M5201 – 2. CVIČENÍ: *Regresní analýza v R*

1 Úvod do regresní analýzy

Regresní analýza je jednou z nejčastěji používaných statistických technik. Zabýváme se při ní vztahy mezi několika veličinami. Snažíme se zjistit, jak hodnoty jedné veličiny (*vysvětlované proměnné, odezvy*) závisí na hodnotách ostatních veličin (*vysvětlujících veličin, regresorů, kovariát*).

Vycházíme z předpokladu, že vysvětlovaná veličina je funkcí vysvětlujících proměnných a náhodné složky, což je náhodná veličina s nulovou střední hodnotou, která reprezentuje náhodné odchylky, chyby v měření apod.

$$Y = f(x_1, x_2, \ldots, x_m, \varepsilon),$$

kde	Y	je vysvětlovaná proměnná
	$x_1, \ldots, x_m$	jsou vysvětlující proměnné (kovariáty, regresory)
	ε	je náhodná složka

Funkce f patří do předem zvolené třídy funkcí a závisí na jednom či více parametrech. Ty jsou neznámé a odhadují se na základě zjištěných dat.

Obvykle máme k dispozici n pozorování vysvětlované proměnné

$$Y_1, Y_2, \ldots, Y_n$$

a jim odpovídající hodnoty vysvětlujících proměnných

$$x_{11}, x_{12}, \ldots, x_{1n}, x_{21}, x_{22}, \ldots, x_{2n}, \ldots, x_{m1}, \ldots, x_{mn}$$

a pomocí regresního modelu můžeme popsat jednotlivá pozorování

$$Y_i = f(x_{i1}, x_{i2}, \ldots, x_{im}, \varepsilon_i), \qquad i = 1, \ldots, n$$

Nejjednodušším případem regresního modelu je klasický lineární regresní model. Je charakteristický tím, že vysvětlovanou proměnnou se v něm snažíme popsat pomocí lineární kombinace známých funkcí vysvětlujících proměnných s neznámými regresními koeficienty. Pomocí rovnice můžeme lineární regresní model zapsat takto

$$Y_i = \beta_0 + \beta_1 f_1(x_{i1}, x_{i2}, \ldots, x_{im}) + \beta_2 f_2(x_{i1}, x_{i2}, \ldots, x_{im}) + \cdots + \beta_k f_k(x_{i1}, x_{i2}, \ldots, x_{im}) + \varepsilon_i,$$

kde pro $i = 1, \ldots, n$

Y_i	jsou vysvětlované proměnné
$x_{i1}, \ldots, x_{im}$	jsou vysvětlující proměnné (kovariáty, regresory)
$f_1, f_2, \ldots, f_k$	jsou známé (předem zvolené) funkce vysvětlujících proměnných
$\beta_0, \beta_1, \ldots, \beta_k$	jsou neznámé regresní koeficienty
ε_i	jsou nekorelované náhodné veličiny s nulovou střední hodnotou a konstatním rozptylem

V lineárním regresním modelu tedy vysvětlovanou proměnnou modelujeme jako součet *systematické složky* (což je funkce, která je lineární v neznámých regresních koeficientech) a *náhodné složky.*

V lineárním regresním modelu se dále předpokládá, že pro různé realizace Y_i, $i = 1, \ldots, n$ vysvětlované proměnné jsou odpovídající náhodné veličiny ε_i navzájem nezávislé a mají stejné rozdělení s nulovou střední hodnotou a konstatním rozptylem, což značíme $\varepsilon_i \sim IID(0, \sigma^2)$. Pro testování hypotéz je třeba přidat předpoklad, že rozdělení chybové složky je normální.

Pro odhad neznámých parametrů regresních modelů existuje více metod. Pro lineární regresní modely se obvykle používá *metoda nejmenších čtverců*. Odhady parametrů se obvykle označují stříškou, např. $\widehat{\beta}_0, \widehat{\beta}_1, \ldots$. Hodnoty vysvětlované proměnné $\widehat{Y}_i$, které získáme po dosazení odhadnutých parametrů do regresní rovnice

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 f_1(x_{i1}, x_{i2}, \ldots, x_{im}) + \widehat{\beta}_2 f_2(x_{i1}, x_{i2}, \ldots, x_{im}) + \cdots + \widehat{\beta}_k f_k(x_{i1}, x_{i2}, \ldots, x_{im})$$

se nazývají *vyrovnané hodnoty*. Rozdíly mezi skutečnými a vyrovnanými hodnotami vysvětlované proměnné $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$ se nazývají *rezidua*.

Veličiny v modelu mohou být jak spojité, tak kategorické (sem patří nominální, ordinální a diskrétní veličiny). Kategorické proměnné se také nazývají faktory a jejich hodnoty se nazývají úrovně.

Poznámka

Při maticovém popisu regresního modelu obvykle provedeme určité přeznačení

$$Y_i = \overbrace{\beta_0 + \beta_1 \underbrace{f_1(z_{i1}, z_{i2}, \ldots, z_{ip})}_{\text{označíme } x_{i1}} + \beta_2 \underbrace{f_2(z_{i1}, z_{i2}, \ldots, z_{ip})}_{\text{označíme } x_{i2}} + \cdots + \beta_m \underbrace{f_m(z_{i1}, z_{i2}, \ldots, z_{im})}_{\text{označíme } x_{im}}}^{\text{systematická složka}} + \varepsilon_i,$$

takže pracujeme s modelem

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_m x_{1k} + \varepsilon_1 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_m x_{nm} + \varepsilon_n \end{aligned} \quad \begin{array}{c}\text{tj.}\\ \text{maticově}\end{array} \quad \underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix}}_{\mathbf{X}\text{(matice plánu)}} \underbrace{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_m \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

