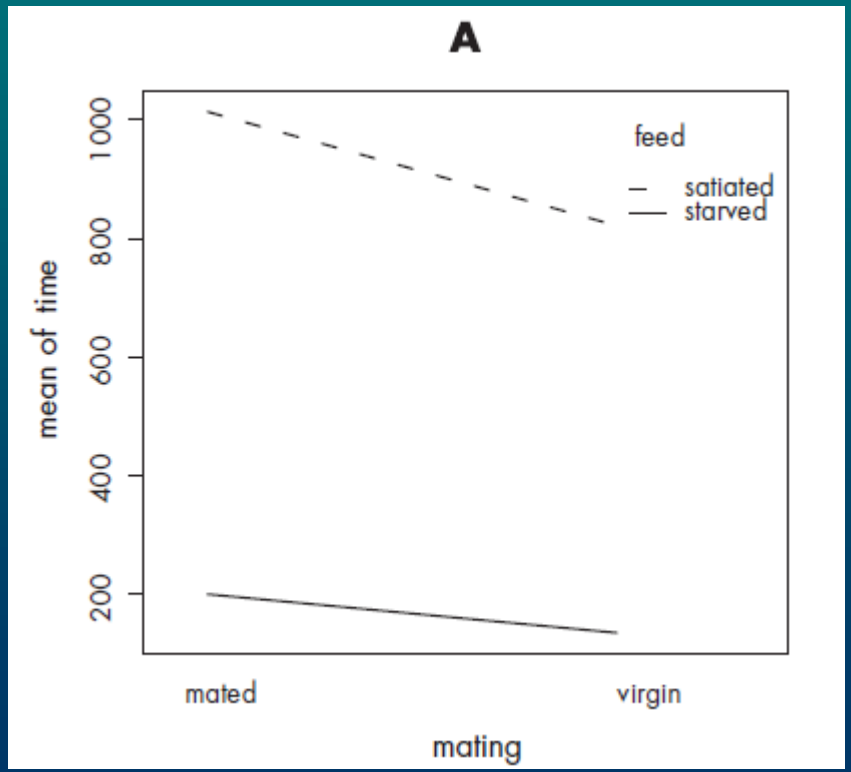
If it is what is the difference between factor levels?

Variables

MATING: mated, virgin

FEED: satiated, starved

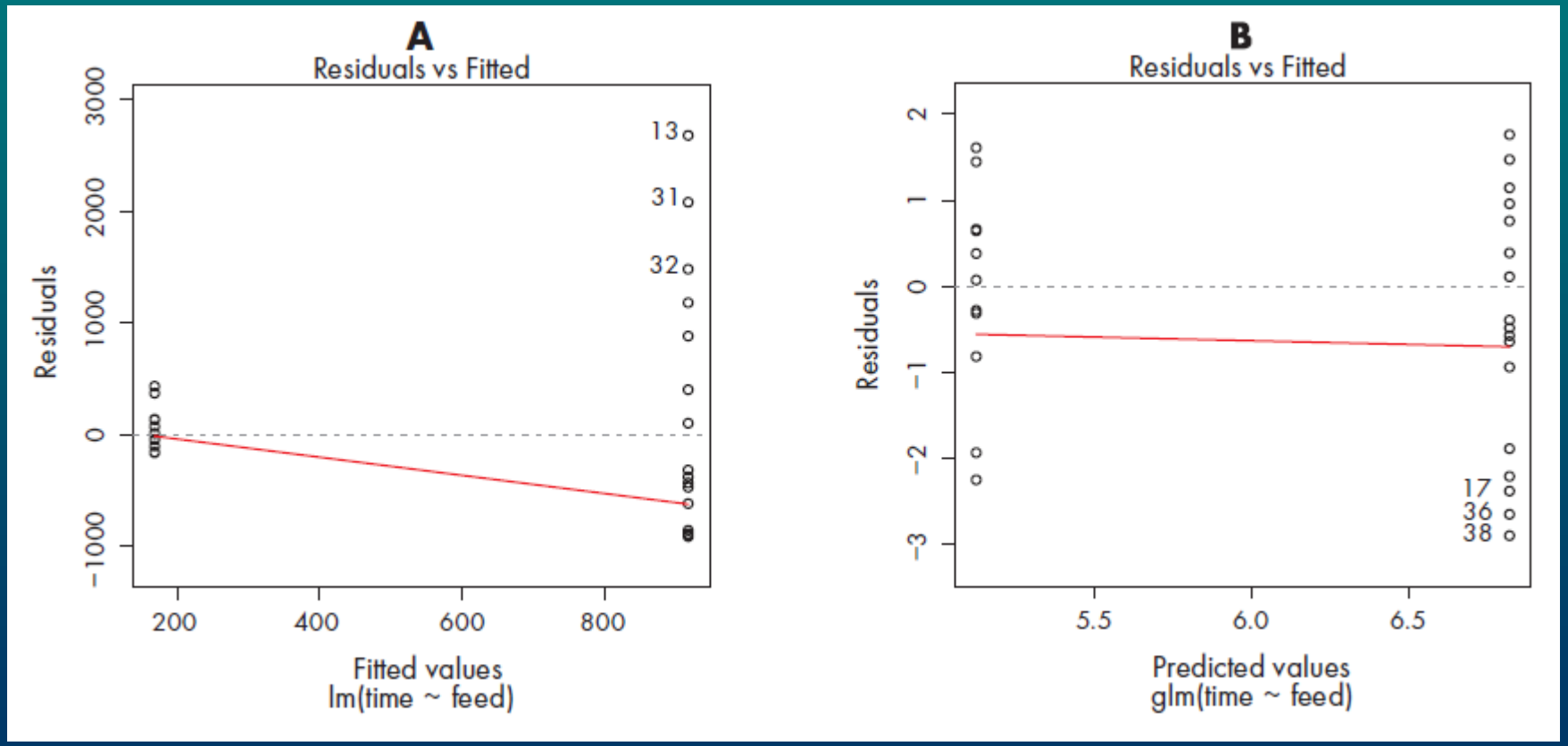
time



$$time_{ijk} = \alpha + MATING_j + FEED_k + MATING{:}FEED_{jk} + \varepsilon_{ijk},$$
s $\varepsilon_{ijk} \sim N(0, \sigma^2)$, nezávisle pro jednotlivá pozorování.

```
> m1 <- lm(time ~ mating*feed)
> anova(m1)
Analysis of Variance Table

Response: time
            Df   Sum Sq  Mean Sq F value   Pr(>F)
mating       1   165122   165122  0.2558 0.616098
feed         1  5625000  5625000  8.7142 0.005528 **
mating:feed  1    40322    40322  0.0625 0.804058
Residuals   36 23237845   645496
```

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

Response: time
          Df   Sum Sq  Mean Sq F value   Pr(>F)
feed       1  5625000  5625000  9.1177 0.004507 **
Residuals 38 23443290   616929
---
```


A
Residuals vs Fitted
Residuals
3000
2000
1000
0
−1000
13
31
32
200
400
600
800
Fitted values
lm(time ~ feed)
B
Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
17
36
38
5.5
6.0
6.5
Predicted values
glm(time ~ feed)

$$\log(\mu_{jk}) = \alpha + MATING_j + FEED_k + MATING{:}FEED_{jk},$$
s $time_{jk} \sim Gama(\mu_{jk}, \varphi)$, nezávisle pro jednotlivá pozorování,

```
> m4 <- glm(time ~ mating*feed, Gamma(link=log))
> anova(m4, test="F")
...
            Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                           39    122.018
mating       1    0.564         38    121.454  0.3888 0.5368618
feed         1   26.258         37     95.196 18.1021 0.0001425 ***
mating:feed  1    0.083         36     95.113  0.0570 0.8126218
---
```

```
> anova(m6, test="F")
...
     Df Deviance Resid. Df Resid. Dev    F    Pr(>F)
NULL                    39    122.018
feed  1   25.922         38     96.096 20.3 6.138e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(m6)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.8222     0.2527  26.999  < 2e-16 ***
feedstarved  -1.6982     0.3573  -4.752 2.87e-05 ***
---
```

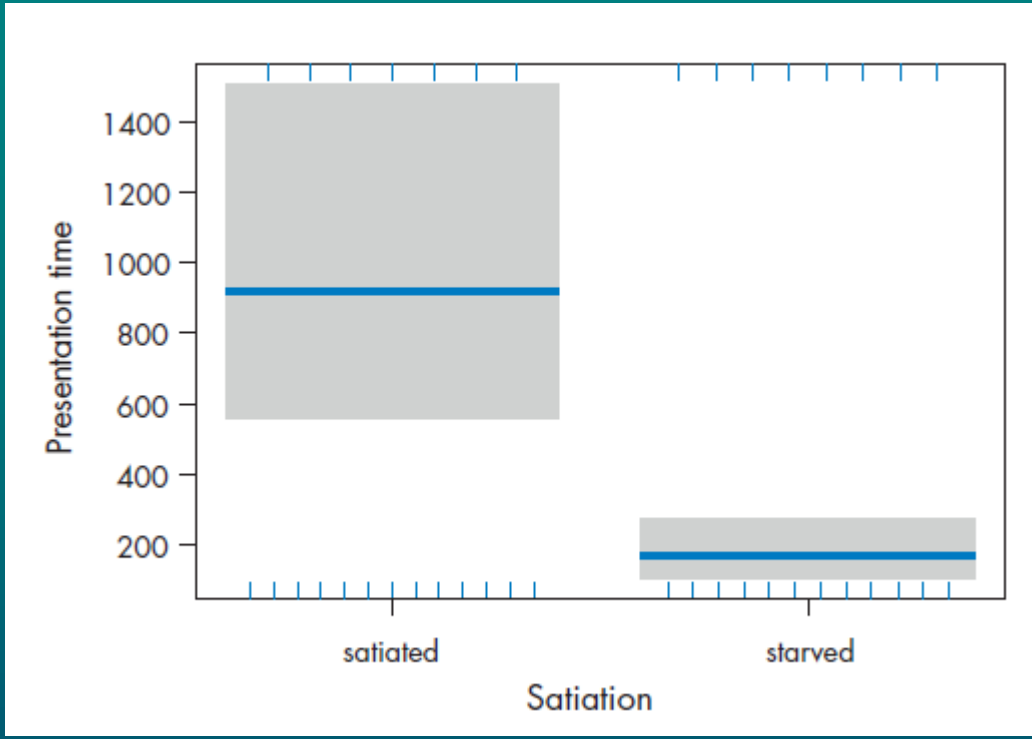
```
> exp(6.8222)
[1] 918.0024
> exp(6.8222-1.6982)
[1] 168.0061
> tapply(time, feed, mean)
satiated  starved
     918      168
```

$$\log(time_{ijk}) = \alpha + MATING_j + FEED_k + MATING{:}FEED_{jk} + \varepsilon_{iik},$$
$$\text{s } \varepsilon_{ijk} \sim N(0, \sigma^2), \text{ nezávisle pro jednotlivá pozorování.}$$

```
> m7 <- lm(log(time) ~ mating*feed)
> anova(m7)
Analysis of Variance Table

Response: log(time)
             Df  Sum Sq Mean Sq F value  Pr(>F)
mating        1  11.432  11.432  2.7578 0.10547
feed          1  19.262  19.262  4.6468 0.03787 *
mating:feed   1   0.019   0.019  0.0045 0.94681
Residuals    36 149.226   4.145
---
```

```
> m8 <- lm(log(time) ~ feed)
> summary(m8)
```

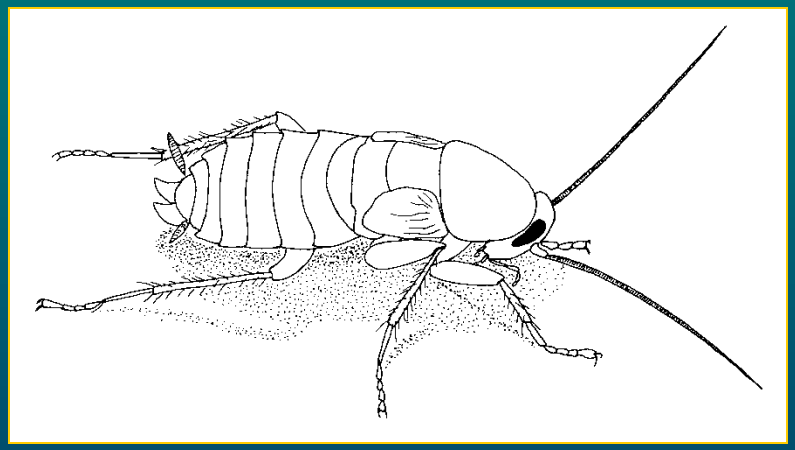
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.4658     0.4598  11.887 2.27e-14 ***
feedstarved   -1.3879     0.6503  -2.134   0.0393 *
```


Presentation time
1400
1200
1000
800
600
400
200
satiated
starved
Satiation

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?

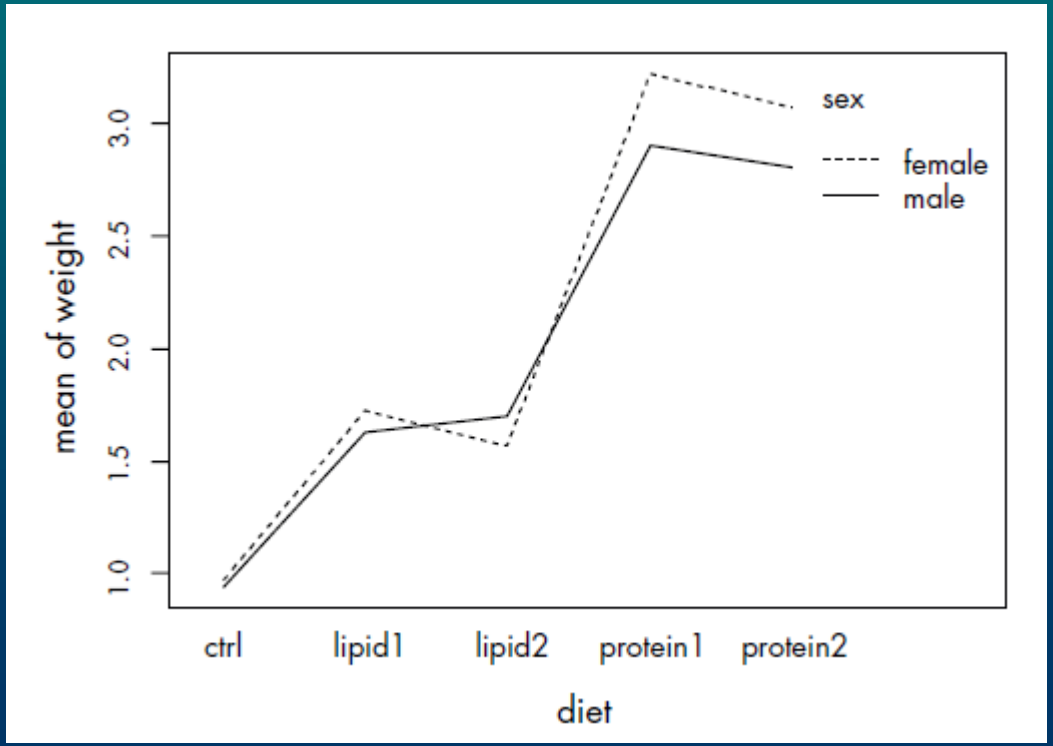
Variables

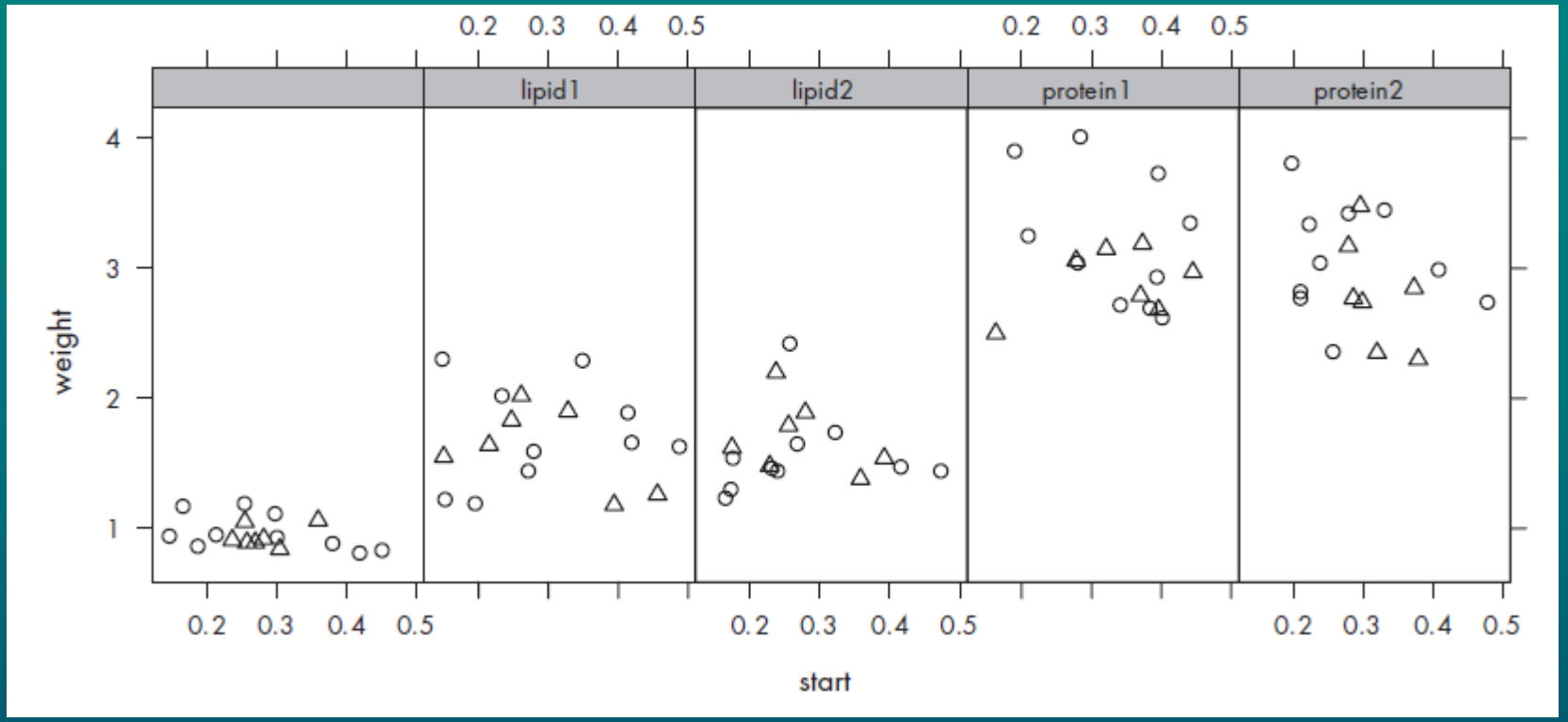
DIET: control, lipid1, lipid2, protein1, protein2

