MV011 Statistika I – cvičení 11

Lineární regresní model

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

jaro 2016

Pro řešení lineárního regresního modelu $Y_i = m(x_i) + \varepsilon_i$, $i = 1, \ldots, n$, v *R* slouží příkaz `lm` (*linear model*):

`model <-lm (formule, data = DatovaTabulka)`, příp.

`model <-lm (formule, data = DatovaTabulka, weights = VektorVah)`.

Pro tzv. *formuli* se používá speciální syntaxe, kde `Y` je název sloupce závisle proměnné, `x` je název sloupce nezávisle proměnné:

$m(x)$	*formule*
$\beta_0 + \beta_1 x$	`Y ~ x` nebo `Y ~ 1 + x`, člen β_0 je totiž vkládán implicitně
$\beta_1 x$	`Y ~ 0 + x`, odstranění členu β_0 nutno zapsat explicitně
$\beta_0 + \beta_1 x + \beta_2 x^2$	`Y ~ x + I(x^2)`
$\beta_2 x^2$	`Y ~ 0 + I(x^2)`
$\beta_1 \lvert x \rvert$	`Y ~ 0 + I(abs(x))`
$\beta_0 + \beta_1 e^x$	`Y ~ I(exp(x))`
$\beta_0 + \beta_1 \ln x$	`Y ~ I(log(x))`
$\beta_0 + \beta_1 \sqrt{x}$	`Y ~ I(sqrt(x))`

Detailní výsledky a další číselné charakteristiky získáme příkazem

`prehled <-summary (model)`,

příp. `prehled <-summary (model, correlation=TRUE)` pro výběrovou korelační matici parametrů.

$\widehat{\boldsymbol{\beta}}$	MNČ-odhady parametrů	`model$coefficients` `coef(model)`
$\left(\widehat{\beta}_j, \mathrm{SD}(\widehat{\beta}_j), T_j, p_j\right)$	odhady, směrodatné odchylky, testy významnosti, p-hodnoty	`prehled$coefficients` `coef(prehled)`
$\widehat{\boldsymbol{Y}}$	aproximované hodnoty	`model$fitted.values` `fitted.values(model)`
$\boldsymbol{r}$	rezidua	`model$residuals` `residuals(model)`
$n-k$	stupně volnosti modelu	`model$df.residual`
$\boldsymbol{X}$	matice plánu	`model.matrix(model)`
$\boldsymbol{w}$	váhy	`model$weights`
s	odhad sm. odchylky chyb ε_i	`prehled$sigma`
R^2	index determinace	`prehled$r.squared`
$\overline{R}^2$	korigovaný index determinace	`prehled$adj.r.squared`
$(F, k-1, n-k)$	celkový F-test	`prehled$fstatistic`
$(k, n-k, k)$	stupně volnosti	`prehled$df`
$\mathrm{R}(\widehat{\boldsymbol{\beta}})$	korelační matice odhadů $\widehat{\boldsymbol{\beta}}$	`prehled$correlation`

Úkoly v příkladech:

- MNČ-odhady parametrů $\widehat{\boldsymbol{\beta}}$ regresní funkce $m(x)$, sledujte i jejich významnost,
- zapište matematický tvar regresní funkce $m(x)$,
- reziduální součet čtverců S_e a odhad směrodatné odchylky s náhodných chyb,
- index determinace R^2, proveďte celkový F-test,
- vykreslete data a grafy regresních funkcí (`predict`), příp. s pásy spolehlivosti,
- vykreslete boxploty reziduí,
- modely porovnejte (mj. `anova`), zvolte z nich nejvhodnější.