Uvažujme **klasický regresní model** plné hodnosti, kde chyba má normální rozdělení:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \wedge \quad h(\mathbf{X}) = h(\mathbf{X}'\mathbf{X}) = m + 1 \quad \wedge \quad n > m + 1 \quad \wedge \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

Odhad neznámých parametrů $\boldsymbol{\beta}$ provedený *metodou nejmenších čtverců* je řešením normálních rovnic

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$$

a platí

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Označme

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} = \underbrace{\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'}_{\mathbf{H}}\,\mathbf{Y} = \mathbf{H}\mathbf{Y} \\ \hat{\boldsymbol{\varepsilon}} &= \mathbf{Y} - \hat{\mathbf{Y}} = (\underbrace{\mathbf{I} - \mathbf{H}}_{\mathbf{M}})\mathbf{Y} = \mathbf{M}\mathbf{Y} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) = \underbrace{\mathbf{M}\mathbf{X}}_{=\mathbf{0}}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \\ s^2 &= \tfrac{S_e}{n-p-1} = \tfrac{1}{n-m-1}(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \tfrac{1}{n-m-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \tfrac{1}{n-m-1}\hat{\mathbf{Y}}'(\mathbf{I} - \mathbf{H})\hat{\mathbf{Y}} = \tfrac{1}{n-m-1}\boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \end{aligned}$$

Pak připomeňme následující vlastnosti klasického regresního modelu (plné hodnosti)

- $E\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ (nestrannost odhadu)

- $Es^2 = \frac{E(S_e)}{n-m-1} = \sigma^2$, tj. s^2 je nestranným odhadem rozptylu

- $D\widehat{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

- $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$

- $\widehat{\boldsymbol{\varepsilon}} \sim N_n(\mathbf{O}, \sigma^2(\mathbf{I} - \mathbf{H}))$

- $\widehat{\boldsymbol{\beta}} \sim N_{m+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

- $\frac{S_e}{\sigma^2} \sim \chi^2(n-m-1)$

- $\widehat{\boldsymbol{\beta}}$ a s^2 jsou stochasticky nezávislé

- $T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 v_{jj}}} \sim t(n-m-1)$, kde $(\mathbf{X}'\mathbf{X})^{-1} = (v_{ij})_{i,j=0,\dots,m}$

- $F = \frac{1}{qs^2}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)'\mathbf{W}^{-1}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \sim F(q, n-m-1)$,

  kde $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \mathbf{V} & \mathbf{U} \\ \mathbf{U} & \mathbf{W} \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$ $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$ a $h(\mathbf{W}) = q$

- $T = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}} - \mathbf{c}'\boldsymbol{\beta}}{\sqrt{s^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim t(n-m-1)$, kde $\mathbf{c} = (c_0, c_1, \dots, c_m)'$ a $E(\mathbf{c}'\widehat{\boldsymbol{\beta}}) = \mathbf{c}'\boldsymbol{\beta}$

- $\begin{array}{l} Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \\ \hat{Y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i) \end{array} \Rightarrow Y_i - \hat{Y}_i \sim N(0, \sigma^2(1 + \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i))$

  kde $\mathbf{x}_i' = (x_{i0}, \dots, x_{im})$ *je* i-tý řádek matice plánu $\mathbf{X}$

Na základě předchozích vlastností lze odvodit následující intervaly spolehlivosti

Intervaly spolehlivosti	
pro parametry β_j $j=0,\dots,p$	$\left(\beta_j - t_{1-\frac{\alpha}{2}}(n-p-1)s\sqrt{v_{jj}},\ \beta_j + t_{1-\frac{\alpha}{2}}(n-p-1)s\sqrt{v_{jj}}\right)$
pro střední hodnotu predikce $E\hat{Y}_i = E\mathbf{x}_i'\widehat{\boldsymbol{\beta}} = \mathbf{x}_i'\boldsymbol{\beta}$	$\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}} - t_{1-\frac{\alpha}{2}}(n-m-1)s\sqrt{\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i},\ \mathbf{x}_i'\widehat{\boldsymbol{\beta}} + t_{1-\frac{\alpha}{2}}(n-m-1)s\sqrt{\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}\right)$
pro predikci $\hat{Y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ $i=1,\dots,n$	$\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}} - t_{1-\frac{\alpha}{2}}(n-m-1)s\sqrt{1+\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}, \mathbf{x}_i'\widehat{\boldsymbol{\beta}} + t_{1-\frac{\alpha}{2}}(n-m-1)s\sqrt{1+\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}\right)$

kde $t_{1-\frac{\alpha}{2}}(n-m-1)$ je $1-\frac{\alpha}{2}$ kvantil Studentova rozdělení o $n-m-1$ stupních volnosti

2 Použití funkce `lm()`

K odhadu parametrů lineárního regresního modelu je v systému R k dispozici funkce `lm`. Tato funkce má následující argumenty

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Nejdůležitější je argument `formula`, kterým se zadává, jakou podobu má mít rovnice popisující model. Modelový vztah mezi proměnnými se obecně zadává ve tvaru

vysvětlovaná proměnná $\sim$ *systematická složka*

Zatímco na levé straně můžeme normálně používat obvyklé matematické funkce jako logaritmus nebo matematické operátory $+, -$ apod., na pravé straně mají matematické operátory poněkud odlišný význam. Konkrétně:

`+`	přidat vysvětlující proměnnou
`-`	odstranit vysvětlující proměnnou
`:`	interakce mezi veličinami
`*`	všechny komponenty
`1`	absolutní člen
`-1`	odebrat absolutní člen

Pokud potřebujeme při specifikaci modelu použít matematické operátory v obvyklém smyslu, musíme výraz s těmito operátory napsat jako argument funkce `I()`. Rozdíl můžeme ukázat na příkladu.

`y` $\sim$ `x1` $+$ `x2` definuje model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$

zatímco

`y` $\sim$ `I(x1` $+$ `x2)` definuje model $Y_i = \beta_0 + \beta_1(x_{1i} + x_{2i}) + \varepsilon_i.$

Konkrétní příklady zadávání modelů v systému R jsou ukázány v následující tabulce.

