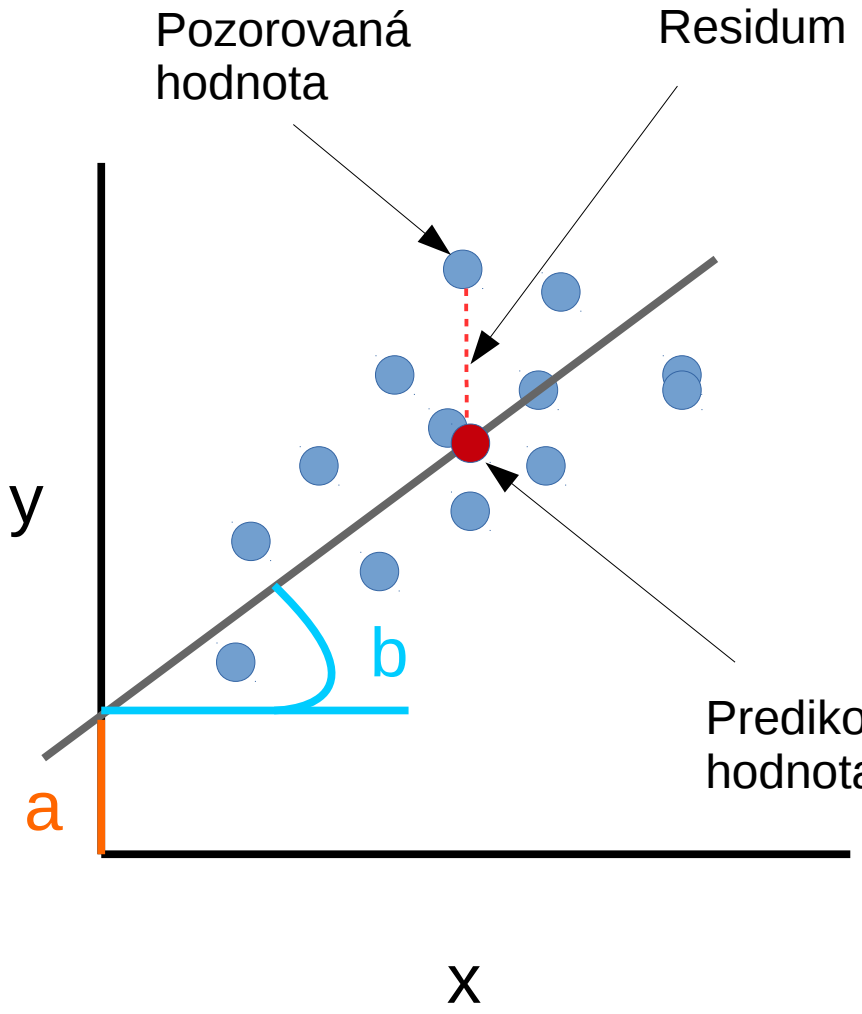
Moderní Statistické Metody II

Jakub Kreisinger
jakubkreisinger@seznam.cz

Program:

1. Opakování
2. Modelování “nelineárních” vztahů
3. Mixované modely
4. Modelování fylogenetických a prostorových autokorelcí

Opakování:



$$y = a + bx + e$$

y - vysvětlovaná proměnná
- závislá
- response variable
- dependent variable

x – vysvlětlující proměnná
- predictor
- nezávislá proměnná

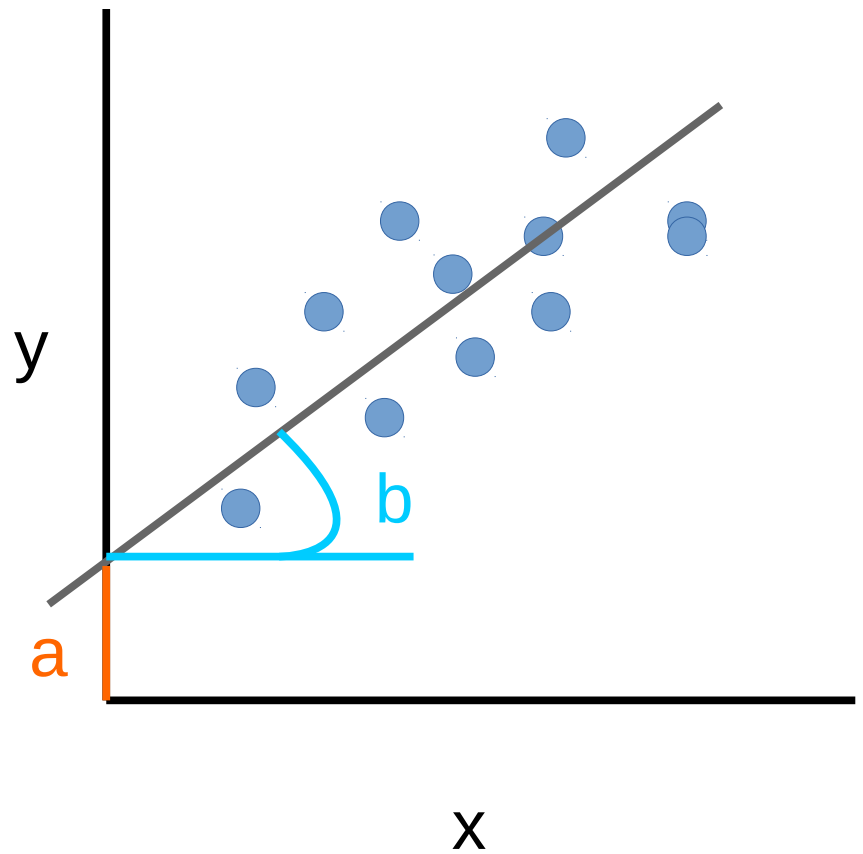
a - intercept
- Tam, kde regresní přímka protíná osu y

b – slope (sklon) regresní přímky

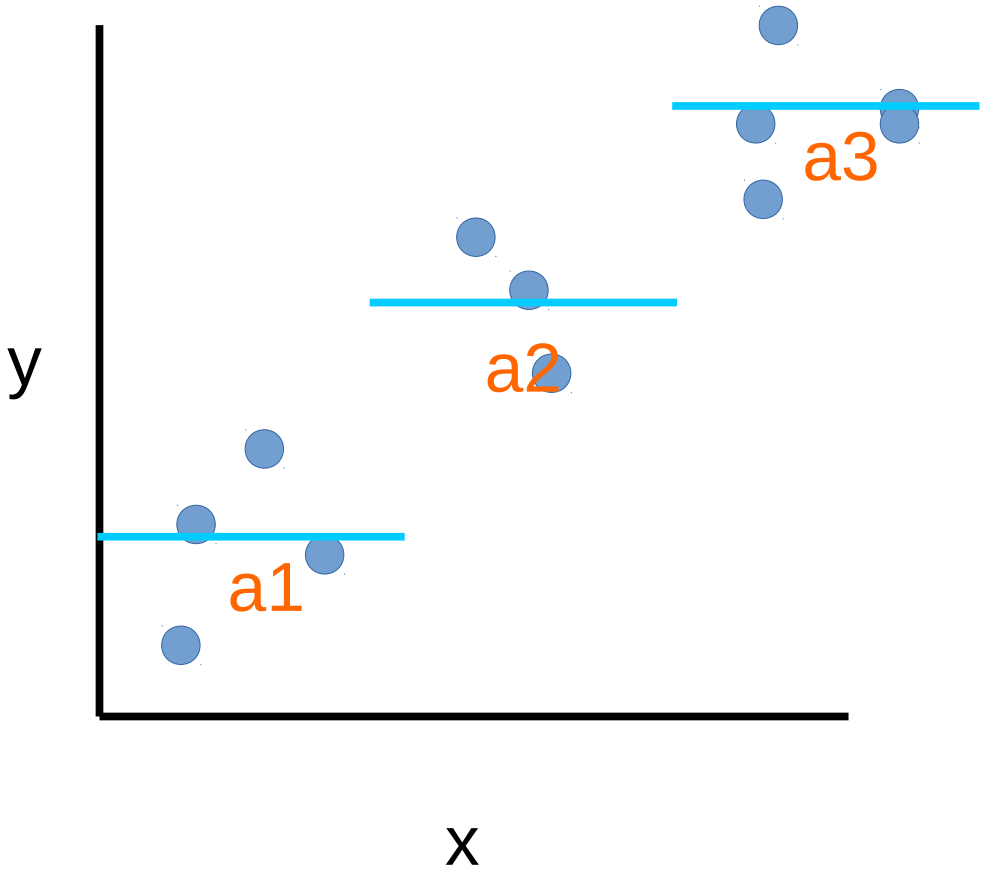
Opakování:

vysvětlující proměnná

Kontinuální
Lineární regrese



Kategorická (faktoriální)
ANOVA

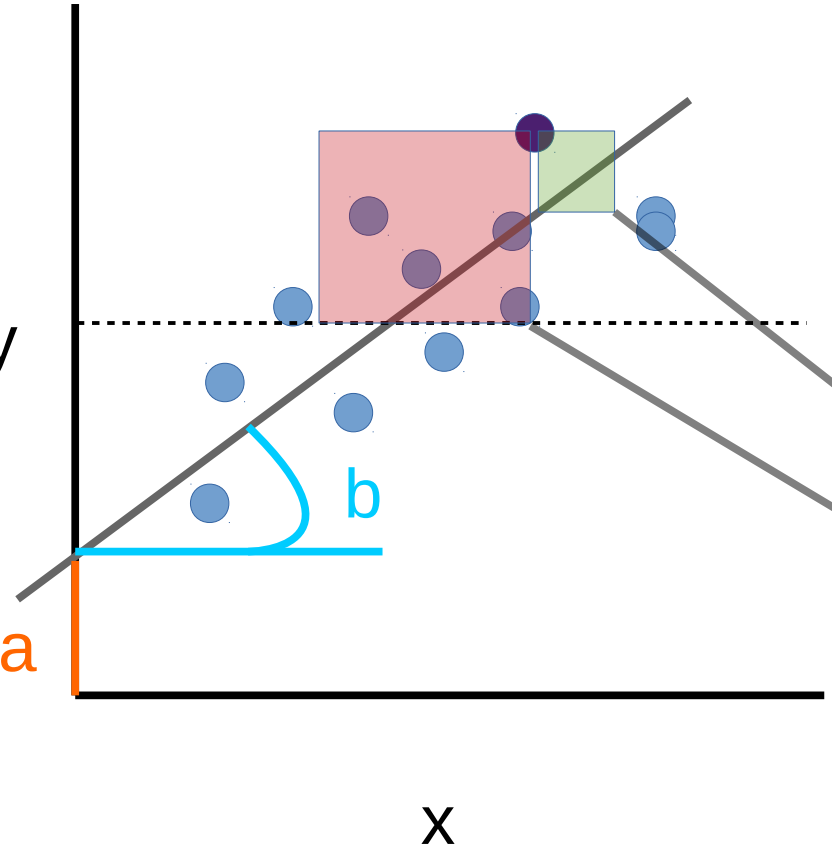


Opakování:

vysvětlující proměnná

Kontinuální

Lineární regrese




```
> data("cars")
> model<-lm(dist~speed,data=cars)
> summary(model)
Call:
lm(formula = dist ~ speed, data = cars)
Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 â€˜***â€™ 0.001 â€˜**â€™ 0.01
â€˜*â€™ 0.05 â€˜.â€™ 0.1 â€˜ â€™ 1
Residual standard error: 15.38 on 48 degrees of
freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:
0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12

> anova(model)
Analysis of Variance Table
Response: dist
          Df Sum Sq Mean Sq F value   Pr(>F)
speed      1  21186 21185.5  89.567 1.49e-12 ***
Residuals 48  11354   236.5
```

Opakování:

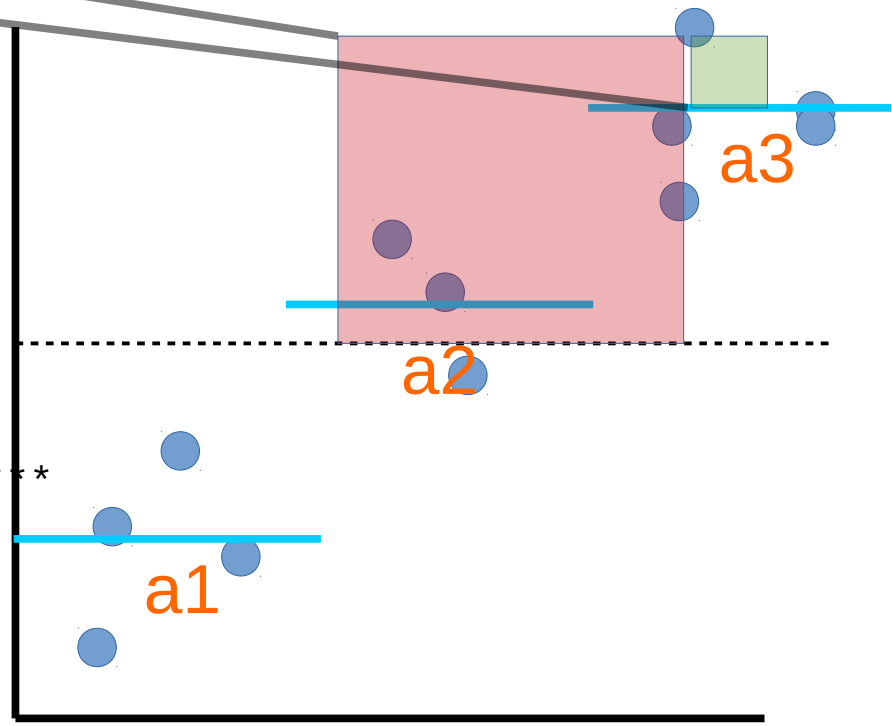
vysvětlující proměnná

```
> npk.aov <- aov(yield ~ block, npk)
> summary(npk.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
block        5  343.3   68.66   2.318 0.0861 .
Residuals   18  533.1   29.61
---

> npk.aov <- lm(yield ~ block, npk)
> summary(npk.aov)
Call:
lm(formula = yield ~ block, data = npk)
Residuals:
    Min      1Q  Median      3Q     Max
-7.2250 -3.4937 -0.5375  2.1062 11.8750
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   54.025      2.721  19.855 1.09e-13 ***
block2         3.425      3.848   0.890   0.3852
block3         6.750      3.848   1.754   0.0964 .
block4        -3.900      3.848  -1.013   0.3243
block5        -3.500      3.848  -0.910   0.3751
block6         2.325      3.848   0.604   0.5532
---

Residual standard error: 5.442 on 18 degrees of
freedom
Multiple R-squared:  0.3917,    Adjusted R-squared:
```