Příklad 1

Datový soubor `KysMlecna.csv`*: zkoumejte závislost množství kyseliny mléčné u novorozence na množství stejné látky u matky-prvorodičky (v mg ve 100 ml krve) pomocí regresní přímky a paraboly.*

Příklad 2

Datový soubor `prodlouzeni.csv`*: zkoumejte závislost prodloužení měděné trubky v závislosti teplotním rozdílu* Δt *od referenční hodnoty* $t_0 = 20\,^\circ C$ *pomocí vhodné regresní přímky a paraboly. Dle fyzikálních zákonů by při* $\Delta t = 0$ *prodloužení mělo být nulové.*

Příklad 3

Datový soubor `spotreba2.csv`*: zkoumejte závislost spotřeby paliva motorového vozidla (v l/100 km) na rychlosti (v km/h) pomocí regresní přímky a paraboly.*

Příklad 4

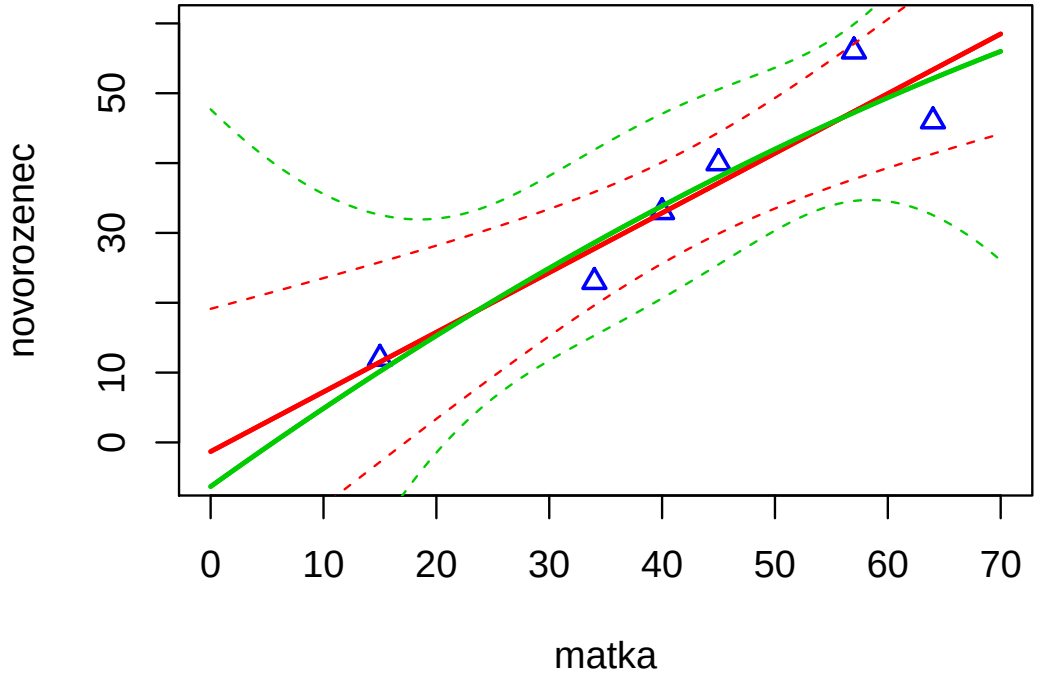
Datový soubor `CO2.csv`*: zkoumejte závislost koncentrace* CO_2 *(v ppm) v atmosféře v letech 1764–1995 pomocí několika polynomických regresních funkcí.*