Formální zápis	***Syntaxe v R***	***Popis***
$Y_i = \mu + \varepsilon_i$	`y ~ 1`	Model pouze s absolutním členem
$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	`y ~ x`	Lineární model se spojitou vysvětlující proměnnou x (regresní přímka)
$Y_i = \beta_1 x_i + \varepsilon_i$	`y ~ x - 1`	Lineární model se spojitou vysvětlující x bez absolutního členu
$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$	`log(y) ~ x`	Lineární model se spojitou vysvětlující x a s logaritmicky transformovanou vysvětlovanou proměnnou
$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$	`y ~ x + I(x^2)`, `y ~ poly(x, 2)`	Kvadratický model se spojitou vysvětlující proměnnou x (regresní parabola)
$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$	`y ~ x1+x2`	Model se dvěma spojitými vysvětlujícími proměnnými x_1 a x_2, v každé z nich je lineární (vícenásobná regrese)
$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$	`y ~ x1*x2`, `y ~ x1+x2+I(x1*x2)`	Model se dvěma spojitými vysvětlujícími proměnnými x_1 a x_2 s křížovým členem
$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$	`y ~ A`	Model s jedním faktorem (jednofaktorová analýza rozptylu, ANOVA)
$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$	`y ~ A + B`	Model se dvěma faktory bez interakce (dvoufaktorová analýza rozptylu)
$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$	`y ~ A + B + A:B`, `y ~ A * B`	Model se dvěma faktory s interakcí (dvoufaktorová analýza rozptylu)
$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijkl}$	`y ~ A + B + C`	Model se třemi faktory bez interakcí (třífaktorová analýza rozptylu)
$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}$	`y ~ A+B+C+A:B+A:C+B:C+A:B:C` `y ~ A*B*C`	Model se třemi faktory s interakcemi (třífaktorová analýza rozptylu)
$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijkl}$	`y ~ A+B+C+A:B+A:C+B:C` `y ~ (A+B+C)^2`	Model se třemi faktory obsahující kromě hlavních efektů maximálně dvojné interakce (třífaktorová analýza rozptylu)
$Y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{ij}$	`y ~ x+A`	Model se spojitou proměnnou x a faktorem A bez interakce interakce (ANCOVA)
$Y_{ij} = \mu + \alpha_j + \beta x_{ij} + \delta_j x_{ij} + \varepsilon_{ij}$	`y ~ x*A`	Model se spojitou proměnnou x a faktorem A s interakcí (ANCOVA)

Tabulka 1: Přehled zápisu základních modelů (spojité veličiny značíme písmeny z konce abecedy: $x, Y, \ldots$, faktory písmeny ze začátku abecedy: $A, B, \ldots$)

3 Příklady

Příklad 1: Regresní přímka

Máme k dispozici údaje o počtu australských domácností připojených v letech 1998–2008 na internet:

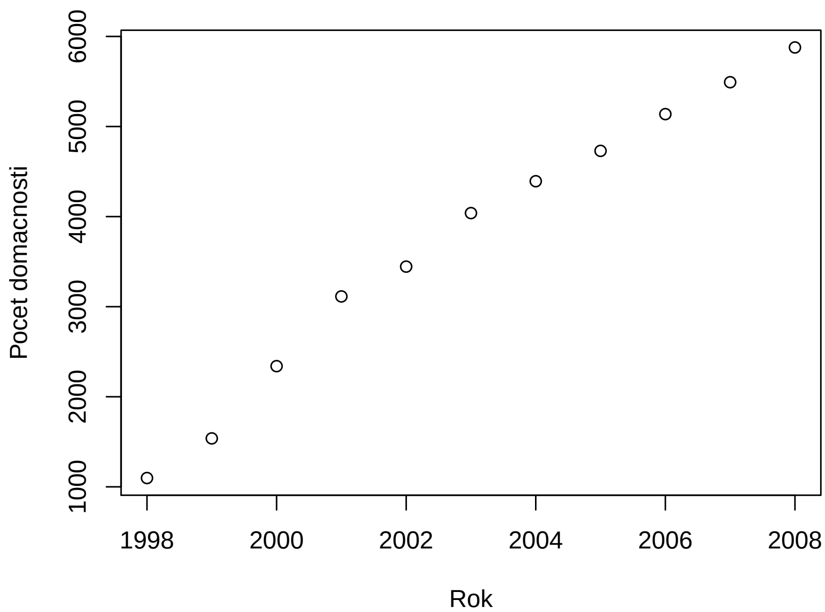
Rok	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Počet domácností	1098	1538	2340	3114	3445	4039	4393	4730	5138	5492	5878

Data nejprve načteme, a to tak, že údaje o počtech domácností zkopírujeme do schránky a použijeme funkci `scan()`.

```
> domacnosti <- scan("clipboard", sep = " ")
> time <- 1998:2008
```

Načtená data vykreslíme do grafu.

```
> TxtX <- "Rok"
> TxtY <- "Pocet domacnosti"
> plot(time, domacnosti, type = "p", xlab = TxtX, ylab = TxtY)
```



Obrázek 1: Počet domácností s připojením na internet v letech 1998–2008

Budeme uvažovat regresní model, ve kterém počet domácností s připojením na internet závisí lineárně na čase. Model můžeme popsat regresní rovnicí

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \dots, 11$$

kde

Y_i	je počet domácností s připojením na internet v i-tém roce
x_i	je rok odpovídající i-tému pozorování
β_0, β_1	jsou neznámé regresní koeficienty, které budeme odhadovat
ε_i	je náhodná odchylka příslušející i-tému pozorování