Kategorická (faktoriální)
ANOVA



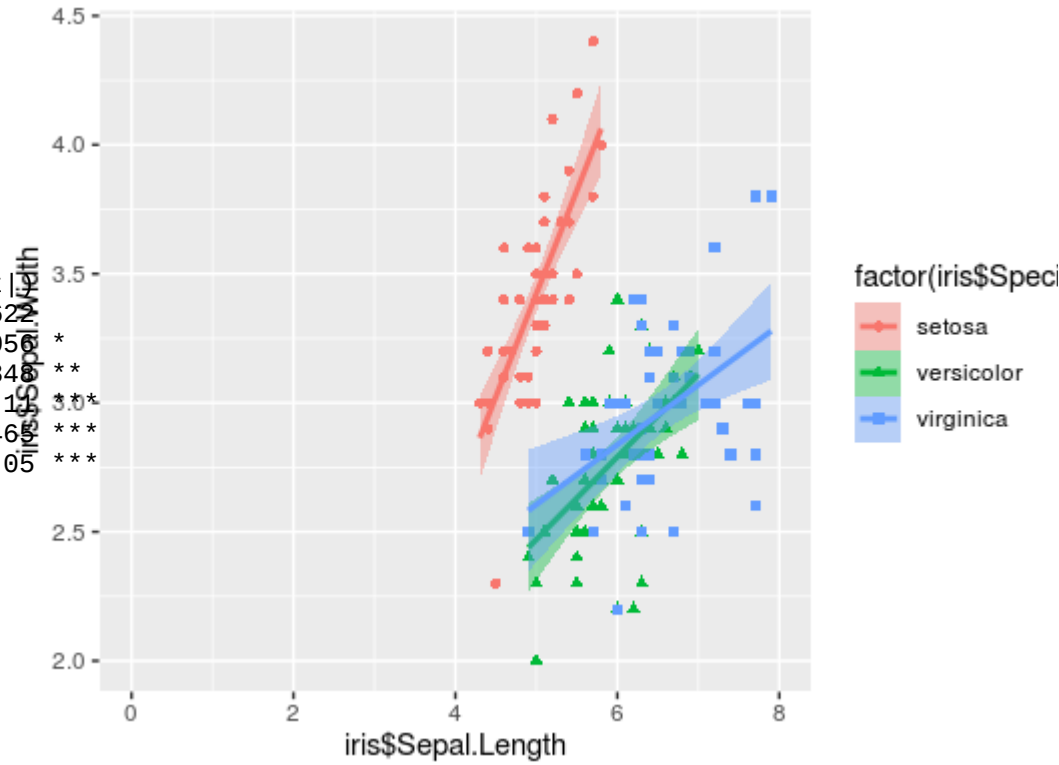
Opakování:

vysvětlující proměnná

Kontinuální + Kategorická (faktoriální)

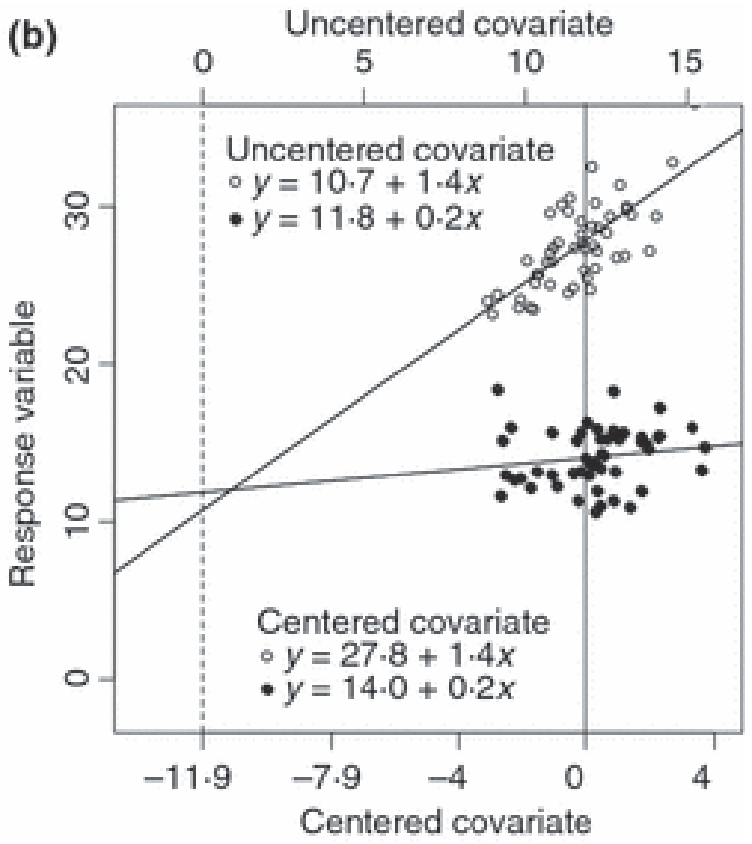
ANCOVA – (general linear model)

```
> data("iris")
> library(ggplot2)
> g <- ggplot(iris, aes(x=iris$Sepal.Length, y=iris$Sepal.Width, color=factor(iris$Species), shape=factor(iris$Species)))
> g1 <- g + geom_point()
> g2 <- g1 + geom_smooth(method=lm, aes(fill=factor(iris$Species))) + xlim(0, 8)
> #g2
>
> model1 <- lm(Sepal.Width ~ Species*Sepal.Length, data=iris)
> summary(model1)
Call:
lm(formula = Sepal.Width ~ Species * Sepal.Length, data = iris)
Residuals:
     Min       1Q   Median       3Q      Max
-0.72394 -0.16327 -0.00289  0.16457  0.60954
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -0.5694     0.5539  -1.028 0.305622
Speciesversicolor             1.4416     0.7130   2.022 0.045056 *
Speciesvirginica              2.0157     0.6861   2.938 0.003848 **
Sepal.Length                  0.7985     0.1104   7.235 2.55e-11 ***
Speciesversicolor:Sepal.Length  -0.4788     0.1337  -3.582 0.000465 ***
Speciesvirginica:Sepal.Length   -0.5666     0.1262  -4.490 1.45e-05 ***
---
Signif. codes:
Residual standard error: 0.2723 on 144 degrees of freedom
Multiple R-squared:  0.6227,    Adjusted R-squared:  0.6096
F-statistic: 47.53 on 5 and 144 DF,  p-value: < 2.2e-16
```