SEX: male, female

start

weight





$$\log(weight_{ijk}) = \alpha + DIET_j + SEX_k + \beta start_i + DIET{:}SEX_{jk} + \delta_{j1} start_i + \delta_{1k} start_i + \delta_{jk} start_i + \varepsilon_{ijk},$$

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

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

Response: log(weight)
               Df  Sum Sq Mean Sq  F value Pr(>F)
diet            4 16.1349  4.0337 150.3981 <2e-16 ***
sex             1  0.0261  0.0261   0.9732 0.3275
start           1  0.0455  0.0455   1.6956 0.1975
diet:sex        4  0.0866  0.0217   0.8073 0.5250
diet:start      4  0.0244  0.0061   0.2272 0.9222
sex:start       1  0.0315  0.0315   1.1743 0.2825
diet:sex:start  4  0.1829  0.0457   1.7048 0.1596
Residuals      65  1.7433  0.0268
---
```

```
> anova(lm(log(weight) ~ sex*diet*start))
Analysis of Variance Table

Response: log(weight)
               Df  Sum Sq Mean Sq  F value Pr(>F)
sex             1  0.0261  0.0261   0.9732 0.3275
diet            4 16.1349  4.0337 150.3981 <2e-16 ***
start           1  0.0455  0.0455   1.6956 0.1975
sex:diet        4  0.0866  0.0217   0.8073 0.5250
sex:start       1  0.0196  0.0196   0.7302 0.3959
diet:start      4  0.0363  0.0091   0.3382 0.8512
sex:diet:start  4  0.1829  0.0457   1.7048 0.1596
Residuals      65  1.7433  0.0268
```

$$\log(weight_{ijk}) = \alpha + DIET_j + SEX_k + \beta start_i + \gamma start_i^2 + DIET{:}SEX_{jk} + \delta_{j1} start_i + \delta_{1k} start_i + \delta_{jk} start_i + \omega_{j1} start_i^2 + \omega_{1k} start_i^2 + \omega_{jk} start_i^2 + \varepsilon_{ijk}, \quad (9\text{-}13)$$

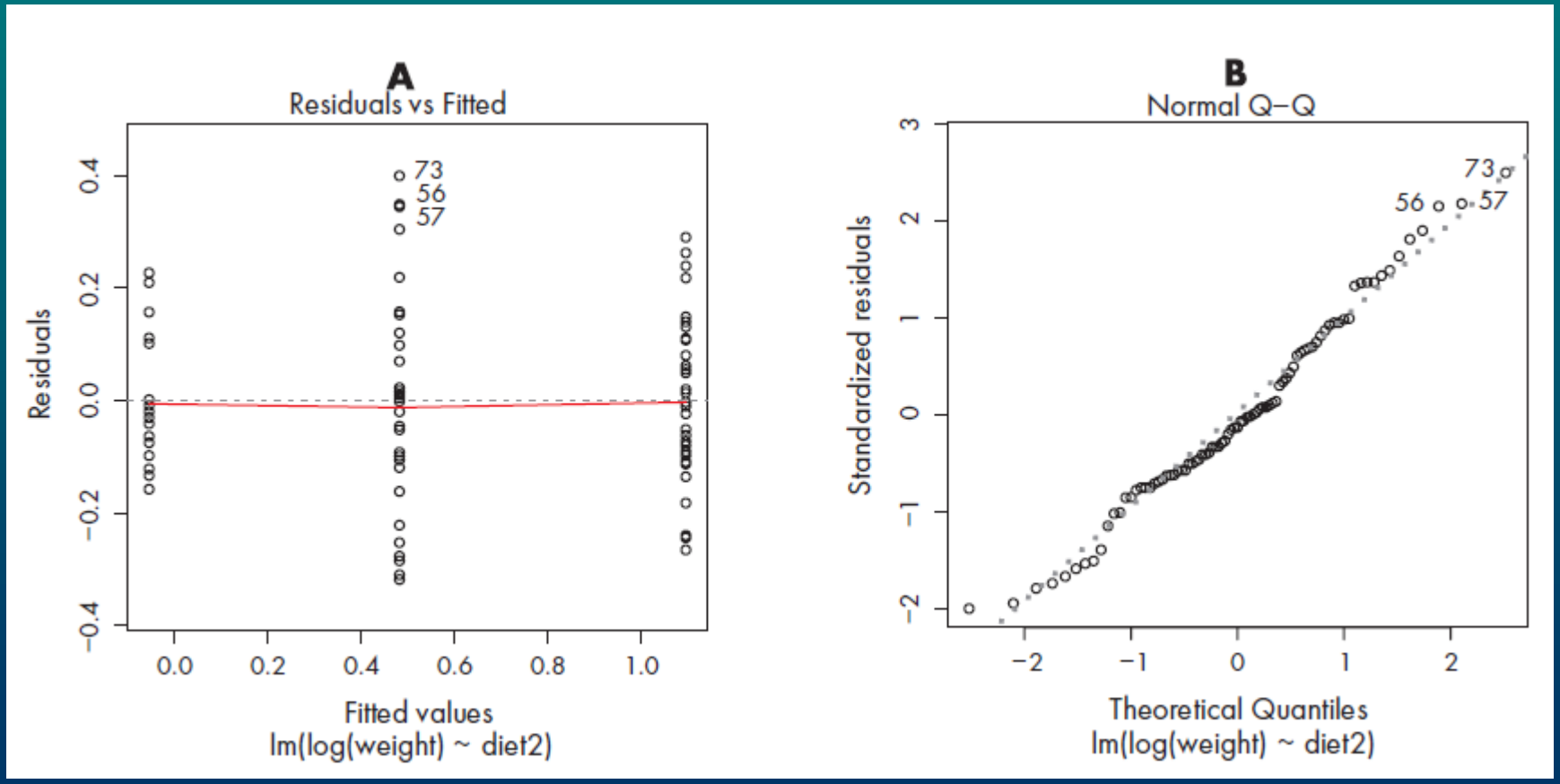
kde $\varepsilon_{ijk} \sim N(0, \sigma^2)$, nezávisle pro jednotlivá měření.

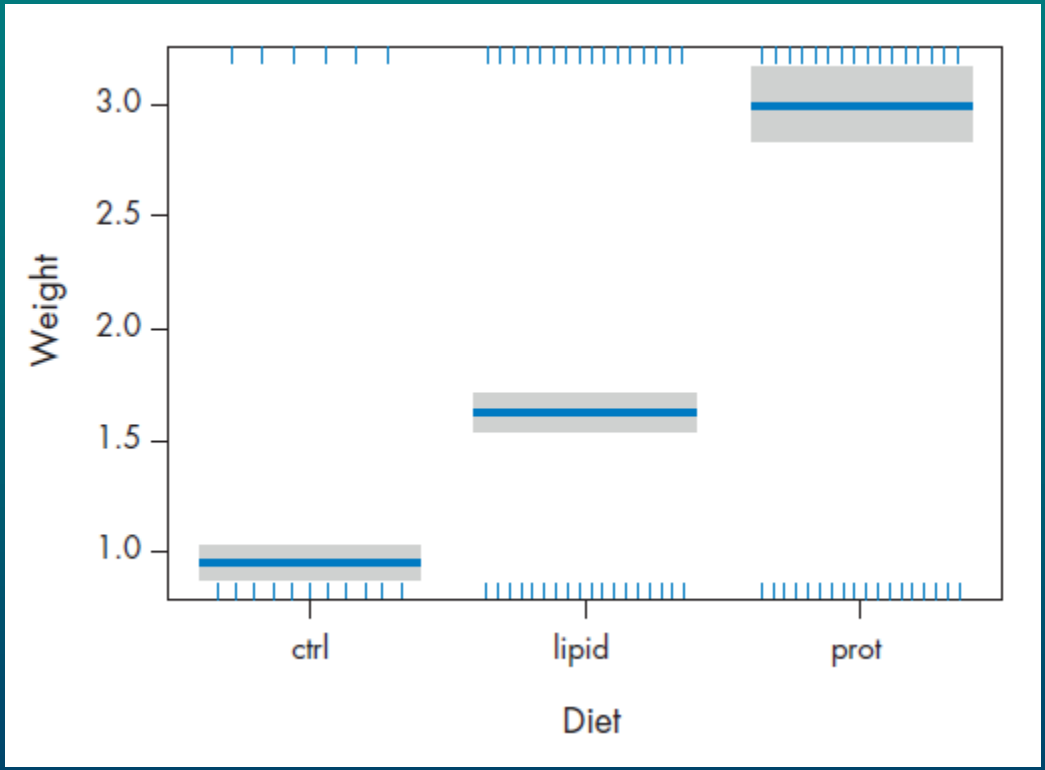
```
> m2 <- lm(log(weight) ~ diet*sex*poly(start,2))
> anova(m1, m2)
Analysis of Variance Table

Model 1: log(weight) ~ diet * sex * start
Model 2: log(weight) ~ diet * sex * poly(start, 2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     65 1.7433
2     55 1.4122 10   0.33113 1.2896 0.2592
```

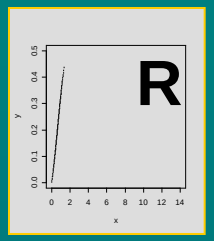
```
> anova(m3)
Analysis of Variance Table

Response: log(weight)
           Df  Sum Sq Mean Sq  F value Pr(>F)
diet        4 16.1349  4.0337 144.4941 <2e-16 ***
sex         1  0.0261  0.0261   0.9350 0.3369
start       1  0.0455  0.0455   1.6290 0.2061
diet:sex    4  0.0866  0.0217   0.7756 0.5448
diet:start  4  0.0244  0.0061   0.2183 0.9274
sex:start   1  0.0315  0.0315   1.1282 0.2919
Residuals  69  1.9262  0.0279
```

```
> summary(m8)

Call:
lm(formula = log(weight) ~ diet)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33311 -0.09764 -0.02934  0.11146  0.41505

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.05319    0.03967  -1.341    0.184
dietlipid1    0.55181    0.05610   9.836 2.02e-15 ***
dietlipid2    0.52190    0.05610   9.303 2.23e-14 ***
dietprotein1  1.17298    0.05610  20.908  < 2e-16 ***
dietprotein2  1.12984    0.05610  20.139  < 2e-16 ***
---
```

```
> summary(m9)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.05319    0.03940   -1.35    0.181
diet2lipid   0.53686    0.04825   11.13   <2e-16 ***
diet2prot    1.15141    0.04825   23.86   <2e-16 ***
```


A
Residuals vs Fitted
Residuals
0.4
0.2
0.0
−0.2
−0.4
73
56
57
0.0
0.2
0.4
0.6
0.8
1.0
Fitted values
lm(log(weight) ~ diet2)
B
Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
73
56
57
−2
−1
0
1
2
Theoretical Quantiles
lm(log(weight) ~ diet2)


3.0
2.5
2.0
1.5
1.0
Weight
ctrl
lipid
prot
Diet



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`)

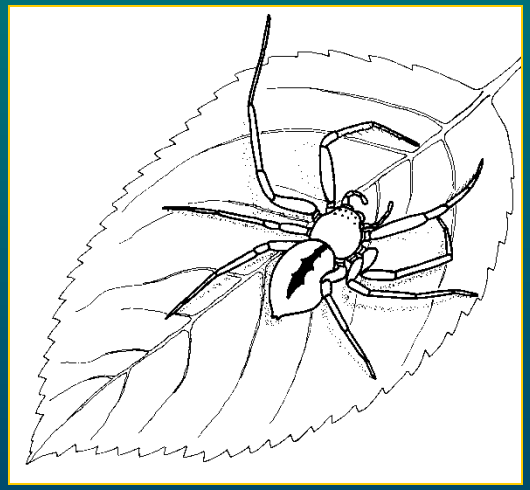
- estimated parameters are on logaritmic scale ($-\infty$, $+\infty$)

- 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

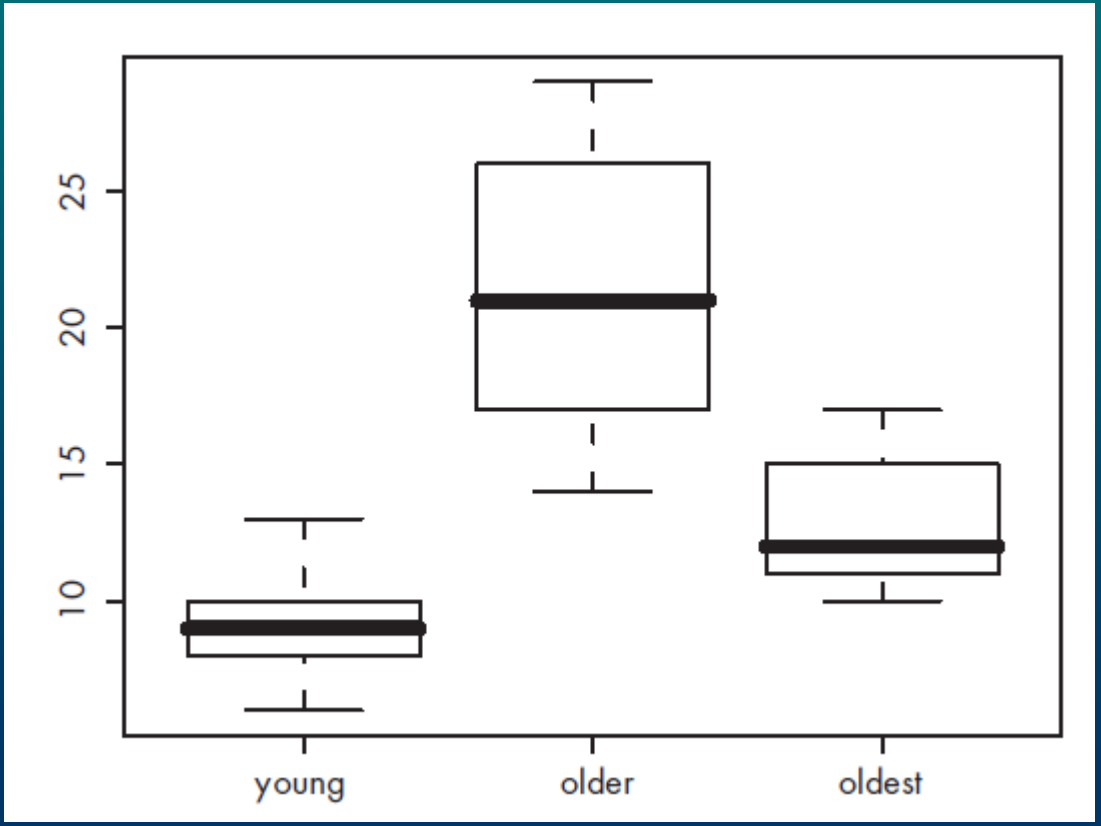
Is diversity related to the age of orchards?

What is the trend of change?

Variables

ORCHARD: young, older, oldest

divers



$$\log(\mu_j) = \alpha + ORCHARD_j,$$

kde $divers_j \sim Poi(\mu_j)$, nezávisle pro jednotlivé sady.