Parametry tohoto modelu odhadneme v systému R pomocí funkce `lm()` a k zobrazení výsledků použijeme funkci `summary()`.

```
> model1 <- lm(domacnosti ~ time)
> summary(model1)


Call:
lm(formula = domacnosti ~ time)

Residuals:
    Min      1Q  Median      3Q     Max 
-306.45 -200.05   20.18  173.09  318.82 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -948407.45   45085.31  -21.04 5.81e-09 ***
time            475.36      22.51   21.12 5.61e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 236.1 on 9 degrees of freedom
Multiple R-squared: 0.9802,     Adjusted R-squared: 0.978 
F-statistic:   446 on 1 and 9 DF,  p-value: 5.615e-09
```

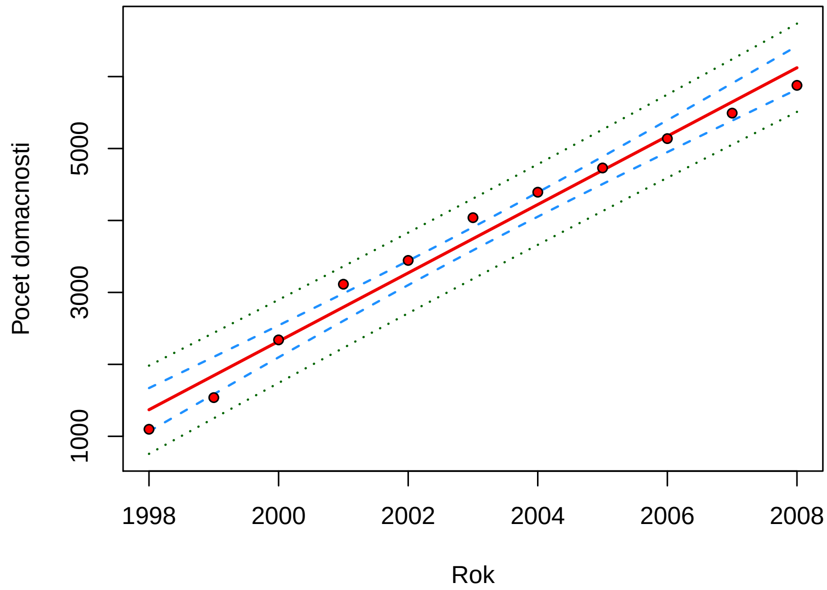
Ve výpisů výsledků odhadu modelu vidíme nejprve zadaný tvar modelu. Dále jsou zde uvedeny informace o reziduích (minimální a maximální reziduum, horní a dolní kvartil a medián). V další tabulce jsou shrnuty odhadnuté parametry

$$\beta_0 = -948407.45 \; (\texttt{Intercept}) \qquad \beta_1 = 475.36 \; (\texttt{time})$$

a v dalších sloupcích jsou směrodatné odchylky odhadnutých parametrů a testové statistiky pro testování některých hypotéz týkajících se parametrů modelu. V závěru jsou uvedeny souhrnné statistiky týkající se celého modelu.

Na závěr zobrazíme data proložená odhadnutou regresní přímkou. Do grafu navíc dokreslíme interval spolehlivosti kolem střední hodnoty a predikční interval spolehlivosti (který je širší a měl by obsahovat cca 95% dat).

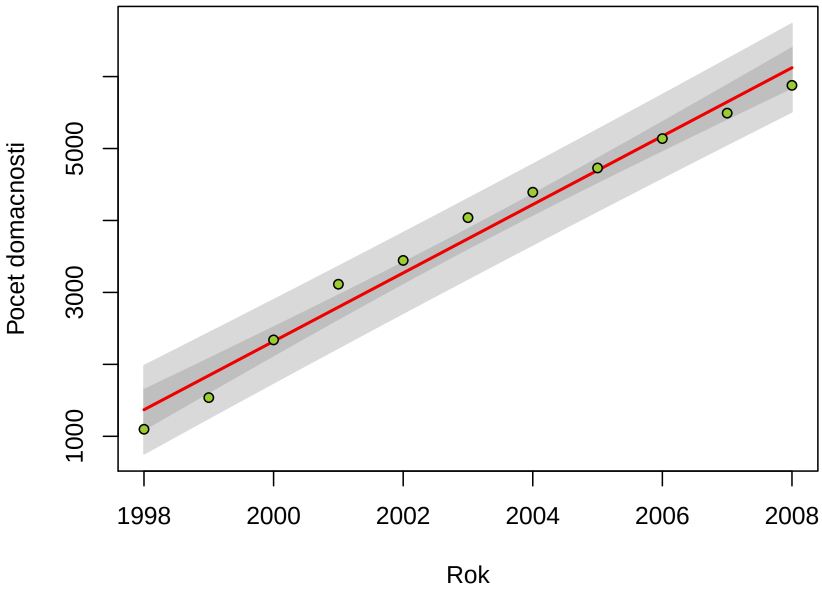
```
> n <- length(time)
> gridt <- seq(time[1], time[n], length.out = 500)
> ci.conf <- predict(model1, newdata = data.frame(time = gridt), interval = "confidence")
> ci.pred <- predict(model1, newdata = data.frame(time = gridt), interval = "prediction")
> yrange <- range(c(domacnosti, ci.pred[, 2], ci.pred[, 3]))
> plot(time[c(1, n)], yrange, type = "n", xlab = TxtX, ylab = TxtY)
> matlines(gridt, ci.conf[, 1], lty = 1, lwd = 2, col = "red2")
> matlines(gridt, ci.conf[, 2:3], lty = 2, lwd = 1.5, col = "dodgerblue")
> matlines(gridt, ci.pred[, 2:3], lty = 3, lwd = 1.5, col = "darkgreen")
> points(time, domacnosti, pch = 21, cex = 0.85, bg = "red1")
```



Obrázek 2: Grafické vyjádření regresní přímky pro data *Počet domácností s připojením na internet v letech 1998–2008* spolu s intervaly spolehlivosti kolem střední hodnoty a s predikčním intervalem