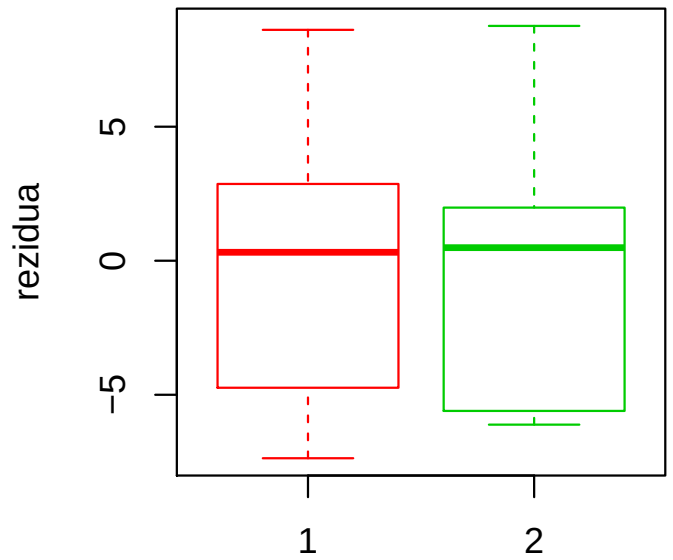
Příklad 5

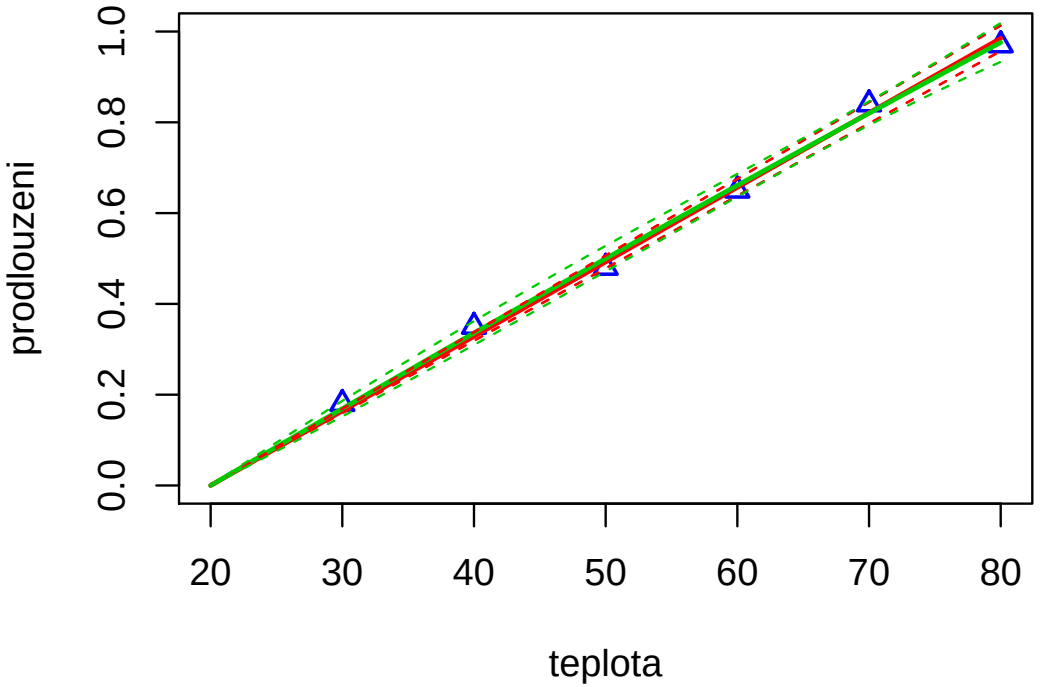
Datový soubor `EmiseUhliku.csv`*: zkoumejte závislost uhlíkových emisí (v milionech tun) v letech 1950–1995 pomocí několika polynomických regresních funkcí.*