```
> m1 <- glm(divers ~ orchard, family=poisson)
> anova(m1, test="Chi")
Analysis of Deviance Table

Model: poisson, link: log

Response: divers

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       14     38.964
orchard  2   26.246         12     12.718 1.999e-06
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.2192     0.1474  15.051  < 2e-16 ***
orchardolder   0.8442     0.1763   4.788 1.68e-06 ***
orchardoldest  0.3457     0.1927   1.794   0.0727 .
---
```

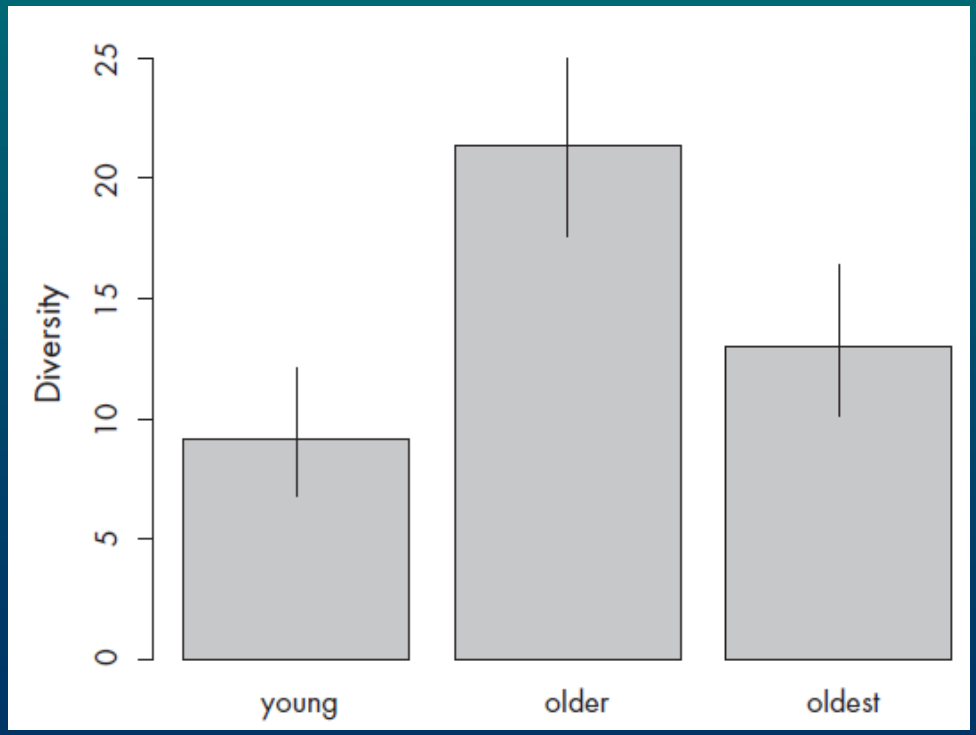
```
> contrasts(orchard) <- "contr.helmert"
> m2 <- glm(divers ~ orchard, family=poisson)
> summary(m2)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.61585    0.07186  36.404  < 2e-16 ***
orchard1     0.42209    0.08815   4.788 1.68e-06 ***
orchard2    -0.02545    0.05072  -0.502    0.616
```

```
> orchard1 <- ordered(orchard)
> m3 <- glm(divers ~ orchard1, family=poisson)
```

```
> summary(m3)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.61585    0.07186  36.404  < 2e-16 ***
orchard1.L   0.24448    0.13624   1.794   0.0727 .
orchard1.Q  -0.54813    0.11144  -4.919 8.71e-07 ***
```

```
> m3 <- glm(divers ~ orchard - 1, poisson)
> summary(m3)
...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
orchardyoung  2.21920    0.14744   15.05   <2e-16 ***
orchardolder  3.06339    0.09667   31.69   <2e-16 ***
orchardoldest 2.56495    0.12403   20.68   <2e-16 ***
```

```
> exp(confint(m3))
Waiting for profiling to be done...
                  2.5%    97.5%
orchardyoung   6.790864 12.12010
orchardolder  17.597063 25.71441
orchardoldest 10.090235 16.42096
```



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

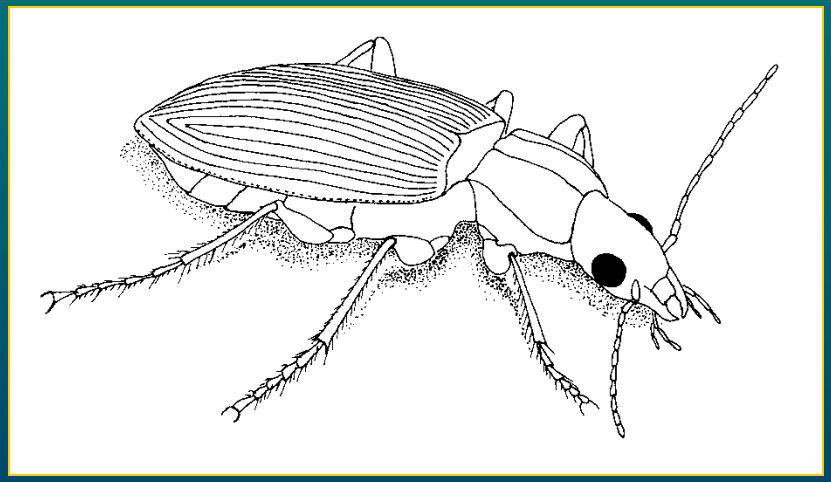
• without correction for overdispersion there would be too many false positive results (in favour of H_A)

• 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

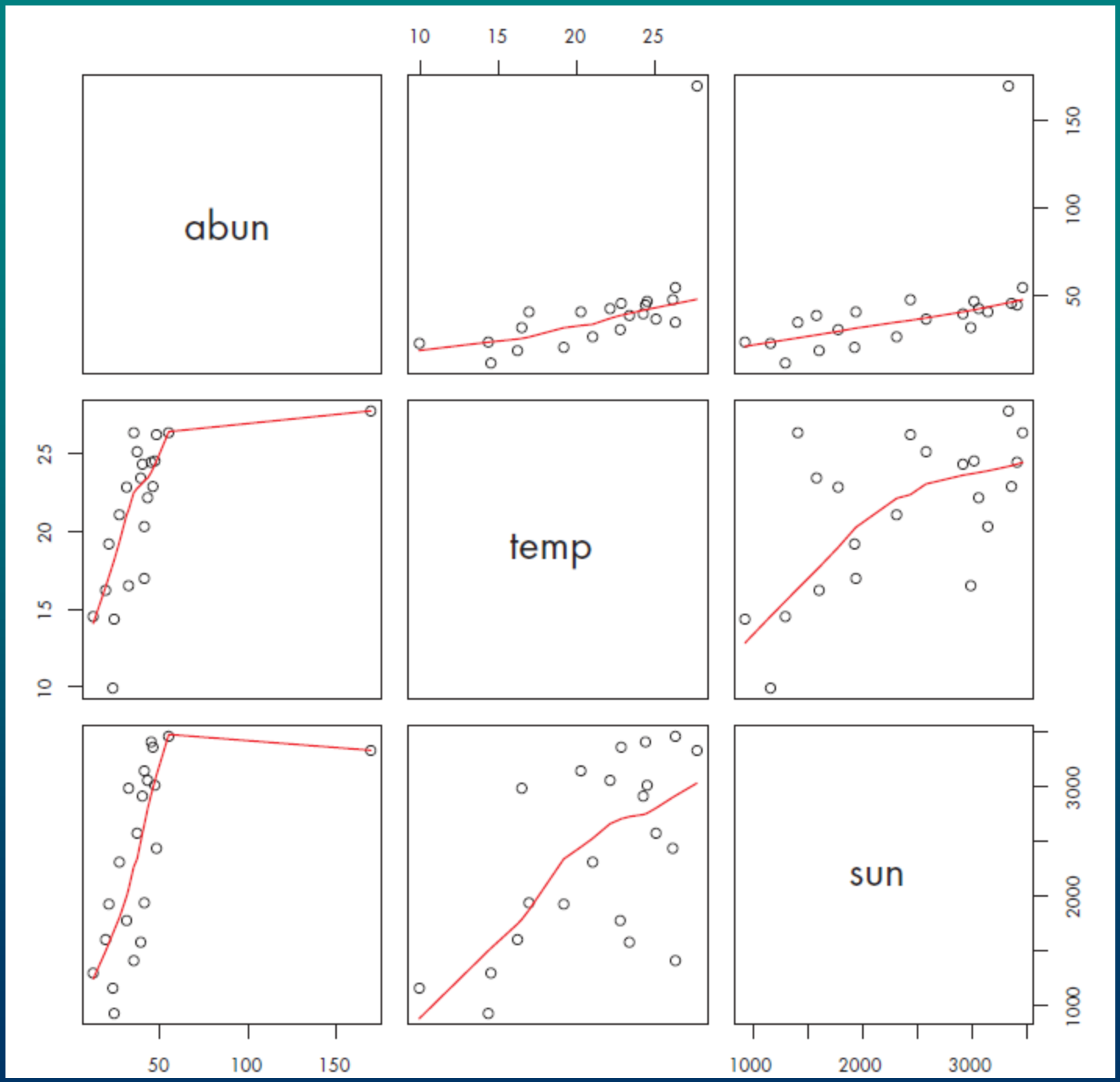
In the field, on 21 wheat plots the abundance of carabid beetles was studied by means of pitfall traps. At every site average day temperature [ºC] and average sun activity [W/m^2] was recorded.

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

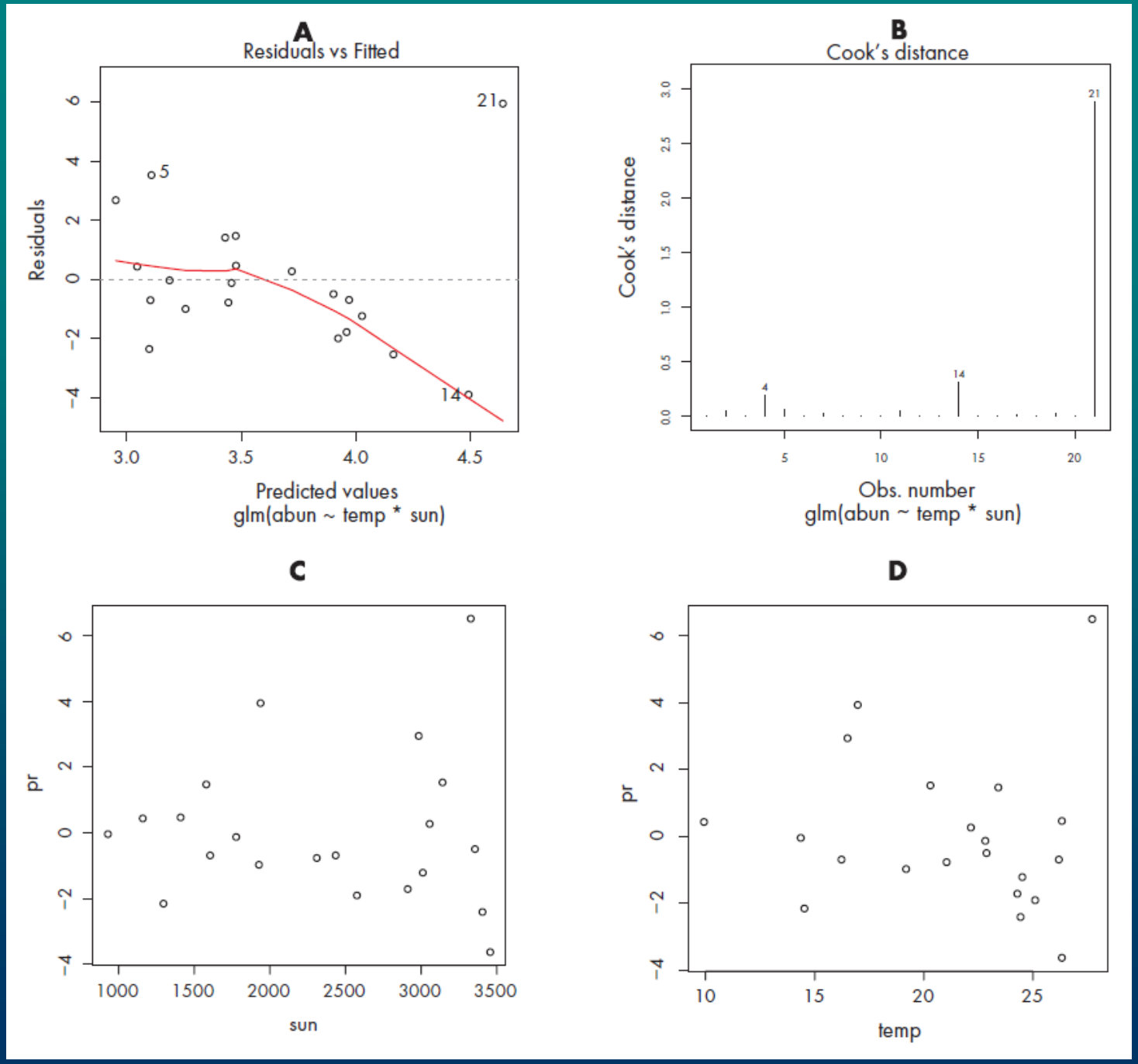

abun
temp
sun
10
15
20
25
50
100
150
1000
2000
3000

$$\log(\mu_i) = \alpha + \beta_1 temp_i + \beta_2 sun_i + \delta temp_i sun_i ,$$

kde $abun_i \sim Poi(\mu_i)$, nezávisle pro jednotlivé porosty.

```
> m1 <- glm(abun ~ temp*sun, family=poisson)
> summary(m1)
...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.195e+00  4.745e-01   8.840  < 2e-16 ***
temp        -5.386e-02  2.258e-02  -2.385   0.0171 *
sun         -1.151e-03  2.364e-04  -4.869 1.12e-06 ***
temp:sun     6.257e-05  1.006e-05   6.221 4.95e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 317.229  on 20  degrees of freedom
Residual deviance:  98.657  on 17  degrees of freedom
```

```
> m2 <- update(m1, family=quasipoisson)
> anova(m2, test="F")
...
         Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                        20     317.23
temp      1   153.10        19     164.12 24.5836 0.0001196 ***
sun       1    27.90        18     136.23  4.4796 0.0493541 *
temp:sun  1    37.57        17      98.66  6.0324 0.0251002 *
---
```


A
Residuals vs Fitted
Residuals
Predicted values
glm(abun ~ temp * sun)
B
Cook's distance
Cook's distance
Obs. number
glm(abun ~ temp * sun)
C
pr
sun
D
pr
temp

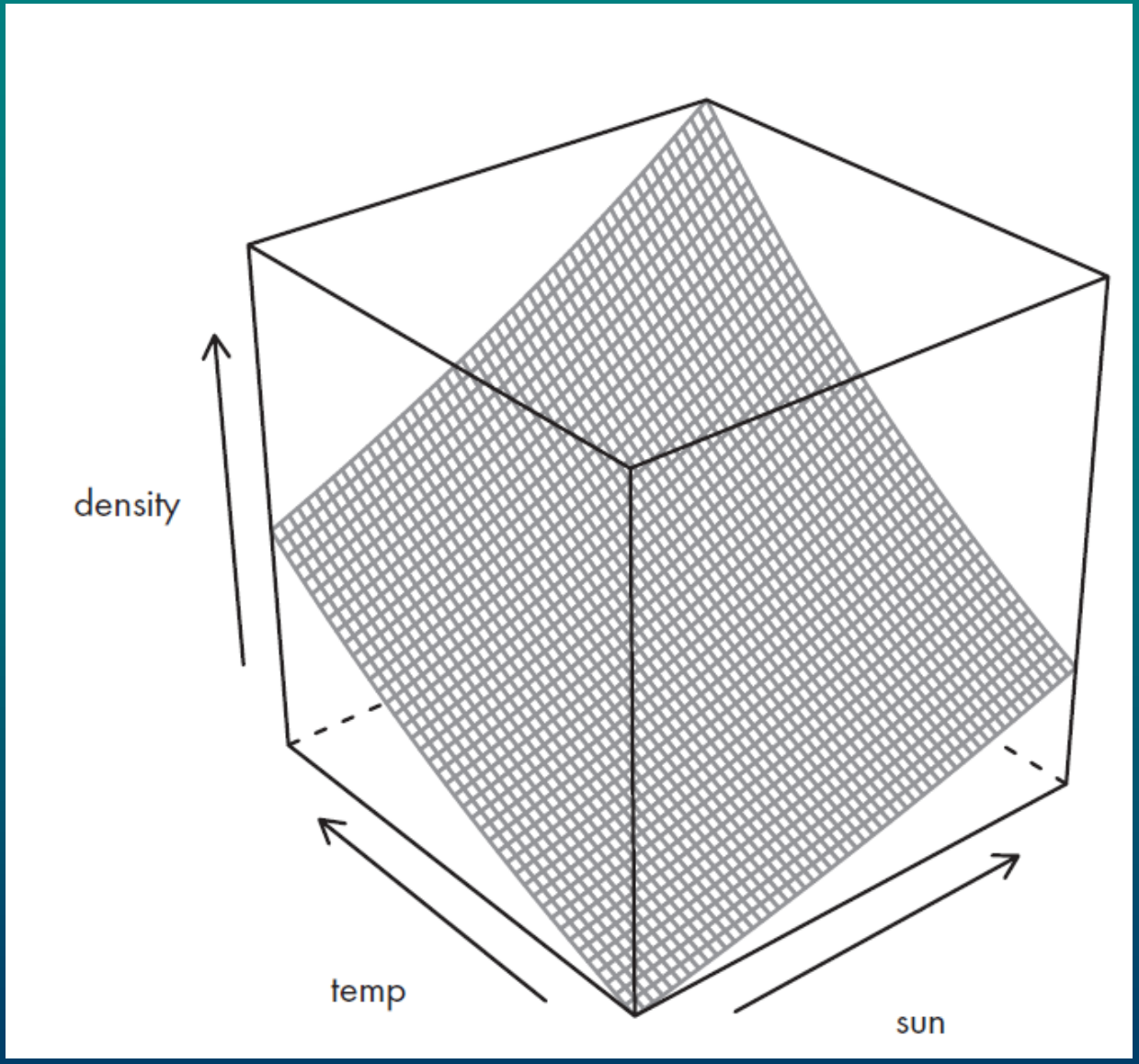
```
> m3 <- glm(abun ~ temp*sun, poisson, subset=-21)
> anova(m3, test="Chi")
...
         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                        19     75.292
temp      1   40.291         18     35.001 2.188e-10
sun       1   12.165         17     22.836 4.870e-04
temp:sun  1    0.117         16     22.719     0.732
> m4 <- update(m3, ~.-temp:sun)
> anova(m4, test="Chi")
...
     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    19     75.292
temp  1   40.291         18     35.001 2.188e-10
sun   1   12.165         17     22.836 4.870e-04
```

```
> library(car)
> Anova(m4)
Analysis of Deviance Table (Type II tests)

Response: abun
     LR Chisq Df Pr(>Chisq)
temp   12.567  1  0.0003926 ***
sun    12.165  1  0.0004870 ***
---
```

```
> vif(m4)
    temp      sun
1.325588 1.325588
```

```
> summary(m4)
...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.283e+00  2.088e-01  10.933  < 2e-16 ***
temp        3.781e-02  1.070e-02   3.534 0.000409 ***
sun         1.954e-04  5.655e-05   3.455 0.000550 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 75.292  on 19  degrees of freedom
Residual deviance: 22.836  on 17  degrees of freedom
AIC: 135.76
```