Pro zajímavost totéž provedeme v trochu jiné grafické úpravě.

```
> plot(time[c(1, n)], yrange, type = "n", xlab = TxtX, ylab = TxtY)
> xx <- c(gridt, rev(gridt))
> yy <- c(ci.conf[, 2], rev(ci.conf[, 3]))
> polygon(xx, yy, col = "gray75", border = "gray75")
> yy <- c(ci.conf[, 2], rev(ci.pred[, 2]))
> polygon(xx, yy, col = "gray85", border = "gray85")
> yy <- c(ci.conf[, 3], rev(ci.pred[, 3]))
> polygon(xx, yy, col = "gray85", border = "gray85")
> matlines(gridt, ci.conf[, 1], lty = 1, lwd = 2, col = "red2")
> points(time, domacnosti, pch = 21, cex = 0.85, bg = "yellowgreen")
```



Obrázek 3: Grafické vyjádření regresní přímky pro data *Počet domácností s připojením na internet v letech 1998–2008* spolu s intervaly spolehlivosti kolem střední hodnoty a s predikčním intervalem – s využitím ploch

Příklad 2: Polynomická regrese

Máme k dispozici průměrné roční průtoky vody v řece Nigeru v Coulicouro (Mali). Údaje se vztahují k období 1907 až 1957 a jsou uvedena v jednotkách cfs.10^{-3}.

Data jsou uložena v souboru `DataNiger.dat`. První řádek obsahuje popis časové řady, pak následuje volný řádek, teprve potom dva sloupce dat: rok a průměrný roční průtok.

Nejprve načteme nadpis.

```
> fileDat <- paste(data.library, "DataNiger.dat", sep = "")
> (TXT <- paste(scan(fileDat, what = "", nlines = 1), collapse = " "))
```

```
[1] "Prumerne rocni prutoky vody v rece Nigeru v Coulicoure (Mali) v letech 1907 az 1957"
```

Pak načteme vlastní data do datového rámce za pomoci funkce `read.table()` s argumentem `skip=2`, který způsobí vynechání prvních dvou řádků. Ve vzniklém datovém rámci pojmenujeme proměnné, vypíšeme jeho strukturu a prvních šest řádků. Nakonec data vykreslíme.

```
> dataNiger <- read.table(fileDat, header = F, skip = 2)
> names(dataNiger) <- c("Rok", "Prutok")
> str(dataNiger)
```

```
'data.frame':	51 obs. of  2 variables:
 $ Rok   : int  1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 ...
 $ Prutok: num  39.8 42.9 69.1 43.7 56.7 ...
```

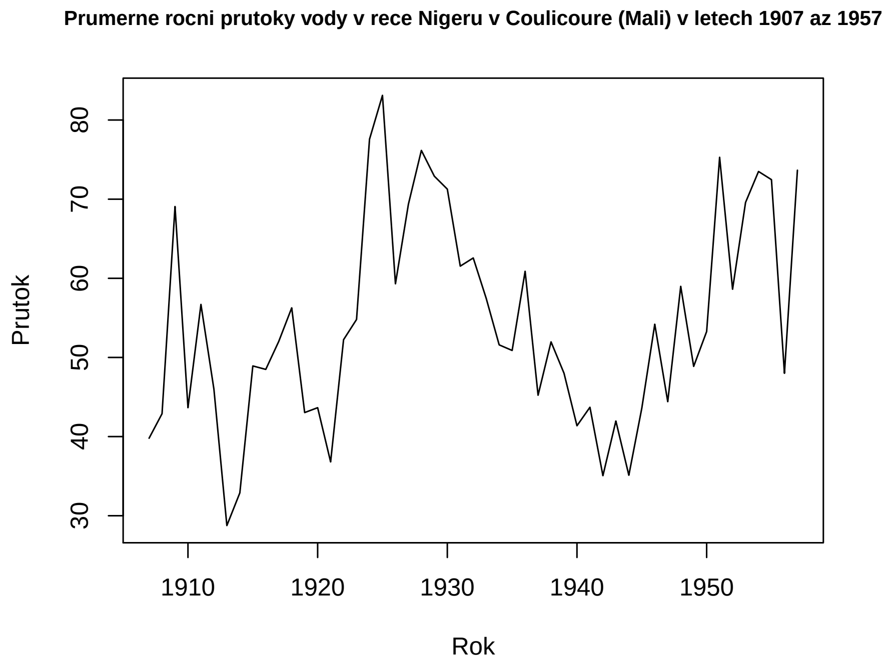
```
> head(dataNiger)
```

```
   Rok Prutok
1 1907 39.793
2 1908 42.892
3 1909 69.070
4 1910 43.653
5 1911 56.700
6 1912 45.990
```

Příkazem `attach()` zpřístupníme proměnné v datovém rámci `dataNiger` a data vykreslíme.

```
> attach(dataNiger)
```

```
> plot(Rok, Prutok, main = TXT, cex.main = 0.85, type = "l")
```



Obrázek 4: Průměrně roční průtoky v řece Niger v Coulicoure (Mali).

Z grafu je patrné, že závislost průtoku na čase bude složitější a proložit grafem regresní přímku by mohlo být příliš zjednodušující. Proto zkusíme daty proložit polynom vyššího stupně. Nejprve zkusíme polynom třetího stupně, kdy budeme předpokládat, že jednotlivá pozorování můžeme popsat vztahem

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i,$$

kde Y_i označuje průtok vody a x_i označuje rok odpovídající danému údaji.