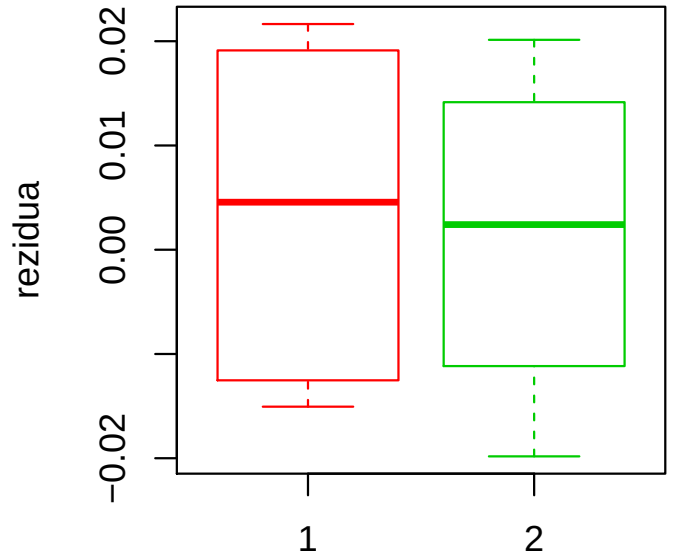
Příklad 6

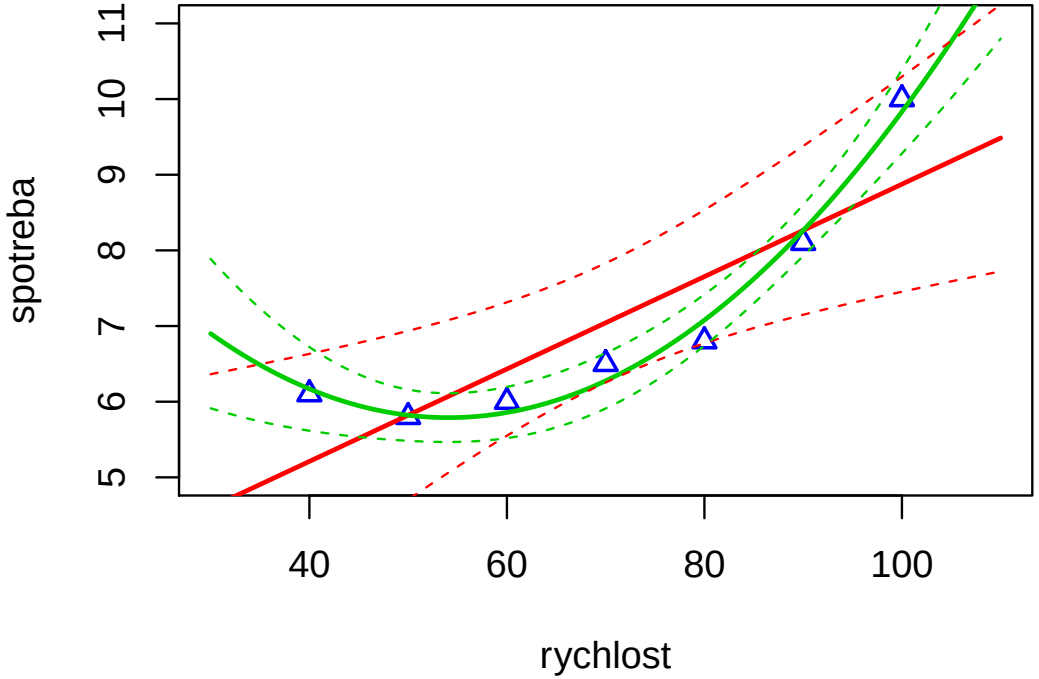
Datový soubor `teplota.csv`*: zkoumejte závislost průměrné teploty (ve* $^\circ C$*) v letech 1866–1996 pomocí několika polynomických regresních funkcí.*