```
> (75.292-22.836)/75.292
[1] 0.6967008
```

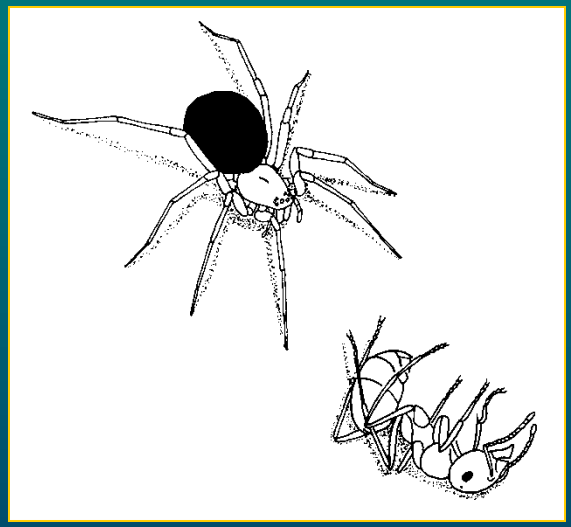


$$\exp(2.283 + 0.038sun + 0.0002temp).$$

1-way ANCOVA

Background

Some spiders are specialised in their diet. Specialisation can involve evolution of physiological and behavioural traits, such as prey-specific venom and number of attacks.



Design

In the lab, the number of attacks of an ant-eating spider on ants of two subfamilies was observed. For each subfamily 20 species of ants were used. Each ant species was tested once. For each ant body size was recorded as it may influence its susceptibility to venom.

Hypotheses

Was the number of attack related to ant size?

Was the number of attacks similar for ants of both subfamilies?

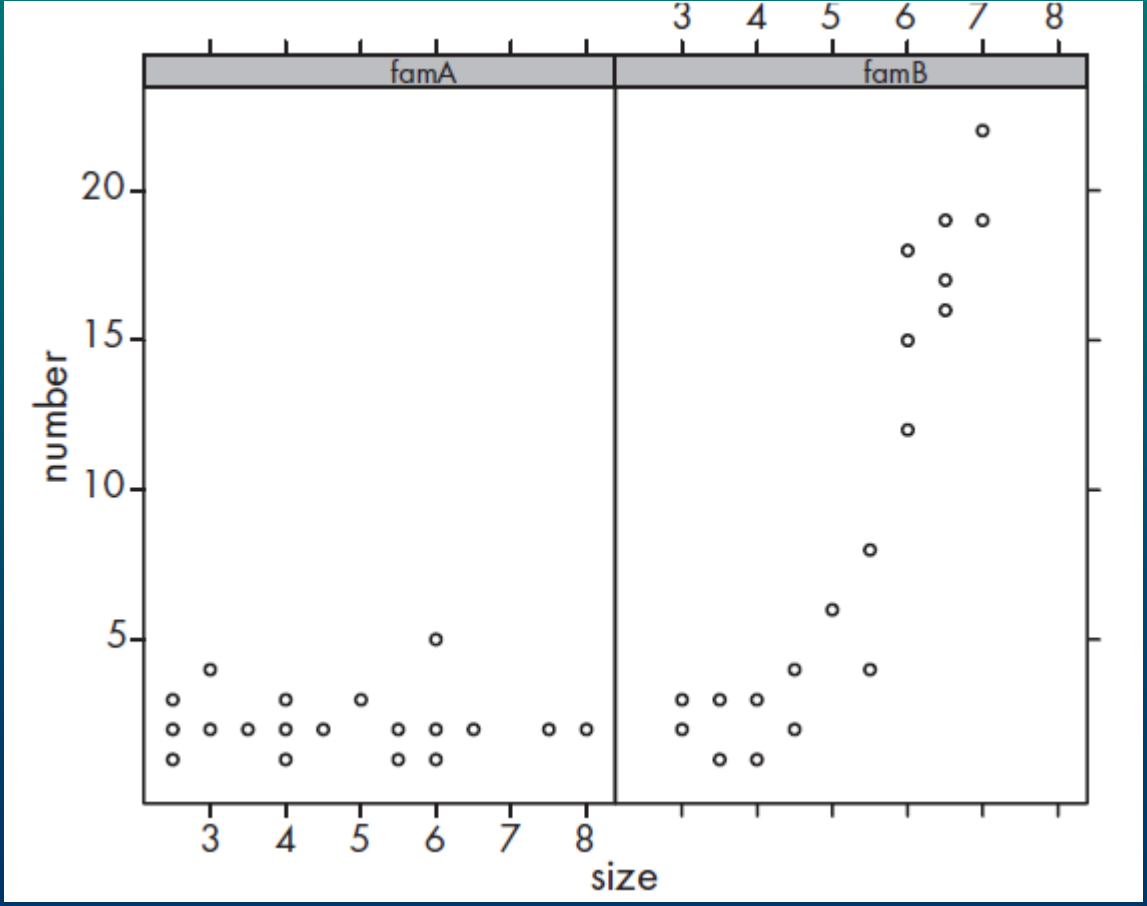
What is the shape of the relationship?

Variables

ANT: famA, famB

size

number

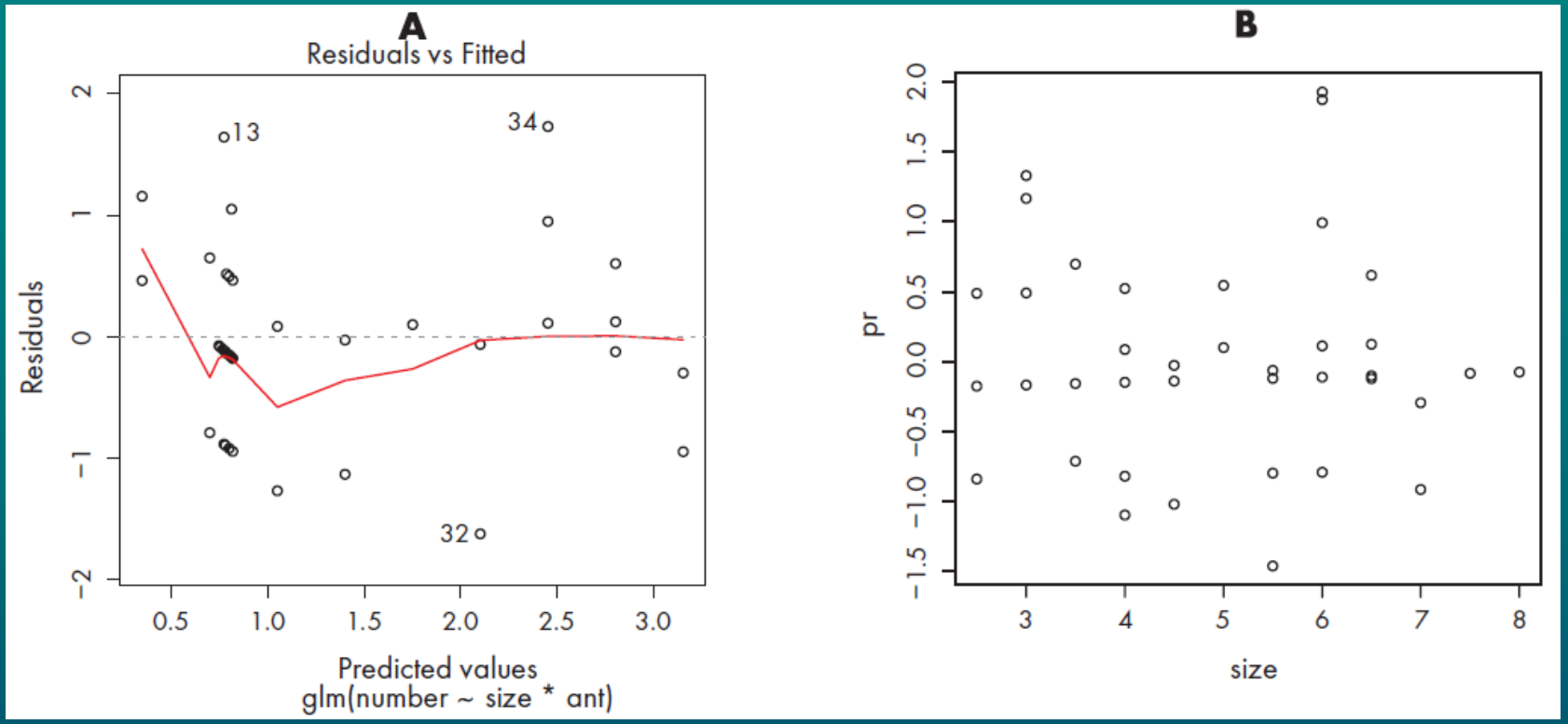


$$\log(\mu_{ij}) = \alpha + ANT_j + \beta size_i + \delta_j size_i ,$$

kde $number_{ij} \sim Poi(\mu_{ij})$, nezávisle pro jednotlivá pozorování.

```
> m1 <- glm(number ~ size*ant, family=poisson)
> anova(m1, test="Chi")
...
         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                        39    215.561
size      1   93.395         38    122.167 4.284e-22
ant       1   75.555         37     46.612 3.554e-18
size:ant  1   25.804         36     20.808 3.779e-07
```

```
> summary(m1)
...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.89794    0.64904   1.383 0.166512
size         -0.02154    0.12456  -0.173 0.862735
antfamB      -2.66924    0.80637  -3.310 0.000932 ***
size:antfamB  0.70407    0.14579   4.829 1.37e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 215.561  on 39  degrees of freedom
Residual deviance:  20.808  on 36  degrees of freedom
AIC: 153.15
```

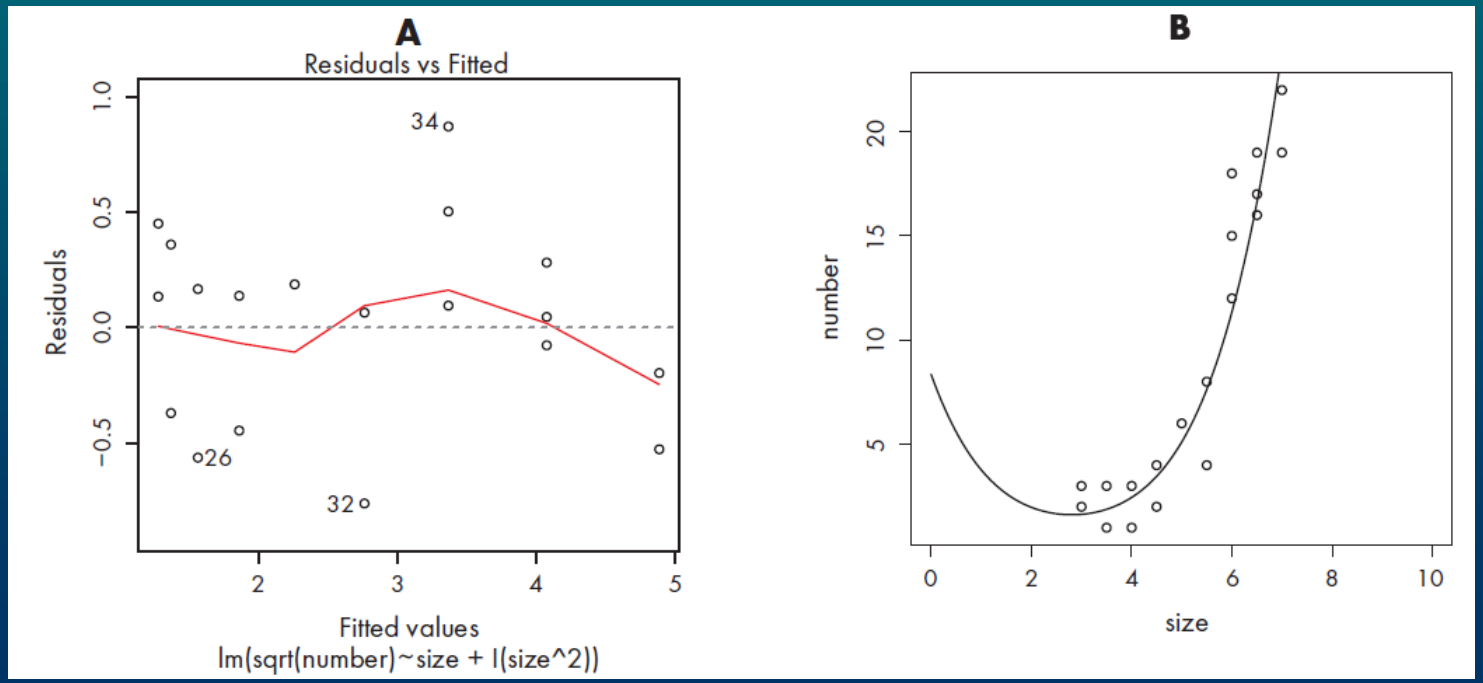


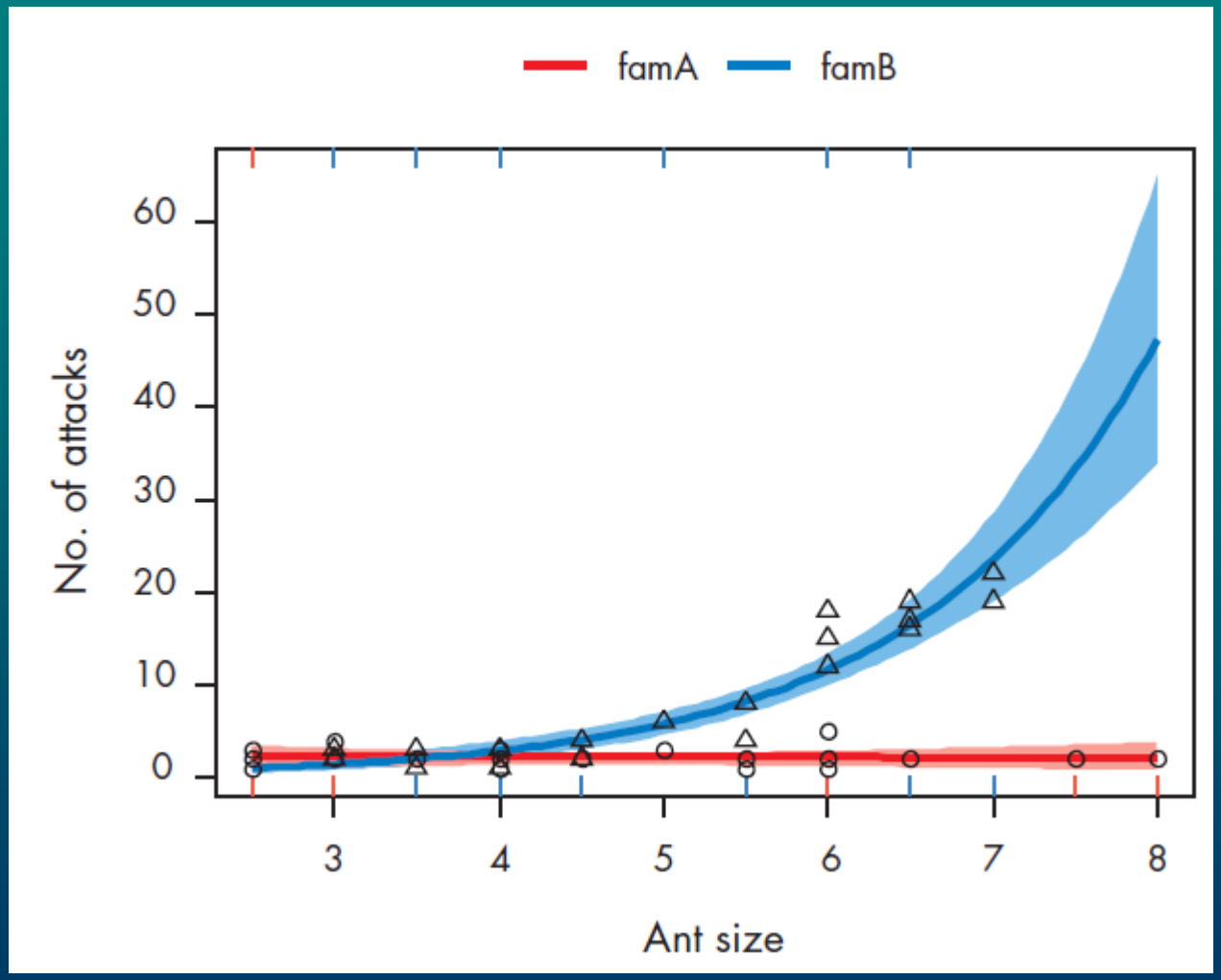

```
> m2 <- glm(number ~ poly(size,2)*ant, poisson)
> anova(m1, m2, test="Chi")
Analysis of Deviance Table

Model 1: number ~ size * ant
Model 2: number ~ poly(size, 2) * ant
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        36    20.8084
2        34    20.7673  2   0.0411    0.9797
```

$$\sqrt{number_i} = \alpha + \beta size_i + \gamma size_i^2 + \varepsilon_i,$$

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