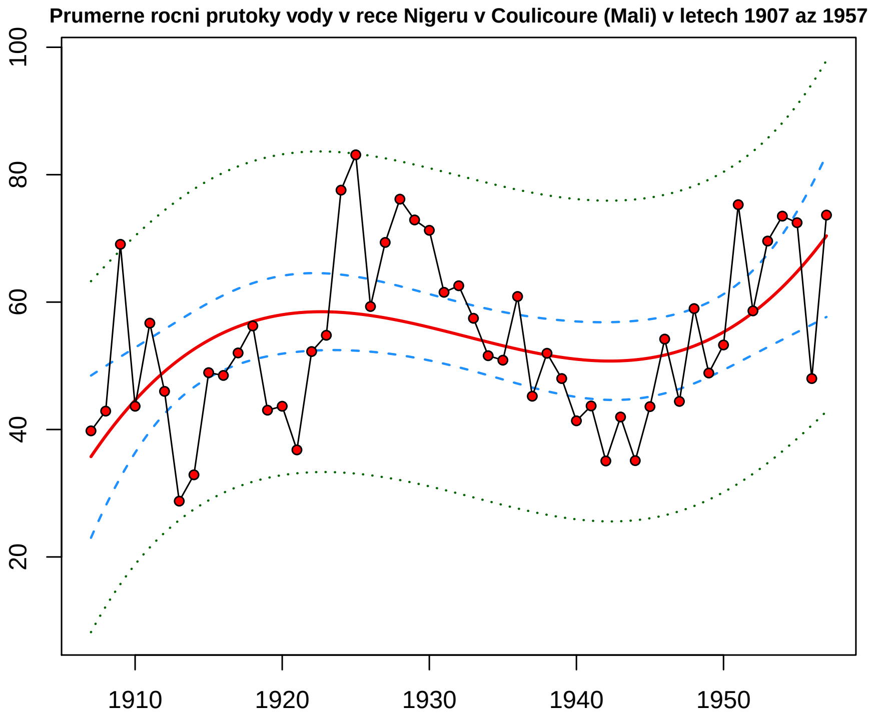
Neznámé parametry tohoto modelu nyní odhadneme.

```
> model3 <- lm(Prutok ~ Rok + I(Rok^2) + I(Rok^3))
> summary(model3)
```

```
Call:
lm(formula = Prutok ~ Rok + I(Rok^2) + I(Rok^3))

Residuals:
     Min       1Q   Median       3Q      Max
-22.2179  -7.0170  -0.9496   7.6885  27.1292

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.483e+07  4.903e+06  -3.024  0.00403 **
Rok          2.302e+04  7.614e+03   3.023  0.00404 **
I(Rok^2)    -1.191e+01  3.941e+00  -3.022  0.00405 **
I(Rok^3)     2.055e-03  6.800e-04   3.022  0.00406 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.14 on 47 degrees of freedom
Multiple R-squared: 0.2108,	Adjusted R-squared: 0.1604
F-statistic: 4.183 on 3 and 47 DF,  p-value: 0.0105
```

Výsledky modelu s polynomem řádu tři zakreslíme do grafu. Odhadnutou regresní křivku zobrazíme do grafu tak, že nejprve vytvoříme vektor `gridt` obsahující 500 hodnot v rozmezí 1907 a 1957. Pak použijeme funkci `predict()`, pomocí níž pro všechny prvky `gridt` spočítáme na základě odhadnutých parametrů modelu vyrovnané hodnoty jednak odhadnuté hodnoty, kromě toho také intervaly spolehlivosti kolem střední hodnoty a predikční intervaly.

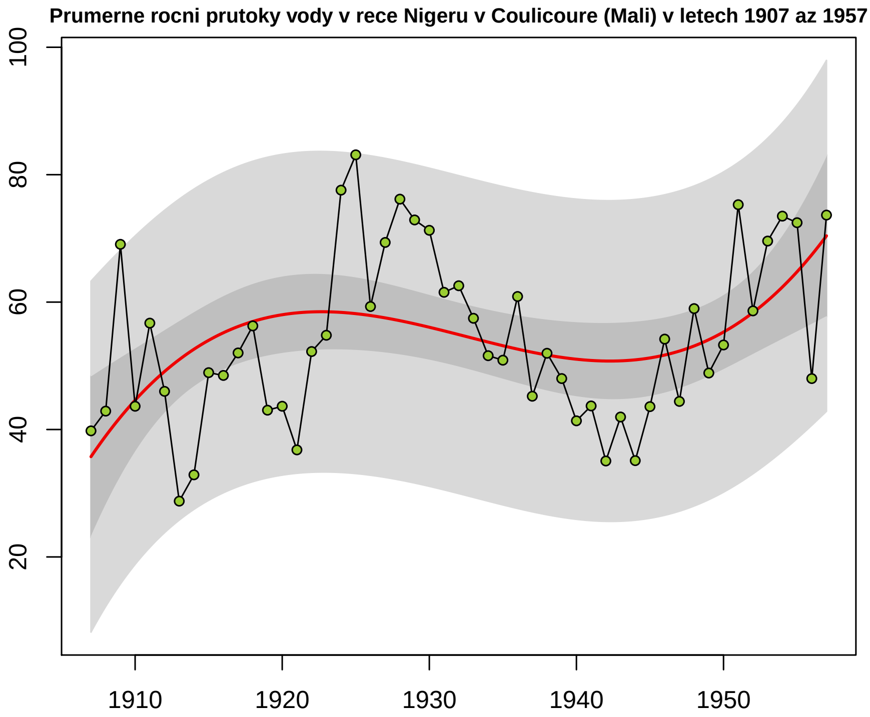
```
> n <- length(Rok)
> gridt <- seq(Rok[1], Rok[n], length.out = 500)
> ci.conf <- predict(model3, newdata = data.frame(Rok = gridt), interval = "confidence")
> ci.pred <- predict(model3, newdata = data.frame(Rok = gridt), interval = "prediction")
> yrange <- range(c(Prutok, ci.pred[, 2], ci.pred[, 3]))
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(Rok[c(1, n)], yrange, type = "n", main = TXT, cex.main = 0.85)
> matlines(gridt, ci.conf[, 1], lty = 1, lwd = 2, col = "red2")
> matlines(gridt, ci.conf[, 2:3], lty = 2, lwd = 1.5, col = "dodgerblue")
> matlines(gridt, ci.pred[, 2:3], lty = 3, lwd = 1.5, col = "darkgreen")
> points(Rok, Prutok, type = "o", pch = 21, cex = 0.85, bg = "red1")
```