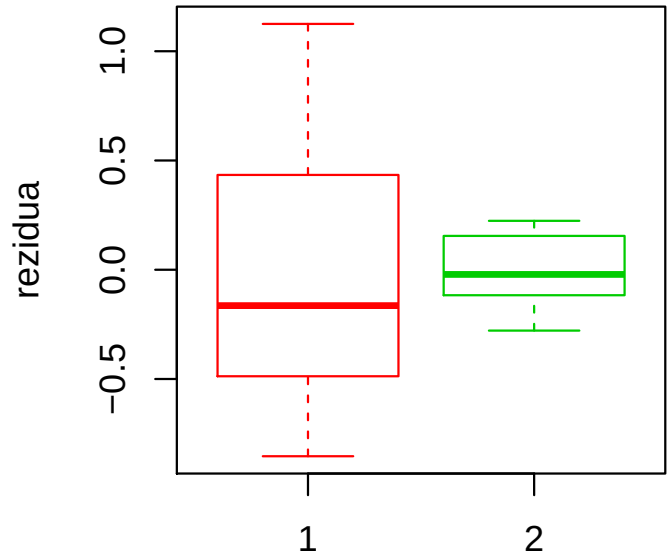

novorozenec
0
10
30
50
0
10
20
30
40
50
60
70
matka

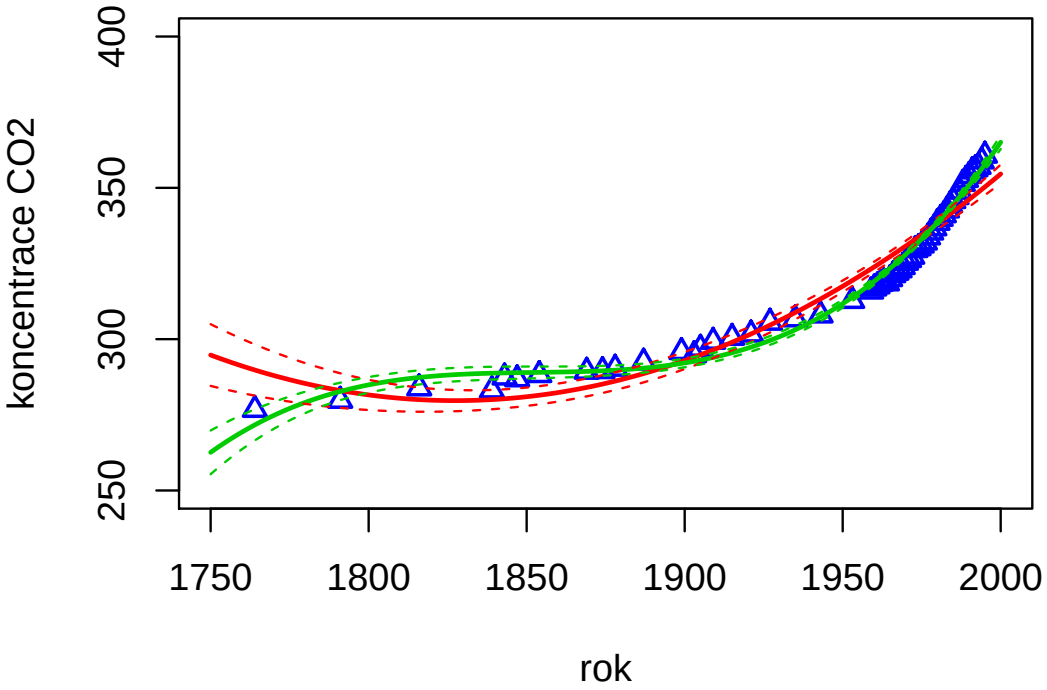

rezidua
−5
0
5
1
2

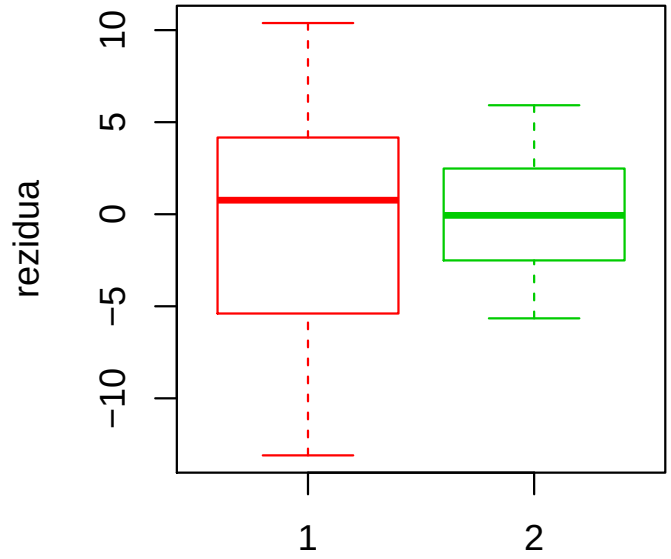

prodlouzeni
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
80
teplota