```
> m3 <- lm(sqrt(number) ~ size + I(size^2), subset=(ant=="famB"))
> anova(m3)
Analysis of Variance Table
Response: sqrt(number)
          Df  Sum Sq Mean Sq  F value    Pr(>F)
size       1 28.3476 28.3476 161.0631 4.253e-10 ***
I(size^2)  1  1.1930  1.1930   6.7783   0.01855 *
Residuals 17  2.9921  0.1760
```





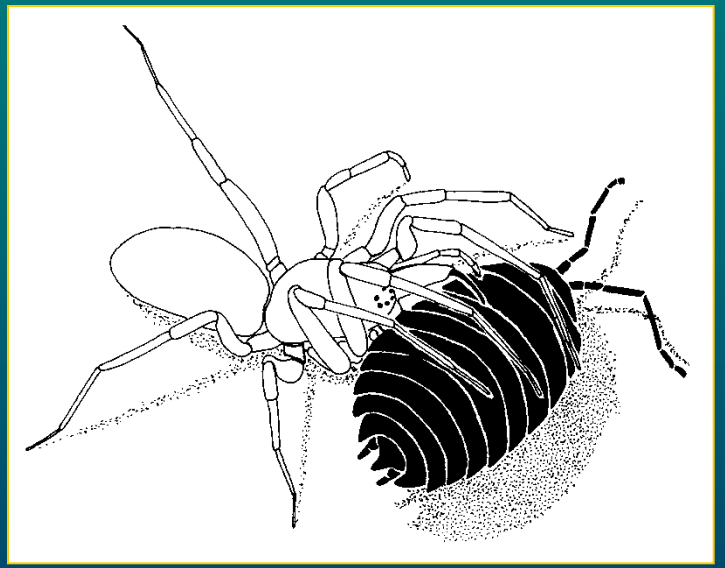
$$number = e^{-1.77 + 0.68size}.$$

$$number = e^{0.89}$$

3-way ANOVA

Background

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



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.

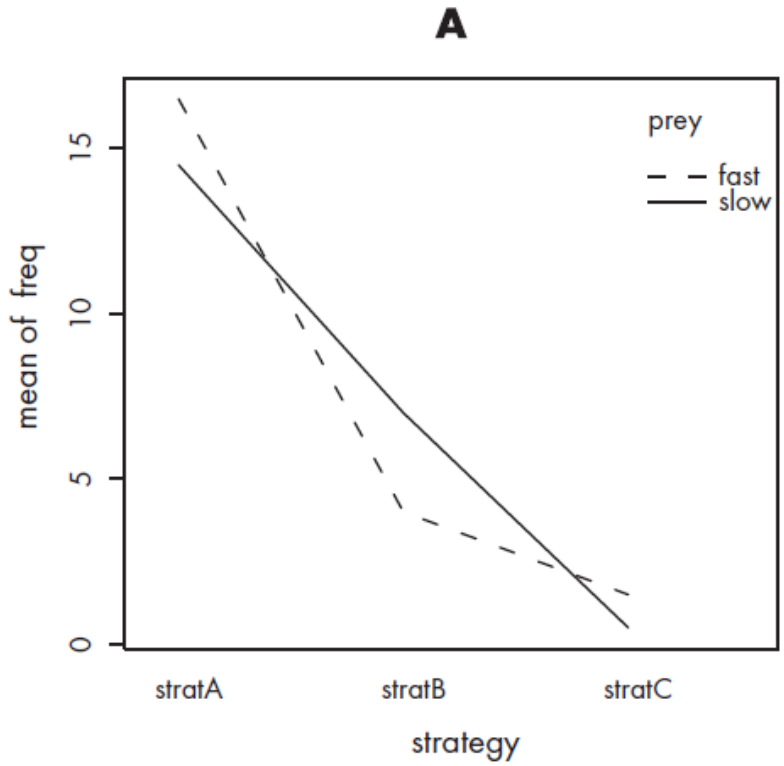
Hypotheses

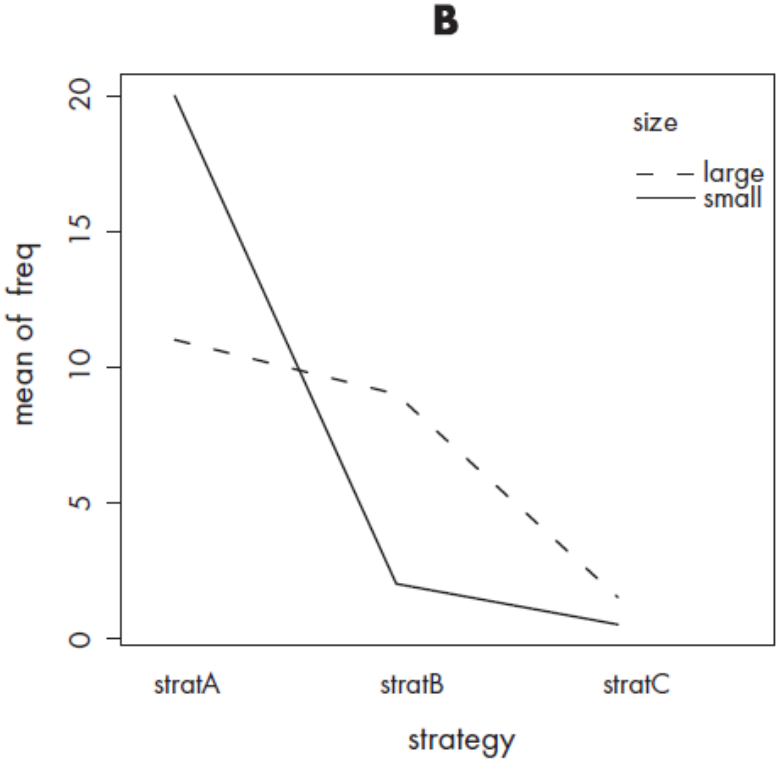
Is use of strategy influenced by prey size and its movement?
If so which prey is captured by strategy A, B and C?

Variables

PREY: fast, slow
SIZE: large, small
STRATEGY: stratA, stratB, stratC
freq

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





$$\log(\mu_{ijk}) = \alpha + STRATEGY_i + SIZE_j + PREY_k + STRATEGY{:}PREY_{ik} + STRATEGY{:}SIZE_{ij} + SIZE{:}PREY_{jk} + STRATEGY{:}SIZE{:}PREY_{ijk},$$

kde $freq_{ijk} \sim Poi(\mu_{ijk})$, nezávisle pro jednotlivá pozorování.

```
> m1 <- glm(freq ~ strategy*size*prey, family=poisson)
> summary(m1)

Call:
glm(formula = freq ~ strategy * size * prey, family = poisson)

Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0  0  0

Coefficients:
                                  Estimate Std. Error    z value Pr(>|z|)
(Intercept)                      2.485e+00  2.887e-01      8.608   <2e-16
strategystratB                  -4.055e-01  4.564e-01     -0.888   0.3744
strategystratC                  -1.792e+00  7.638e-01     -2.346   0.0190
sizesmall                        5.596e-01  3.619e-01      1.546   0.1220
preyslow                        -1.823e-01  4.282e-01     -0.426   0.6702
strategystratB:sizesmall        -2.594e+01  6.965e+04 -0.000372   0.9997
strategystratC:sizesmall        -1.253e+00  1.277e+00     -0.981   0.3266
strategystratB:preyslow          4.055e-01  6.390e-01      0.635   0.5257
strategystratC:preyslow         -5.108e-01  1.297e+00     -0.394   0.6938
sizesmall:preyslow               8.224e-02  5.325e-01      0.154   0.8773
strategystratB:sizesmall:preyslow  2.438e+01  6.965e+04  0.000350   0.9997
strategystratC:sizesmall:preyslow -2.269e+01  6.965e+04 -0.000326   0.9997
```

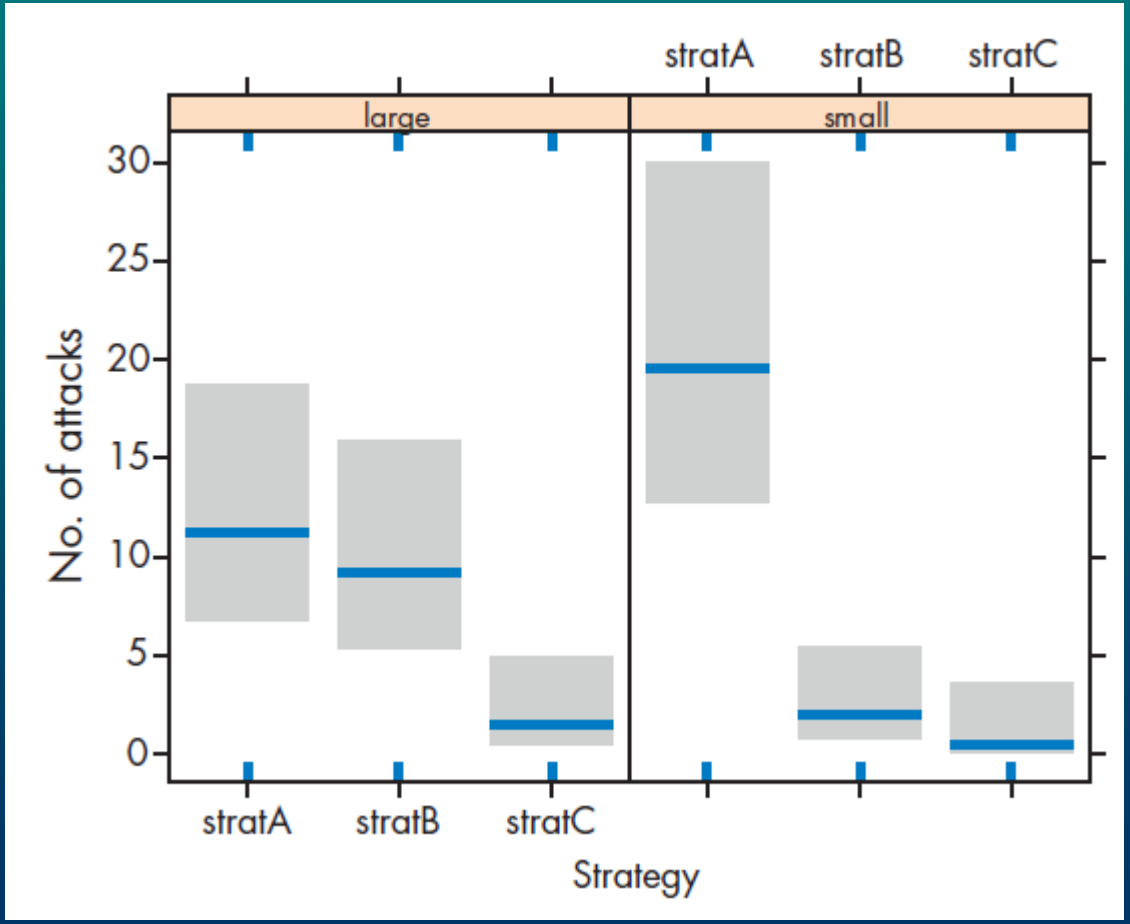
```
> anova(m1, test="Chi")
...
                  Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                 11     87.966
strategy           2   64.205          9     23.761 1.143e-14
size               1    0.045          8     23.715    0.831
prey               1    0.000          7     23.715    1.000
strategy:size      2   15.939          5      7.776 3.458e-04
strategy:prey      2    2.962          3      4.814    0.227
size:prey          1    0.507          2      4.307    0.476
strategy:size:prey 2    4.307          0  3.033e-10    0.116
```

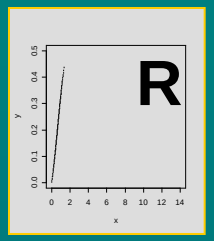
```
> m2 <- update(m1, ~.-strategy:size:prey)
> anova(m2, test="Chi")
...
             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                            11     87.966
strategy      2   64.205          9     23.761 1.143e-14
size          1    0.045          8     23.715     0.831
prey          1    0.000          7     23.715     1.000
strategy:size 2   15.939          5      7.776 3.458e-04
strategy:prey 2    2.962          3      4.814     0.227
size:prey     1    0.507          2      4.307     0.476
```

```
> m3 <- update(m2, ~.-strategy:prey)
> anova(m3, test="Chi")
...
             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                            11     87.966
strategy      2   64.205          9     23.761 1.143e-14
size          1    0.045          8     23.715     0.831
prey          1    0.000          7     23.715     1.000
strategy:size 2   15.939          5      7.776 3.458e-04
size:prey     1    0.045          4      7.731     0.831
```

```
> summary(m3)

Call:
glm(formula = freq ~ strategy + size + prey + strategy:size +
    size:prey, family = poisson)

Deviance Residuals:
      1       2       3       4       5       6       7
-0.3233  1.2076 -1.0111 -0.2297  0.3990 -0.4079  0.3227
      8       9      10      11      12
-1.9777  0.6395  0.2194 -0.4077  0.3585

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                2.42088    0.26010   9.307  < 2e-16 ***
strategystratB            -0.20067    0.31782  -0.631 0.527782
strategystratC            -1.99243    0.61546  -3.237 0.001207 **
sizesmall                  0.55237    0.34042   1.623 0.104669
preyslow                  -0.04652    0.30508  -0.152 0.878805
strategystratB:sizesmall  -2.10191    0.61318  -3.428 0.000608 ***
strategystratC:sizesmall  -1.69645    1.18481  -1.432 0.152193
sizesmall:preyslow         0.09097    0.42662   0.213 0.831142
```

```
> attacks <- tapply(predict(m3,type="response"), list(size,strategy), mean)
> attacks
       stratA stratB stratC
large      11      9    1.5
small      20      2    0.5
```





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:

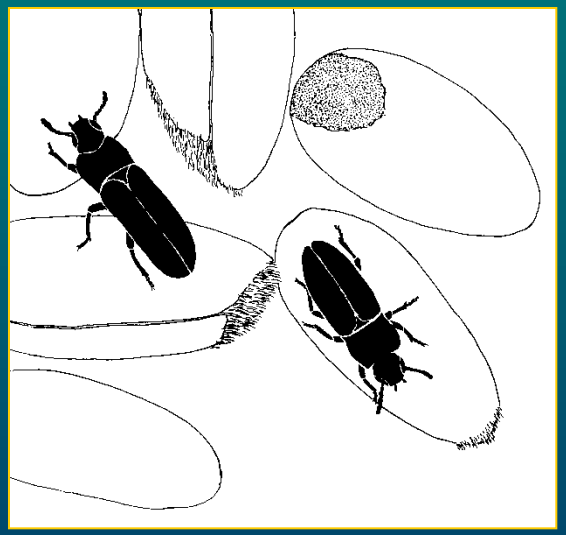
`log` (default)
`sqrt`
`identity`

- 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

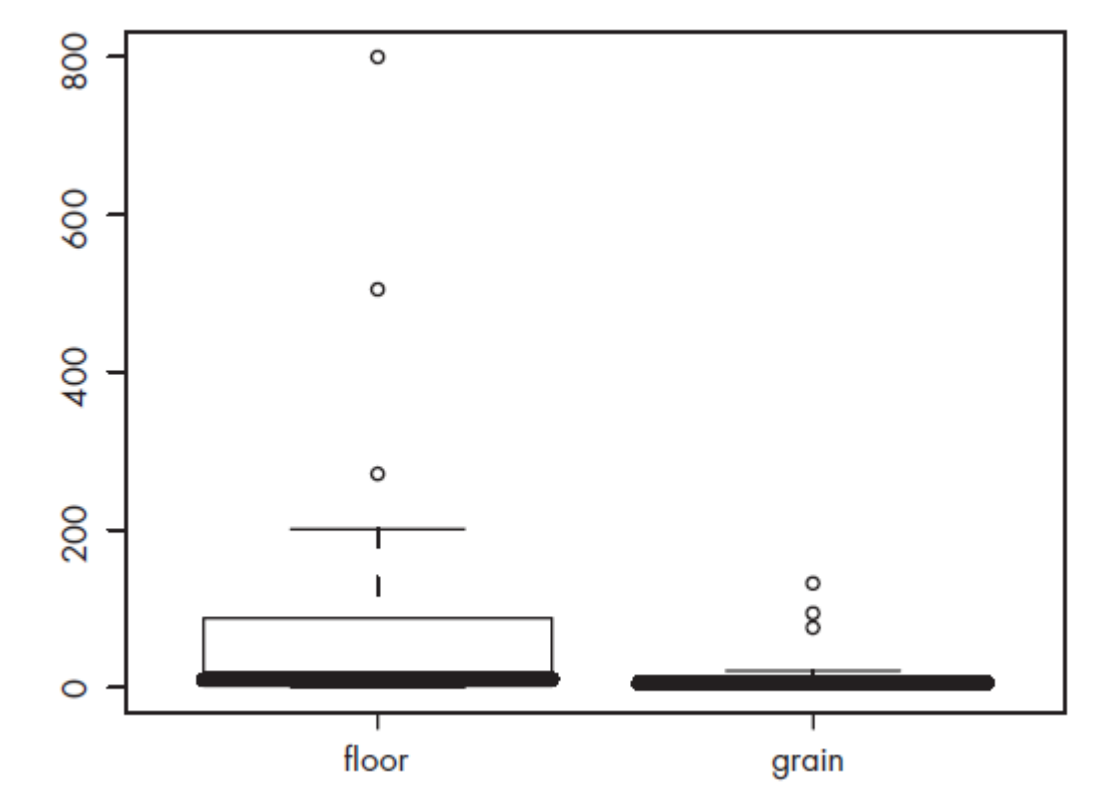
Is density of beetles similar on both places?

If not how different it is?

Variables

PLACE: floor, grain

density



$$\log(\mu_j) = \alpha + PLACE_j ,$$

kde $density_j \sim Poi(\mu_j)$, nezávisle pro jednotlivé pasti.

```
> m1 <- glm(density ~ place, family=quasipoisson)
> anova(m1, test="F")
...
      Df Deviance Resid. Df Resid. Dev      F  Pr(>F)
NULL                     49     8026.5
place  1   1350.1         48     6676.4 6.0434 0.01762 *
---
```

```
> summary(m1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.5161     0.3125   14.45   <2e-16 ***
placegrain   -1.6280     0.7715   -2.11   0.0401 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 223.3983)
```

$$\log(\mu_j) = \alpha + PLACE_j ,$$

kde $density_j \sim NB(\mu_j, \theta)$, nezávisle pro jednotlivé pasti.