Opakování:

Pro lepší interpretovatelnost výsledků se doporučuje centrování (a standardizování) kontinuálních prediktorů (Schielzeth 2010, Methods in Ecology and Evolution)



Opakování:

Výběr modelu (step-wise model selection):

```
> model1 <- lm(Sepal.Width ~ Species*Sepal.Length, data=iris)
> model2 <- lm(Sepal.Width ~ Species+Sepal.Length, data=iris)
> anova(model1,model2)
Analysis of Variance Table

Model 1: Sepal.Width ~ Species * Sepal.Length
Model 2: Sepal.Width ~ Species + Sepal.Length
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    144 10.680
2    146 12.193 -2   -1.5132 10.201 7.19e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Full model → Postupné odstraňování predikrorů → Minimal Adequate Model (MAM)

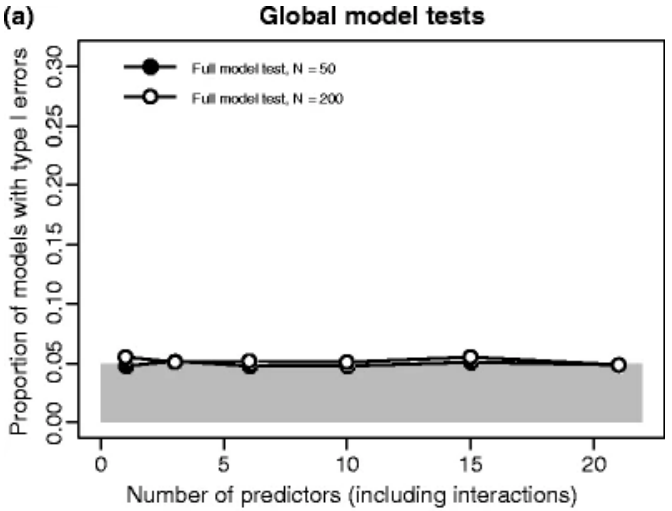
Opakování:

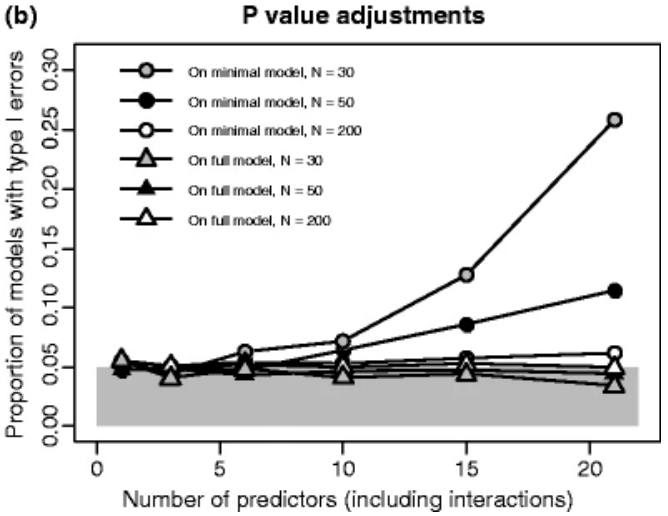
Výběr modelu (step-wise model selection):