Obrázek 5: Grafické znázornění polynomické regrese 3. stupně pro data *Průměrné roční průtoky vody v řece Niger v Coulicoure (Mali)* spolu s intervaly spolehlivosti

Opět intervaly spolehlivost zdůrazníme pomocí ploch.

```
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(Rok[c(1, n)], yrange, type = "n", main = TXT, cex.main = 0.85)
> xx <- c(gridt, rev(gridt))
> yy <- c(ci.conf[, 2], rev(ci.conf[, 3]))
> polygon(xx, yy, col = "gray75", border = "gray75")
> yy <- c(ci.conf[, 2], rev(ci.pred[, 2]))
> polygon(xx, yy, col = "gray85", border = "gray85")
> yy <- c(ci.conf[, 3], rev(ci.pred[, 3]))
> polygon(xx, yy, col = "gray85", border = "gray85")
> matlines(gridt, ci.conf[, 1], lty = 1, lwd = 2, col = "red2")
> points(Rok, Prutok, type = "o", pch = 21, cex = 0.85, bg = "yellowgreen")
```



Obrázek 6: Grafické znázornění polynomické regrese 3. stupně pro data *Průměrné roční průtoky vody v řece Niger v Coulicoure (Mali)* spolu s intervaly spolehlivosti – pomocí ploch

Zkusíme odhadnout jiný model s polynomem šestého stupně, budeme tudíž předpokládat, že jednotlivá pozorování lze popsat vztahem

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5 + \beta_6 x_i^6 + \varepsilon_i.$$

Odhadneme parametry nového modelu.

```
> model6 <- lm(Prutok ~ Rok + I(Rok^2) + I(Rok^3) + I(Rok^4) + I(Rok^5) +
     I(Rok^6))
> summary(model6)


Call:
lm(formula = Prutok ~ Rok + I(Rok^2) + I(Rok^3) + I(Rok^4) +
    I(Rok^5) + I(Rok^6))

Residuals:
     Min       1Q   Median       3Q      Max
-22.2179  -7.0170  -0.9496   7.6885  27.1292

Coefficients: (3 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.483e+07  4.903e+06  -3.024  0.00403 **
Rok          2.302e+04  7.614e+03   3.023  0.00404 **
I(Rok^2)    -1.191e+01  3.941e+00  -3.022  0.00405 **
I(Rok^3)     2.055e-03  6.800e-04   3.022  0.00406 **
I(Rok^4)            NA         NA      NA       NA
I(Rok^5)            NA         NA      NA       NA
```

```
I(Rok^6)            NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.14 on 47 degrees of freedom
Multiple R-squared: 0.2108,        Adjusted R-squared: 0.1604
F-statistic: 4.183 on 3 and 47 DF,  p-value: 0.0105
```

Vidíme, že odhad modelu s vysokým stupněm polynomu je numericky problematický, protože se koeficienty od čtvrtého stupně výše nepodařilo spočítat. V takových případech se doporučuje proměnnou `Rok` centrovat a odhadnout parametry pro model s touto centrovanou proměnnou. Centrování znamená, že od každé hodnoty proměnné `Rok` odečteme průměr všech hodnot. Model se tak změní do podoby

$$Y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \beta_3 z_i^3 + \beta_4 z_i^4 + \beta_5 z_i^5 + \beta_6 z_i^6 + \varepsilon_i,$$

kde $z_i = x_i - \bar{x}$.

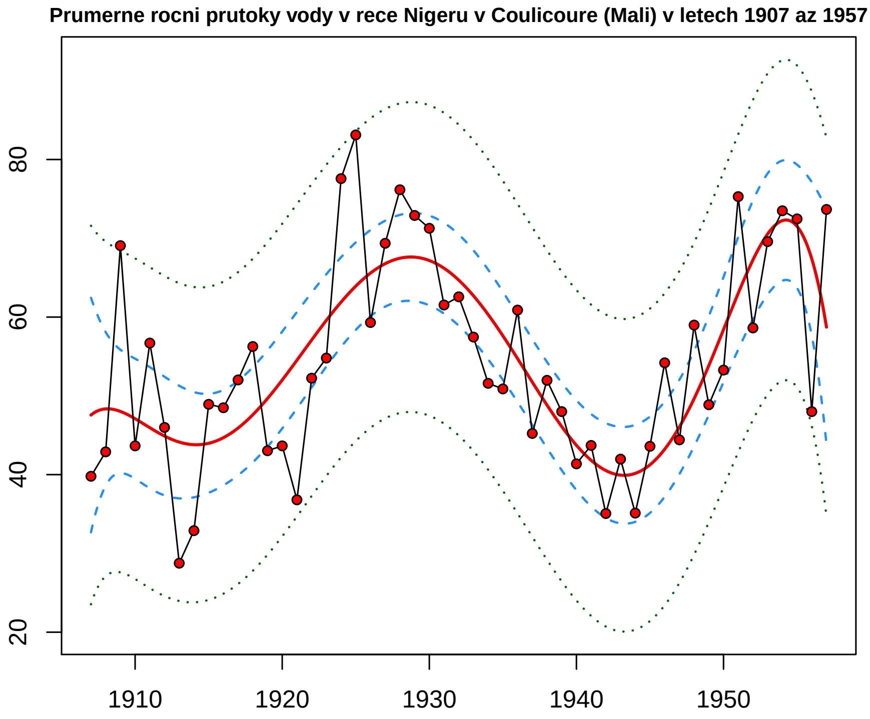
Provedeme nový odhad modelu s centrovanou proměnnou.

```
> delta <- mean(Rok)
> RokC <- Rok - delta
> model6C <- lm(Prutok ~ RokC + I(RokC^2) + I(RokC^3) + I(RokC^4) + I(RokC^5) +
     I(RokC^6))
> summary(model6C)


Call:
lm(formula = Prutok ~ RokC + I(RokC^2) + I(RokC^3) + I(RokC^4) +
    I(RokC^5) + I(RokC^6))

Residuals:
     Min       1Q   Median       3Q      Max
-19.3877  -5.4541  -0.9673   5.1213  21.0478

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.469e+01  2.872e+00  22.523  < 2e-16 ***
RokC        -1.757e+00  3.909e-01  -4.494 5.02e-05 ***
I(RokC^2)   -2.385e-01  5.367e-02  -4.444 5.90e-05 ***
I(RokC^3)    1.045e-02  2.371e-03   4.408 6.63e-05 ***
I(RokC^4)    8.809e-04  2.268e-04   3.884 0.000341 ***
I(RokC^5)   -1.165e-05  3.209e-06  -3.631 0.000733 ***
I(RokC^6)   -8.461e-07  2.526e-07  -3.350 0.001666 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.36 on 44 degrees of freedom
Multiple R-squared: 0.5608,        Adjusted R-squared: 0.5009
F-statistic: 9.364 on 6 and 44 DF,  p-value: 1.278e-06
```