```
> tapply(density, place, var)/tapply(density, place, mean)
     floor      grain
386.58096   60.20546
```

```
> tapply(density, place, function(x) mean(x)^2/(var(x)-mean(x)))
     floor      grain
0.2372524 0.3033504
```

```
> library(MASS)
> m2 <- glm.nb(density ~ place)
> anova(m2)
Analysis of Deviance Table

Model: Negative Binomial(0.3318), link: log

Response: density

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     49     70.174
place  1    9.877        48     60.297     0.002
Warning message:
In anova.negbin(m2) : tests made without re-estimating 'theta'
```

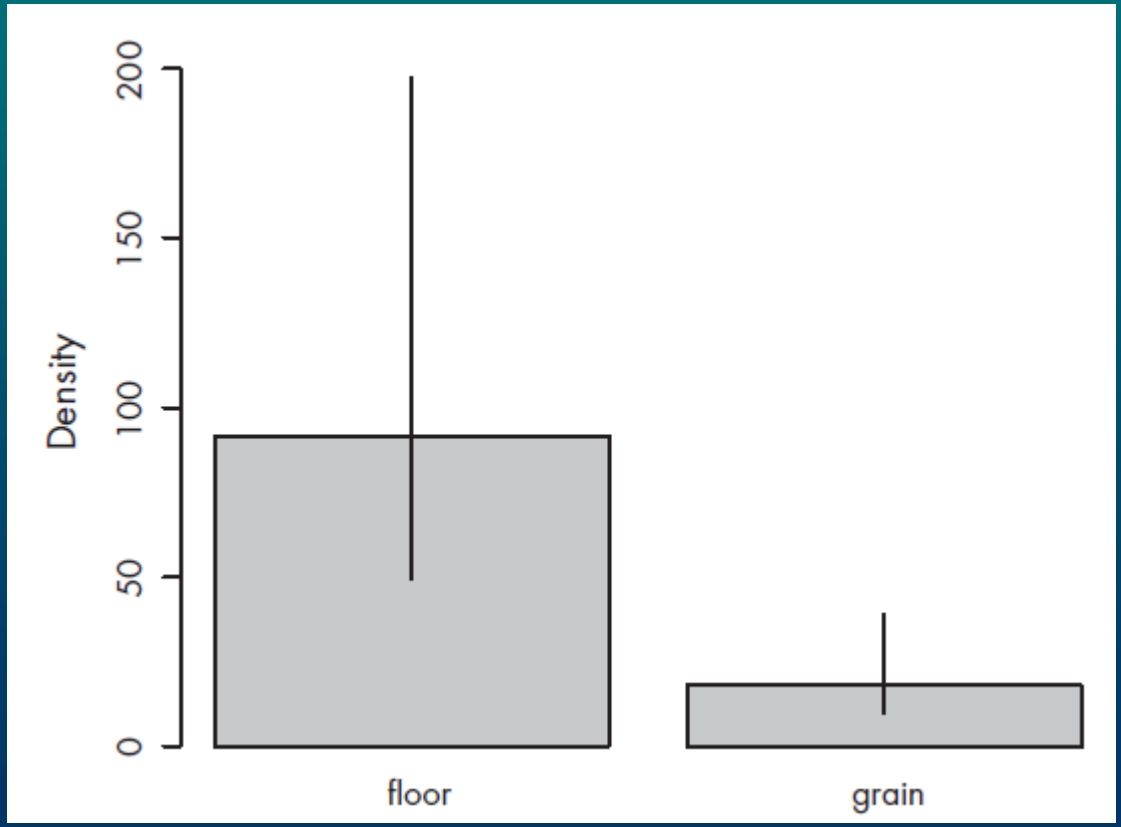
```
> summary(m2)

Call:
glm.nb(formula = density ~ place, init.theta = 0.331844006124825,
    link = log)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.5161     0.3478  12.984  < 2e-16 ***
placegrain   -1.6280     0.4937  -3.297 0.000976 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.3318) family taken to be 1)

    Null deviance: 70.174  on 49  degrees of freedom
Residual deviance: 60.297  on 48  degrees of freedom
AIC: 430.95

Number of Fisher Scoring iterations: 1


              Theta:  0.3318
          Std. Err.:  0.0610

 2 x log-likelihood:  -424.9480
```

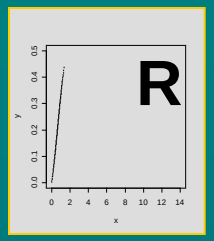
```
> a <- split(x=density, f=place)
> m3 <- glm.nb(floor ~ 1)
> summary(m3)
...
    Null deviance: 31.307  on 24  degrees of freedom
Residual deviance: 31.307  on 24  degrees of freedom
AIC: 245.47

Number of Fisher Scoring iterations: 1


              Theta:  0.2915
          Std. Err.:  0.0719

 2 x log-likelihood:  -241.4670
```

```
> m4 <- glm.nb(grain ~ 1)
> summary(m4)
...
    Null deviance: 29.197  on 24  degrees of freedom
Residual deviance: 29.197  on 24  degrees of freedom
AIC: 186.78

Number of Fisher Scoring iterations: 1

              Theta:  0.399
          Std. Err.:  0.111

2 x log-likelihood:  -182.780
```

```
> 1-pchisq(0.701,1)
[1] 0.4024479
```

```
> m5 <- glm.nb(density ~ place-1)
> exp(confint(m5))
Waiting for profiling to be done...
                2.5%     97.5%
placefloor  49.57777 197.24605
placegrain   9.67290  38.87768
```





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 ($-\infty$, $+\infty$)
- 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 anti-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:

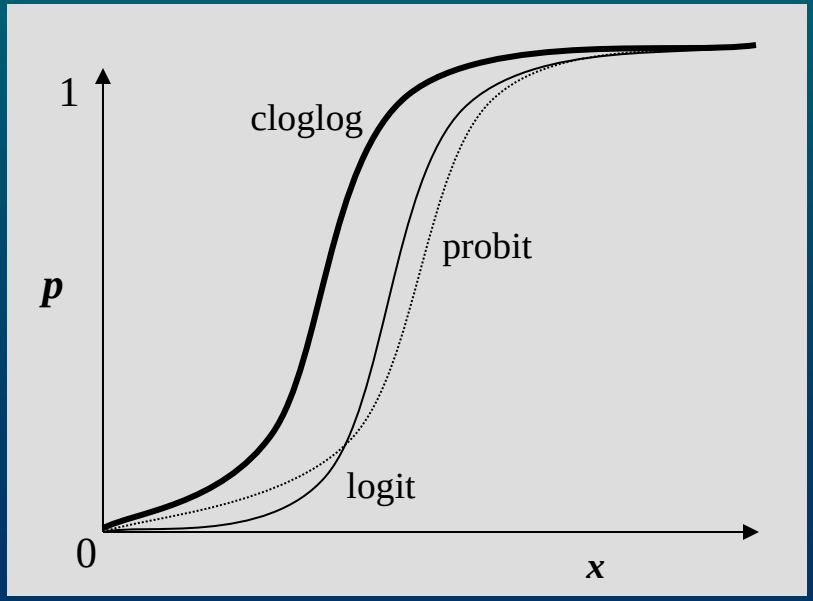
- logit (`logit`)

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

- probit (`probit`)

- complementary logit (`cloglog`)

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



Data format:

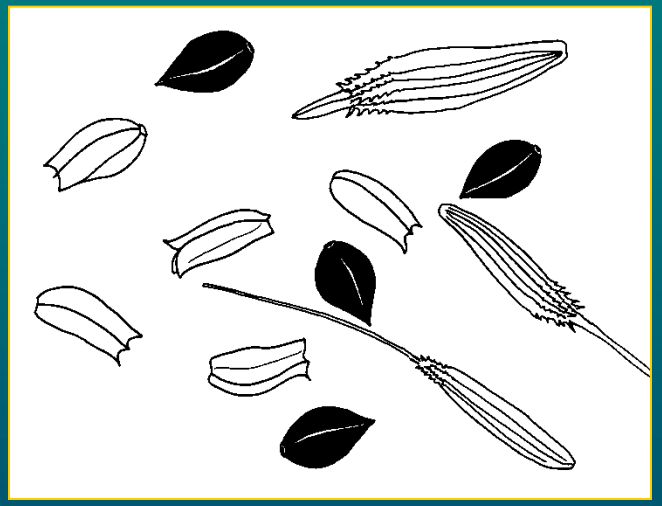
• **Binomial distribution ...** individuals within a group are homogenous
- two vectors (y, n-y) or (y, n) of integers

• **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

1-way ANOVA

Background

Some weed seeds may germinate following water priming (by rain) more than others thus attaining likely competitive advantage.



Design

The effect of water priming on the germination of weed seeds of 4 genera was studied in the laboratory. Each of 5 days 400 seeds of each genus were sown (200 seeds on control and 200 seeds on wet soil). Altogether 2000 seeds per genus were sown. Germination was recorded thereafter. Based on assumption of similar conditions during 5 days, data from 5 days were pooled.

Hypotheses

- Does water priming promote germination?
- If it does was the effect similar for all four genera?
- Which species germinated most and least?

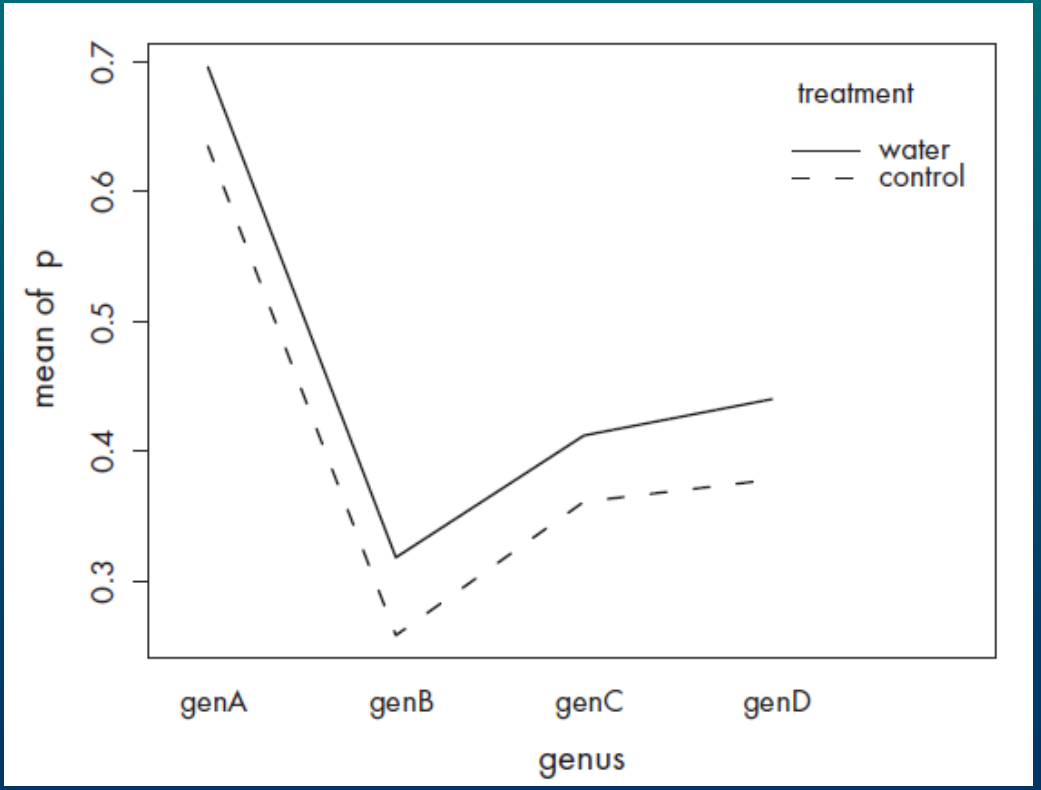
Variables:

TREATMENT: control, water

GENUS: genA, genB, genC, genD

germ

n



$$\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha + TREATMENT_j + GENUS_k + TREATMENT:GENUS_{jk},$$

kde $germ_{jk} \sim Bin(\pi_{jk}, n_{jk})$, nezávisle pro jednotlivé půdy.

```
> y <- cbind(germ, n-germ)
> m1 <- glm(y ~ genus*treatment, family=binomial)
> anova(m1, test="Chi")
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                              7     669.34
genus           3   638.74         4      30.60 4.026e-138
treatment       1    30.23         3       0.37  3.840e-08
genus:treatment 3     0.37         0  1.212e-13       0.95
```

```
> m2 <- update(m1, ~.-genus:pesticide)
> summary(m2)
...
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.56138    0.05256  10.681   <2e-16 ***
genusgenB       -1.59933    0.06860 -23.313   <2e-16 ***
genusgenC       -1.15462    0.06614 -17.457   <2e-16 ***
genusgenD       -1.06030    0.06583 -16.106   <2e-16 ***
treatmentwater   0.25859    0.04710   5.491    4e-08 ***
---
```

```
> 1/(1+ exp(-0.56138))
[1] 0.6367718
> 1/(1 + exp(-0.56138+1.59933))
[1] 0.2615457
```

```
> genus1 <- genus
> levels(genus1)
[1] "genA" "genB" "genC" "genD"
> levels(genus1)[3:4] <- "genCD"
> m3 <- glm(y ~ genus1 + treatment, binomial)
> anova(m2, m3, test="Chi")
Analysis of Deviance Table

Model 1: y ~ genus + treatment
Model 2: y ~ genus1 + treatment
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         3    0.37316
2         4    2.49523 -1 -2.12207   0.14519
```

```
> summary(m3)
...
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.56141    0.05256   10.68  < 2e-16 ***
genus1genB    -1.59933    0.06860  -23.31  < 2e-16 ***
genus1genCD   -1.10723    0.05749  -19.26  < 2e-16 ***
treatmentwater  0.25852    0.04709    5.49 4.02e-08 ***
---
```

```
> genus2 <- genus1
> levels(genus2)
[1] "genA"  "genB"  "genCD"
> levels(genus2)[2:3] <- "genBCD"
> m4 <- glm(y ~ genus2 + treatment, binomial)
> anova(m3, m4, test="Chi")
Analysis of Deviance Table

Model 1: y ~ genus1 + treatment
Model 2: y ~ genus2 + treatment
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4      2.495
2         5     73.684 -1  -71.189 3.246e-17
```

```
> ge <- tapply(predict(m3,type="response"), list(treatment,genus1), mean)
> ge
             genA      genB    genCD
control 0.6367787 0.2615513 0.366835
water   0.6942213 0.3144487 0.428665
```

Effect size

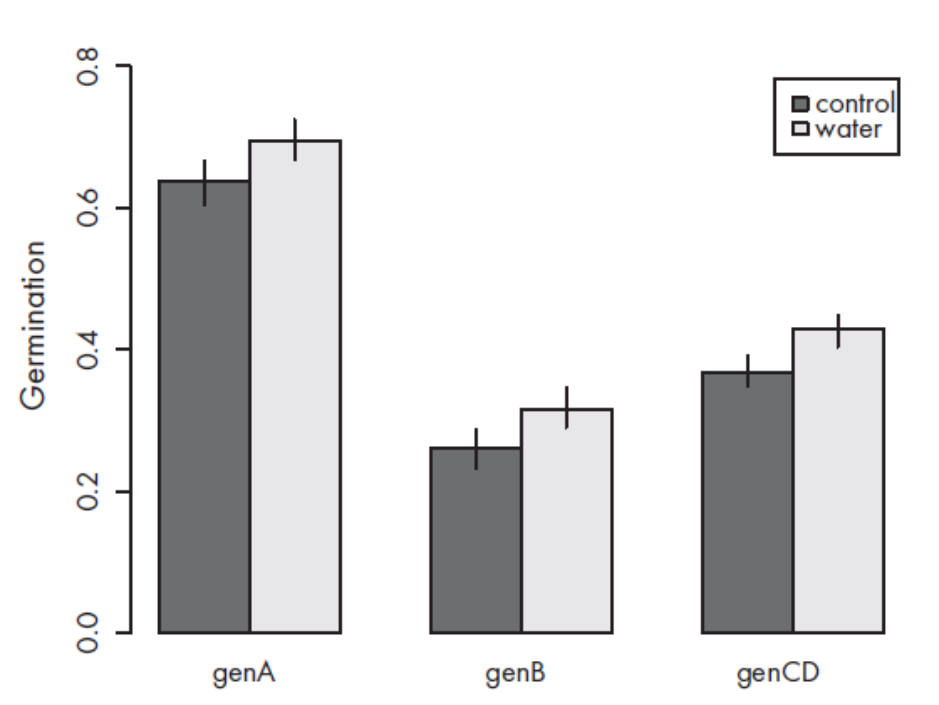
- statistical and biological effects are not identical
- statistical effects are affected by precision of measurements, number of measurements, type of test
- Cohen's coefficient:

$$h = \left| 2 \arcsin \sqrt{p_1} - 2 \arcsin \sqrt{p_2} \right|$$

- h < 0.2 … weak effect
- h > 0.8 … strong effect