Přes velkou oblíbenost, potenciálně poměrně problematický přístup, viz:

Forstmeier & Schielzeth (2010,Behav. Ecol. Socbiol)
Whittingham et al. (2006, J. Anim Ecol)

...ale pro účely tohoto kurzu akceptovatelný





Opakování:

Výběr modelu (alternativy):

- Korekce p hodnot při výběru na mnohonásobná testování,
- viz např. "Multiple-Stage False Discovery Rate procedure" (Benjamini and Gavrilov 2009)

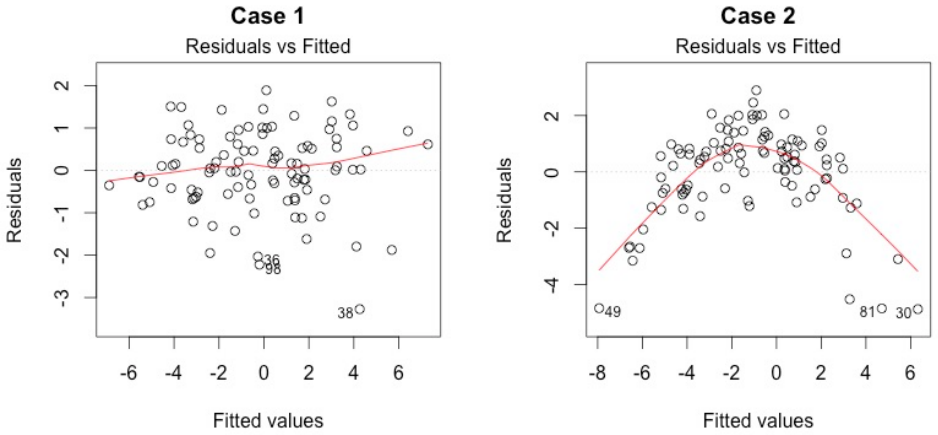
- Selekce na základě informačních kritérií (např. AIC, BIC)

- Použití nástrojů, které umožňují fitování modelu a výběr relevantních prediktorů v jednom kroku – např. **lasso regrese**

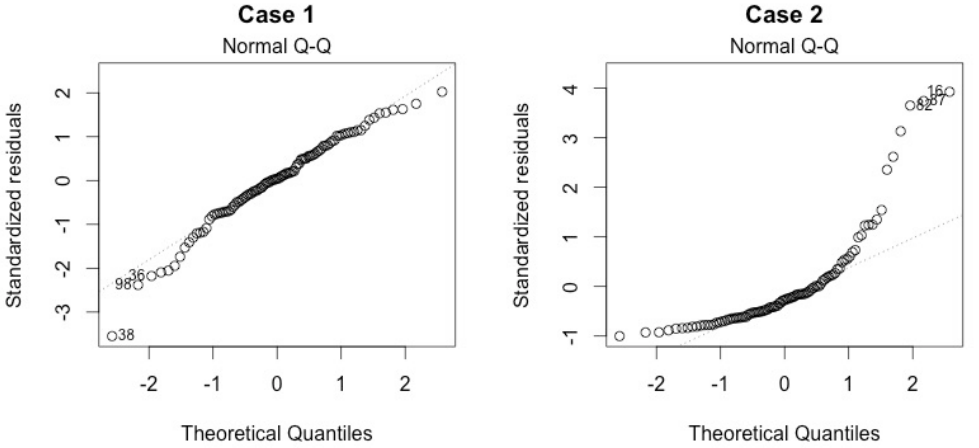
Opakování:

Diagnostika obecného lineárního modelu

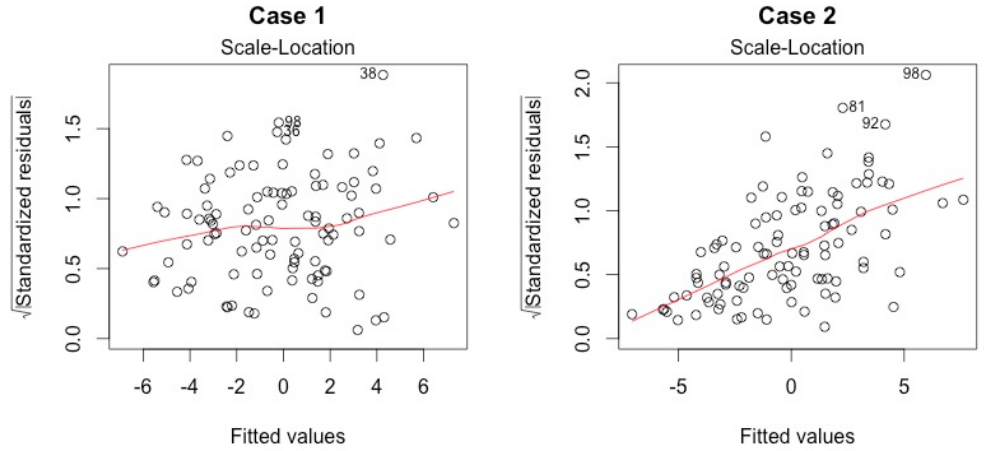
Residuály vs. fitované hodnoty:



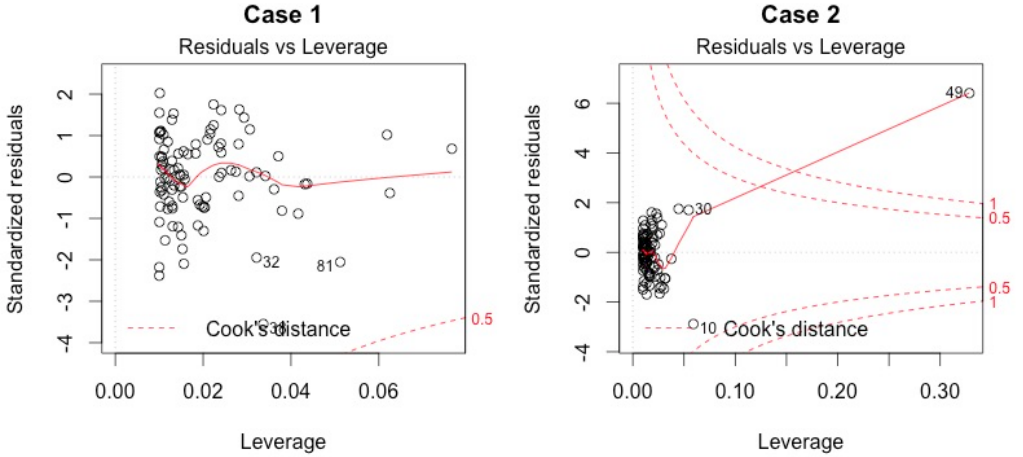
Q-Q plot: Normální rozdělení residuálů:



Scale-Location plot: homoscedasticita



Leverage



Opakování:

GLM: dělá to samé jako lm(), aov() jiná je akorát vysvětlovaná proměnná a výpočetní procedura
glm(y~x,family=XXXXX)

A) family=gaussian – Odpovídá obecnému lineárnímu modelu

B) family=gamma – Kontinuální vysvětlovaná proměnná, nenabývá negativních hodnot, disperze resuduálů roste s její hodnotou (velikost těla, sportovní výkony

C) family=poisson – počty, nezáporné diskrétní (počet parazitů v daném hostitely, počet dopravních nehod ve městech ….)

D) family=binomial – 0/1 přežil nepřežil, pohlaví apod,
- “poměr” - počet vyléčených/nevyléčných v jednotlivých skupinách

V případě binomial a poisson zkontrolovat jestli není overdispeze a pokud ano tak použít jiné rozdělění: C) quasipoisson, neg. binomial, zero inflated D) quasibinomial, batabinomial