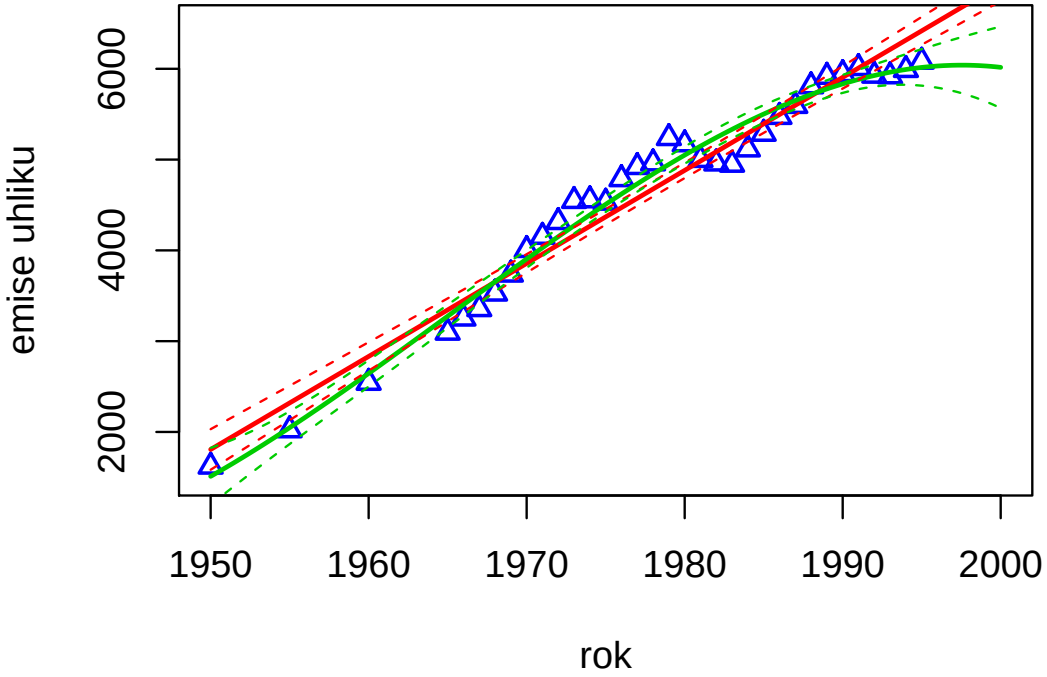

rezidua
−0.02
0.00
0.01
0.02
1
2

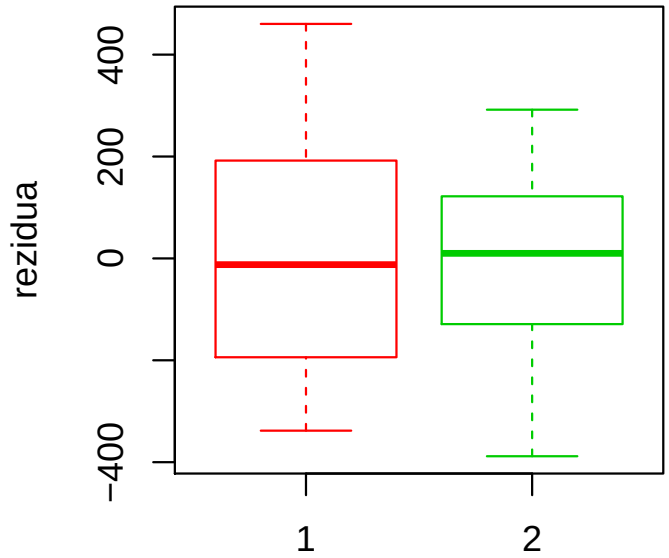

spotreba
5
6
7
8
9
10
11
40
60
80
100
rychlost

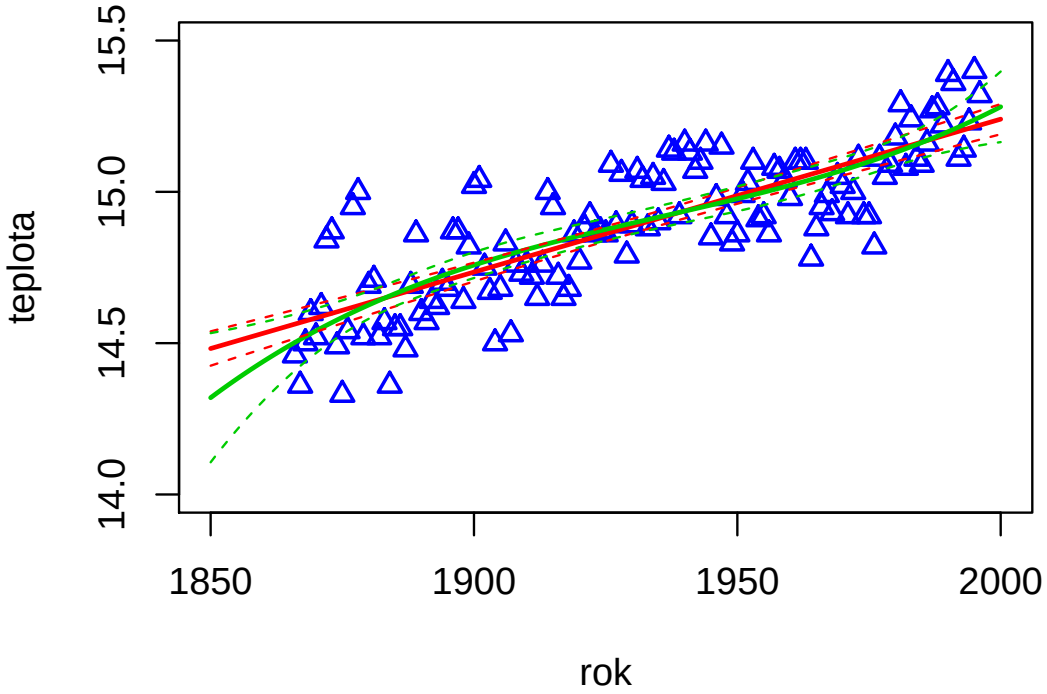

rezidua
−0.5
0.0
0.5
1.0
1
2

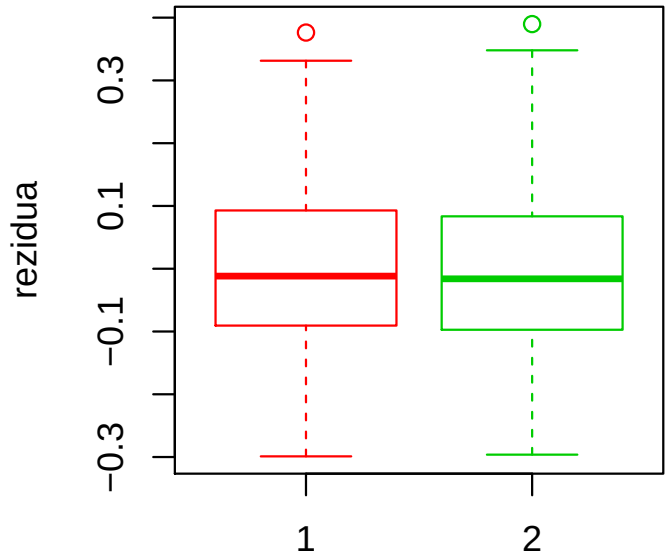

koncentrace CO2
250
300
350
400
1750
1800
1850
1900
1950
2000
rok


rezidua
−10
−5
0
5
10
1
2


emise uhliku
2000
4000
6000
1950
1960
1970
1980
1990
2000
rok


rezidua
−400
0
200
400
1
2


teplota
14.0
14.5
15.0
15.5
1850
1900
1950
2000
rok


rezidua
−0.3
−0.1
0.1
0.3
1
2