```
> abs(2*asin(sqrt(ge[1,1]))-2*asin(sqrt(ge[2,1])))
[1] 0.1218512
```

```
> both <- paste(pesticide,genus1)
> m4 <- glm(y ~ factor(both) - 1, binomial)
> 1/(1+exp(-confint(m4)))
Waiting for profiling to be done...
                                      2.5%      97.5%
factor(both)control genA   0.6048442 0.6644666
factor(both)control genB   0.2315221 0.2857134
factor(both)control genCD  0.3485230 0.3908104
factor(both)water genA     0.6670153 0.7239840
factor(both)water genB     0.2896266 0.3473026
factor(both)water genCD    0.4044330 0.4477560
```



Over- / under-dispersion

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

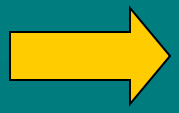
- overdispersion: variance is larger $\rightarrow \varphi > 1$
- underdispersion: variance is smaller $\rightarrow \varphi < 1$

• 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



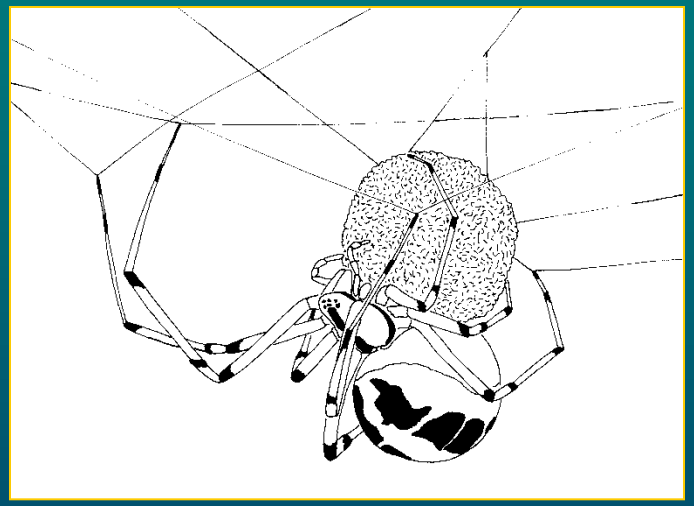
changes *P* values

• 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

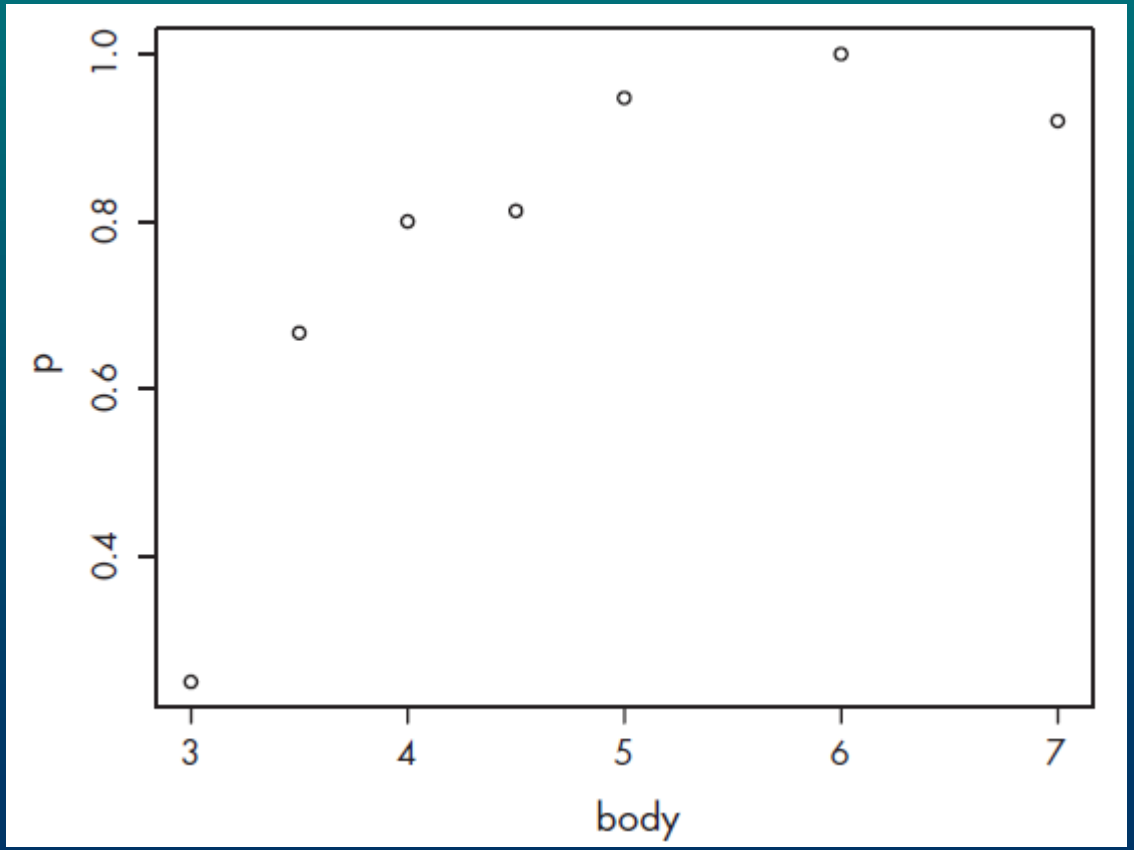
- Is eggsac production related to the body size?
- If it is what is the shape of the relationship?
- What is the model that can be used to predict eggsac production for spider sizes of 3–12 mm?

Variables:

body

n

eggs



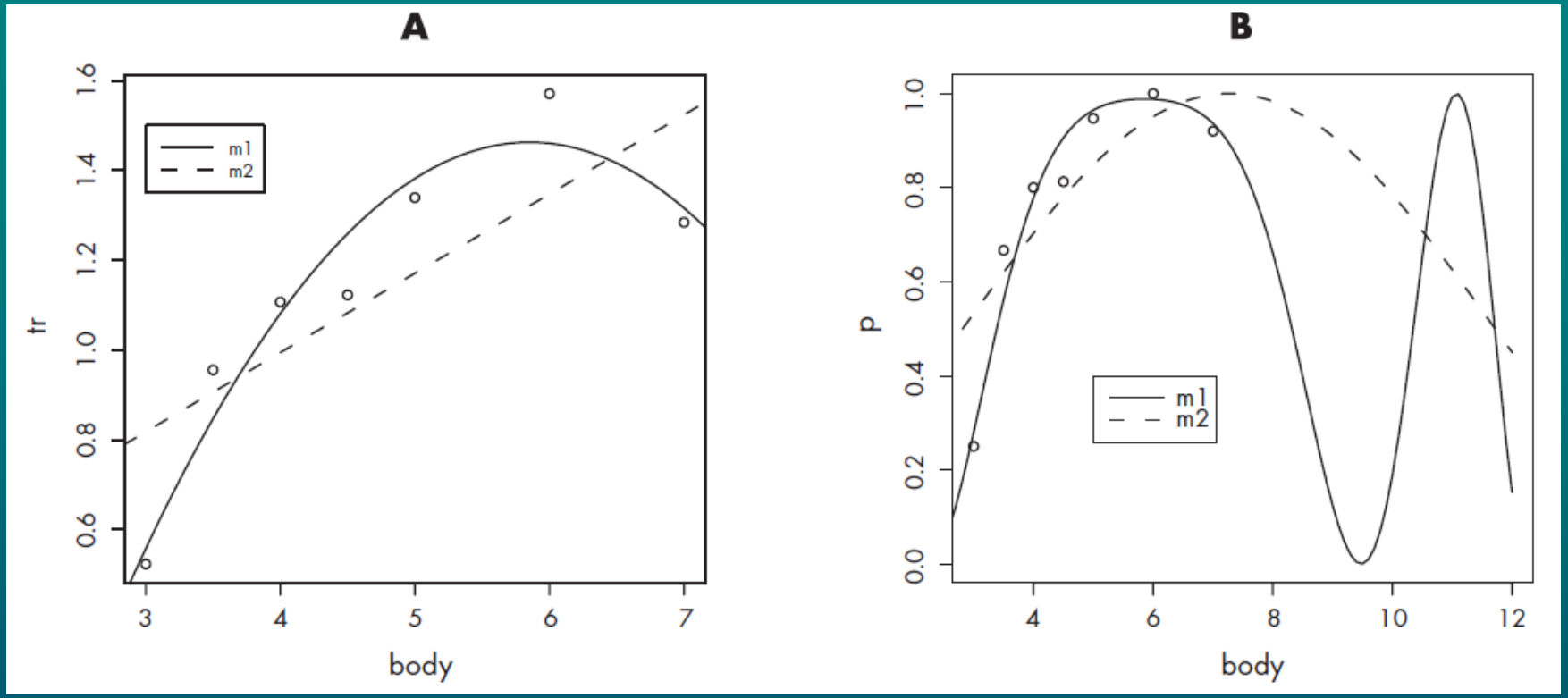
$$\arcsin\sqrt{p_i} = \alpha + \beta body_i + \gamma body_i^2 + \varepsilon_i,$$

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

```
> tr <- asin(sqrt(p))
> m1 <- lm(tr ~ body + I(body^2), weights=n)
> summary(m1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.34592    0.59329  -3.954  0.01676 *
body         1.30161    0.24776   5.254  0.00628 **
I(body^2)   -0.11121    0.02433  -4.571  0.01025 *
```

```
> m2 <- update(m1, ~.-I(body^2))
> summary(m2)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.28836    0.31429   0.918   0.4010
body         0.17649    0.06279   2.811   0.0375 *
```



$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta body_i + \gamma body_i^2,$$

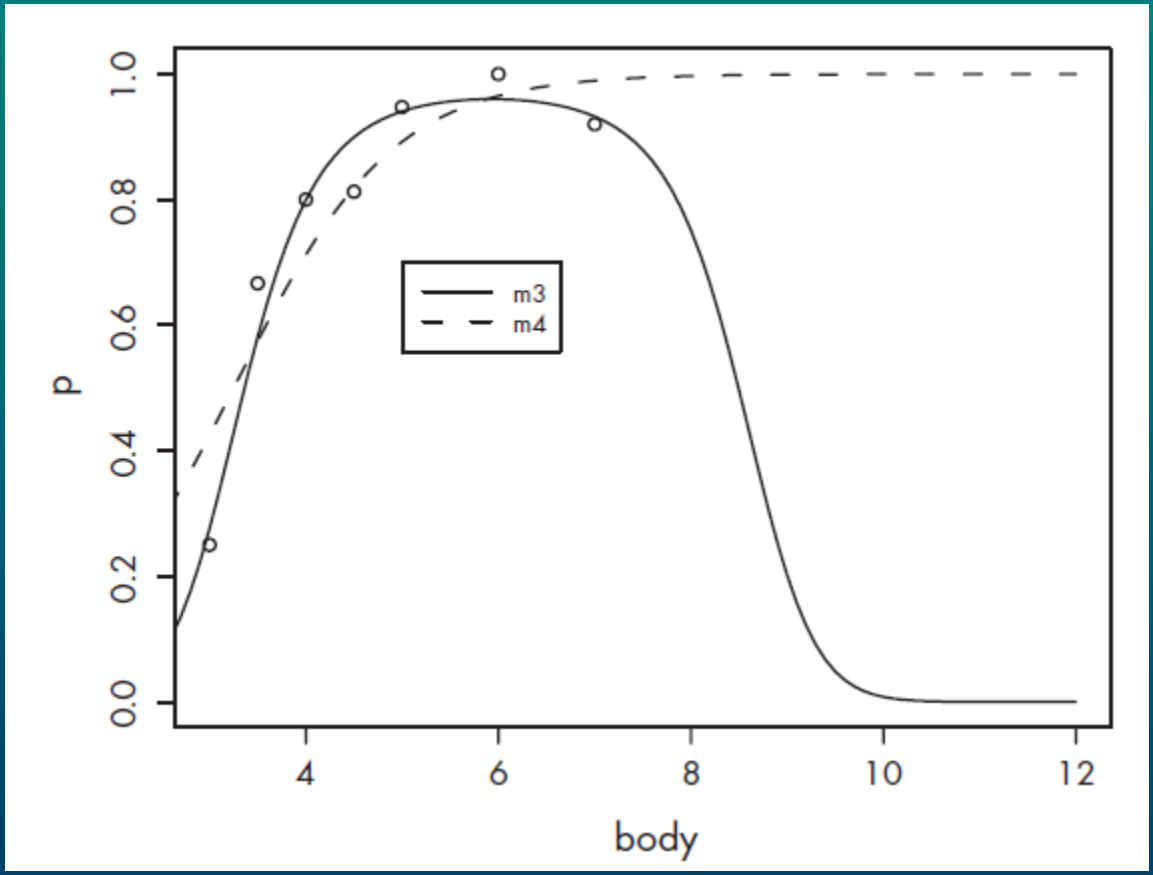
kde $eggs_i \sim Bin(\pi_i, n_i)$, nezávisle pro jednotlivé pavouky.

```
> y <- cbind(eggs, n-eggs)
> m3 <- glm(y ~ body + I(body^2), family=binomial)
> summary(m3)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.7857     3.8482  -3.582 0.000340 ***
body          5.7218     1.6771   3.412 0.000645 ***
I(body^2)    -0.4825     0.1695  -2.846 0.004427 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.2136  on 6  degrees of freedom
Residual deviance:  3.3357  on 4  degrees of freedom
```

```
> summary(m4)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.9270     1.1038  -3.558 0.000374 ***
body          1.2079     0.2756   4.383 1.17e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.214  on 6  degrees of freedom
Residual deviance: 11.072  on 5  degrees of freedom
```

```
> m5 <- update(m4, family=quasibinomial)
> summary(m5)
(Dispersion parameter for quasibinomial family taken to be 3.332466)
```

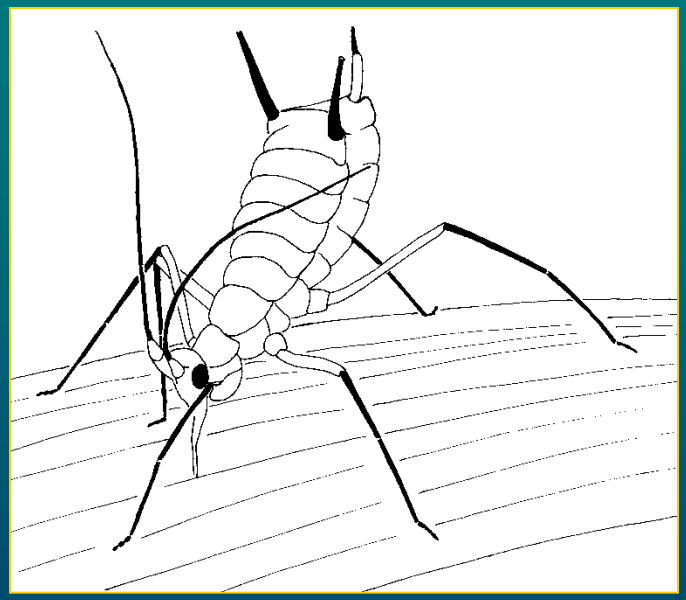
```
> anova(m5, test="F")
...
     Df Deviance Resid. Df Resid. Dev     F  Pr(>F)
NULL                     6     44.214
body  1   33.141         5     11.072 9.945 0.02528 *
```


p
0.0
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
body
m3
m4

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?

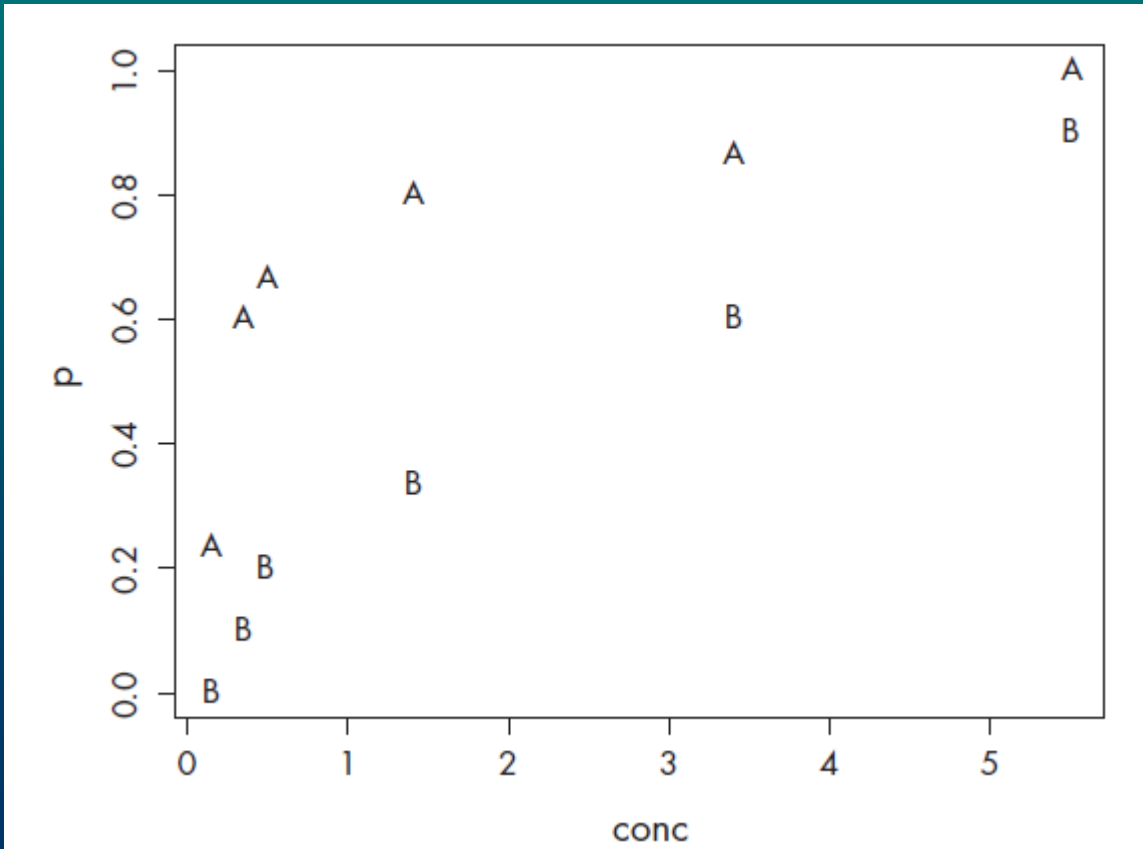
Variables:

SPECIES: A, B

conc

n

dead



$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + SPECIES_j + \beta \log(conc_i) + \delta_j \log(conc_i),$$

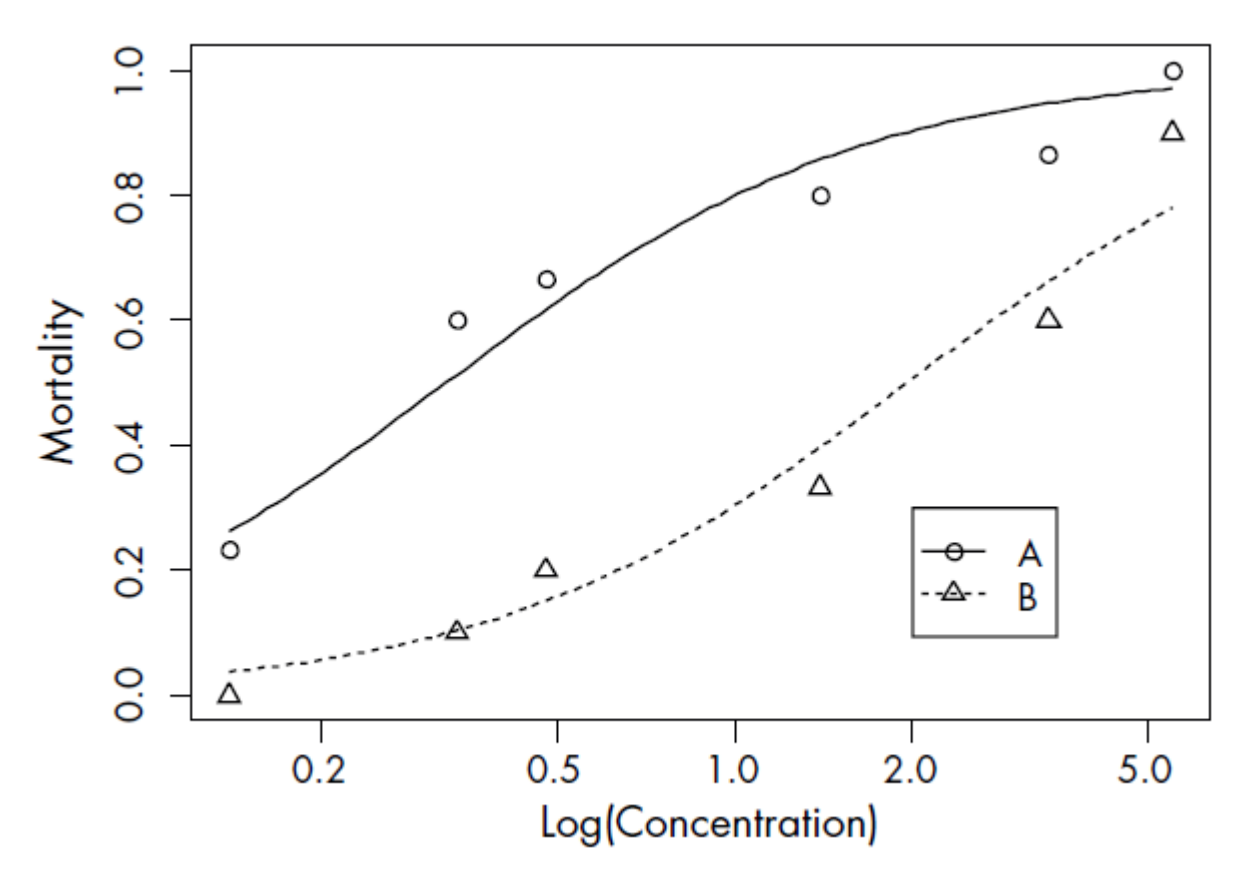
kde $dead_{ij} \sim Bin(\pi_{ij}, n_{ij})$, nezávisle pro jednotlivá pozorování.

```
> y <- cbind(dead, n-dead)
> m1 <- glm (y ~ log(conc)*species, binomial)
> anova(m1)
...
                  Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                    11    185.807
log(conc)          1  110.170         10     75.638 8.996e-26
species            1   62.087          9     13.551 3.286e-15
log(conc):species  1    1.343          8     12.207     0.246
```

```
> m2 <- update(m1, ~.-log(conc):species)
> anova(m2)
...
          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         11    185.807
log(conc)  1  110.170         10     75.638 8.996e-26
species    1   62.087          9     13.551 3.286e-15
```

```
> summary(m2)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.3825     0.2201   6.280 3.39e-10 ***
log(conc)     1.2328     0.1348   9.146  < 2e-16 ***
speciesB     -2.2117     0.3180  -6.955 3.52e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 185.807  on 11  degrees of freedom
Residual deviance:  13.551  on  9  degrees of freedom
```

$$100/(1 + \exp(-1.383 - 1.233\log(conc)))$$



$$100/(1 + \exp(0.829 - 1.233\log(conc)))$$

$$LC_{50} = \exp\left(-\frac{a}{b}\right)$$

```
> m3 <- glm(y ~ species + log(conc) - 1, binomial)
> summary(m3)
...
Coefficients:
          Estimate Std. Error z value Pr(>|z|)
speciesA    1.3825     0.2201   6.280 3.39e-10 ***
speciesB   -0.8293     0.2020  -4.106 4.02e-05 ***
log(conc)   1.2328     0.1348   9.146  < 2e-16 ***
```

```
> library(MASS)
> dose.p(m3, cf=c(1,3), p=0.5)
                Dose         SE
p = 0.5: -1.121418 0.1627097
```

$$\exp(-1.121) = 0.326$$

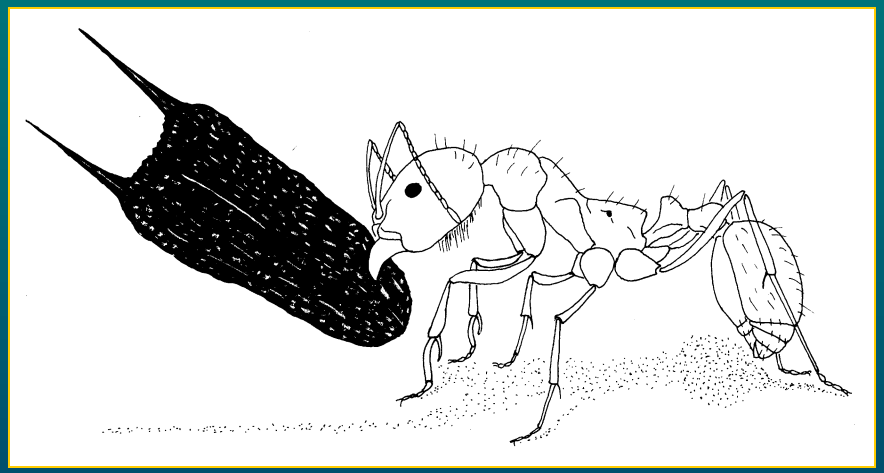
```
> dose.p(m3, cf=c(2,3), p=0.5)
                Dose         SE
p = 0.5: 0.6726813 0.159251
```

$$\exp(0.673) = 1.96.$$

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?

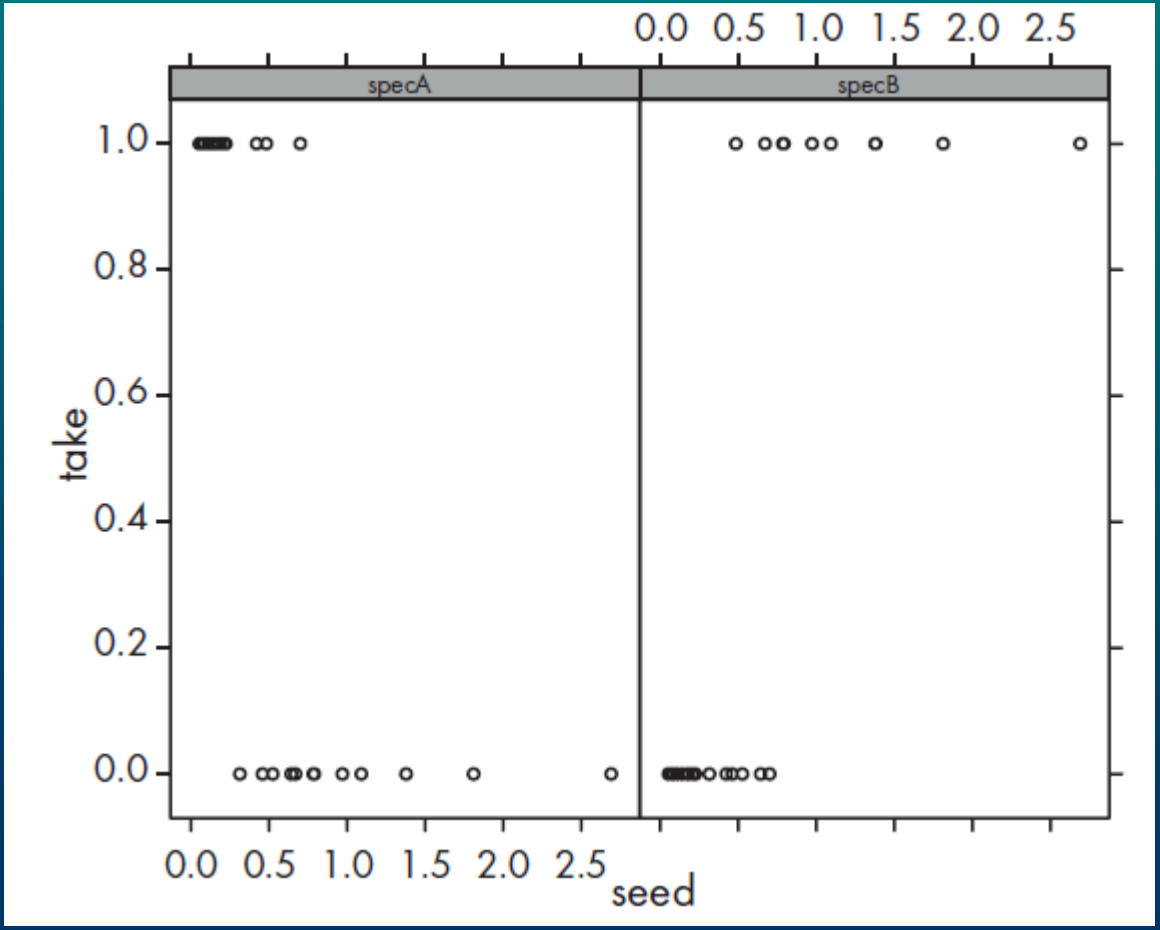
(The threshold size is defined as a size that is accepted with higher than 90% probability)

Variables:

SPECIES: specA, specB

seed

take



$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + SPECIES_j + \beta seed_{ij} + \delta_j seed_{ij},$$

kde $take_{ij} \sim Bin(\pi_{ij}, 1)$, nezávislé pro jednotlivé mravence.

```
> m1 <- glm(take ~ seed*species, family=binomial)
> summary(m1)
...
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.012      1.646   2.437  0.01480 *
seed                -8.346      3.315  -2.517  0.01182 *
speciesspecB       -10.957      3.697  -2.964  0.00304 **
seed:speciesspecB   19.147      6.141   3.118  0.00182 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.593  on 49  degrees of freedom
Residual deviance: 24.726  on 46  degrees of freedom
```

```
> anova(m1, test="Chi")
...
             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                           49     68.593
seed          1    0.054         48     68.539     0.817
species       1    0.325         47     68.214     0.568
seed:species  1   43.488         46     24.726 4.267e-11
```

```
> m2 <- glm(take ~ log(seed)*species, binomial)
> AIC(m1, m2)
   df      AIC
m1  4 32.72631
m2  4 32.23823
```

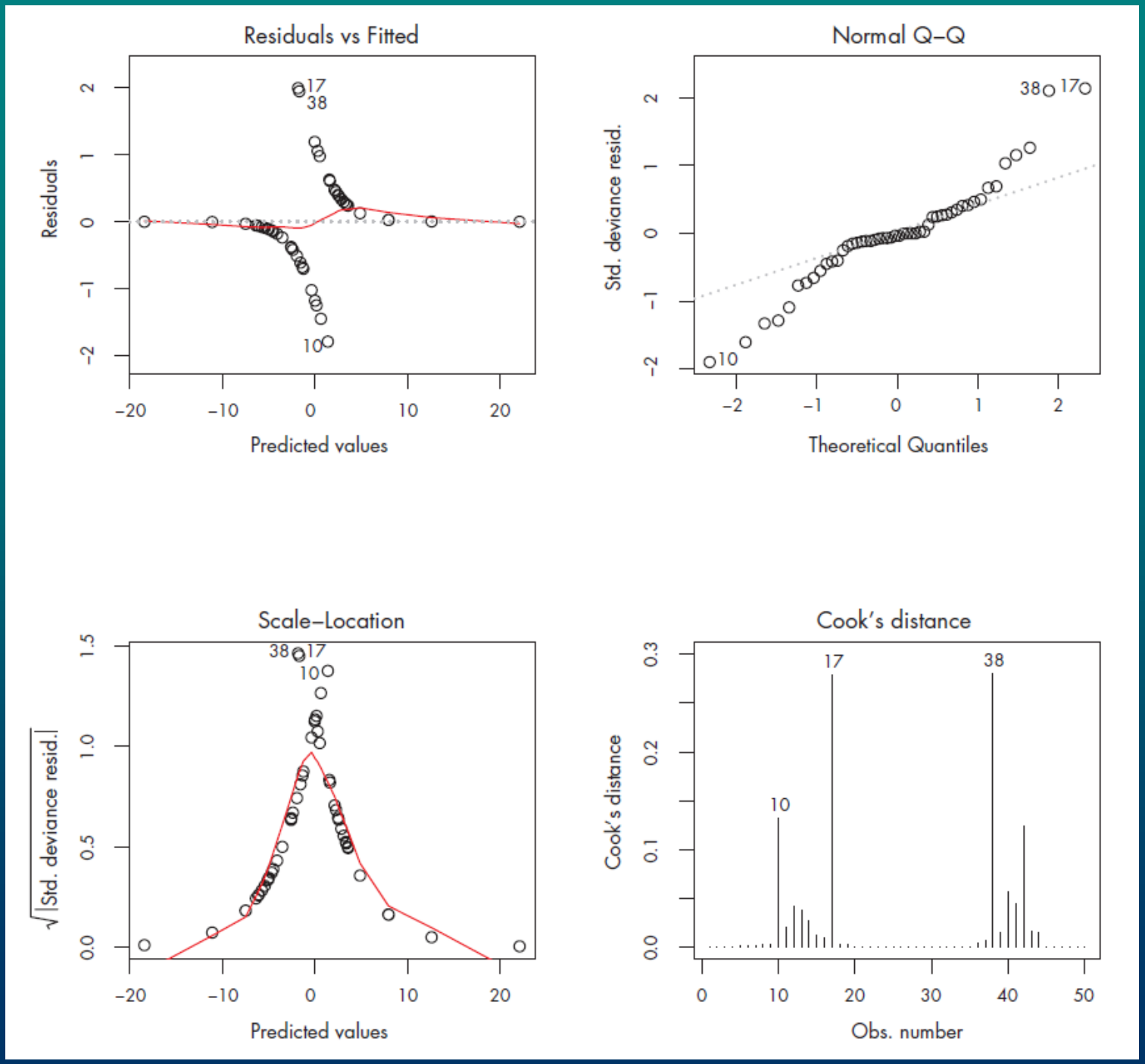
```
> m3 <- glm(take ~ seed*species, binomial(link=cloglog))
> AIC(m3)
[1] 31.63241
```

Coefficient of determination

- several for GLM models
- McFaden's coefficient – based on likelihood of models
- ranges from 0 to 1

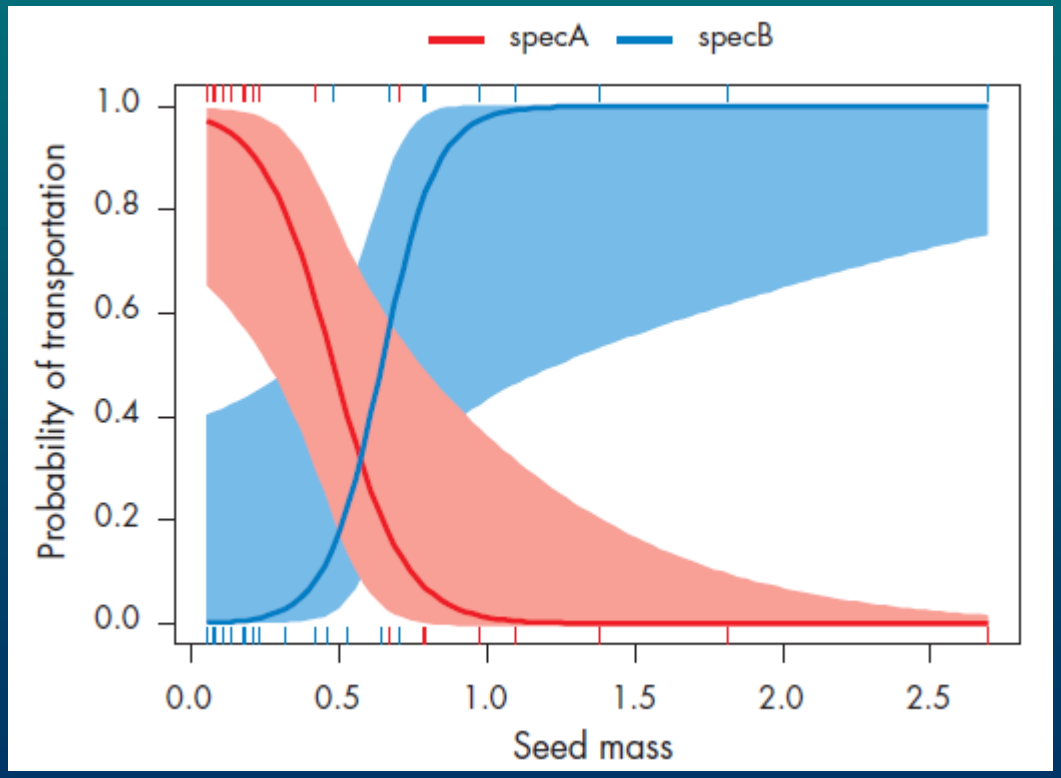
$$\rho^2 = 1 - \frac{\text{LogLik}_M}{\text{LogLik}_{M0}}$$

```
> m4 <- glm(take ~ 1, binomial)
> 1-logLik(m1)/logLik(m4)
'logLik'  0.6395213 (df=4)
```


Residuals vs Fitted
Residuals
Predicted values
17
38
10
Normal Q–Q
Std. deviance resid.
Theoretical Quantiles
38
17
10
Scale–Location
√|Std. deviance resid.|
Predicted values
38
17
10
Cook's distance
Cook's distance
Obs. number
17
38
10

$$x = \frac{\log\left(0.9/0.1\right) - a}{b}$$

```
> (log(0.9/0.1)-4.012)/-8.346
[1] 0.2174425
> (log(0.9/0.1)-4.012+10.957)/(-8.346+19.147)
[1] 0.8464239
```



$$\frac{1}{1+\exp\left(6.945-10.8seed\right)}.$$

$$\frac{1}{1+\exp\left(-4.012+8.346seed\right)}$$