Výsledky opět znázorníme graficky.

```
> n <- length(RokC)
> gridt <- seq(Rok[1], Rok[n], length.out = 500)
> ci.conf <- predict(model6C, newdata = data.frame(RokC = gridt - delta),
     interval = "confidence")
```

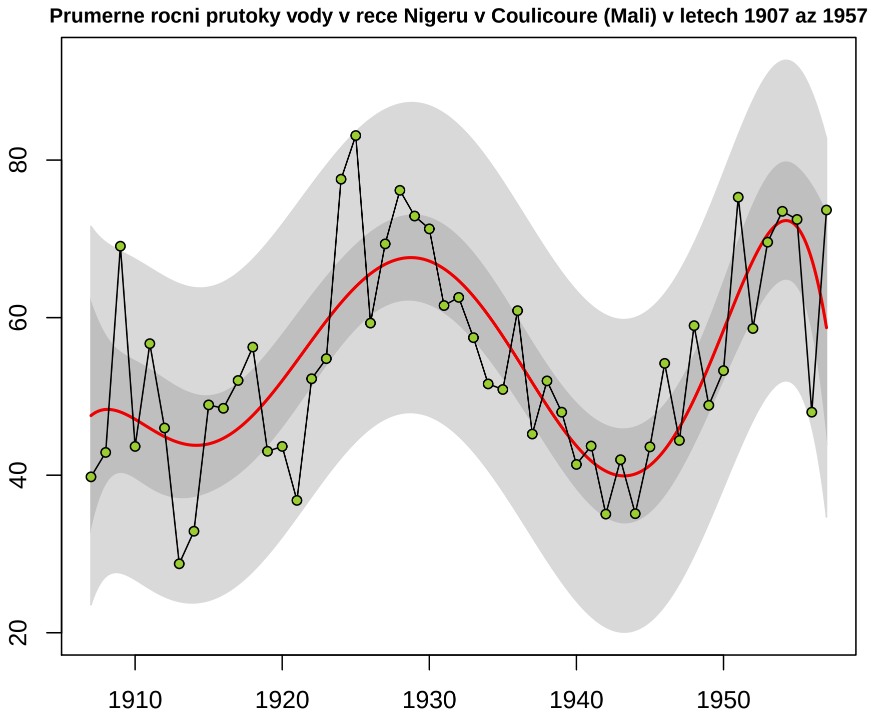
```
> ci.pred <- predict(model6C, newdata = data.frame(RokC = gridt - delta),
     interval = "prediction")
> yrange <- range(c(Prutok, ci.pred[, 2], ci.pred[, 3]))
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(Rok[c(1, n)], yrange, type = "n", main = TXT, cex.main = 0.85)
> matlines(gridt, ci.conf[, 1], lty = 1, lwd = 2, col = "red2")
> matlines(gridt, ci.conf[, 2:3], lty = 2, lwd = 1.5, col = "dodgerblue")
> matlines(gridt, ci.pred[, 2:3], lty = 3, lwd = 1.5, col = "darkgreen")
> points(Rok, Prutok, type = "o", pch = 21, cex = 0.85, bg = "red1")
```



Obrázek 7: Grafické znázornění polynomické regrese 6. stupně pro data *Průměrné roční průtoky vody v řece Niger v Coulicoure (Mali)* spolu s intervaly spolehlivosti

Opět intervaly spolehlivost zdůrazníme pomocí ploch.

```
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(Rok[c(1, n)], yrange, type = "n", main = TXT, cex.main = 0.85)
> xx <- c(gridt, rev(gridt))
> yy <- c(ci.conf[, 2], rev(ci.conf[, 3]))
> polygon(xx, yy, col = "gray75", border = "gray75")
> yy <- c(ci.conf[, 2], rev(ci.pred[, 2]))
> polygon(xx, yy, col = "gray85", border = "gray85")
> yy <- c(ci.conf[, 3], rev(ci.pred[, 3]))
> polygon(xx, yy, col = "gray85", border = "gray85")
> matlines(gridt, ci.conf[, 1], lty = 1, lwd = 2, col = "red2")
> points(Rok, Prutok, type = "o", pch = 21, cex = 0.85, bg = "yellowgreen")
```



Obrázek 8: Grafické znázornění polynomické regrese 6. stupně pro data *Průměrné roční průtoky vody v řece Niger v Coulicoure (Mali)* spolu s intervaly spolehlivosti – pomocí ploch

Na závěr příkladu nesníme zapomenout zrušit zpřístupnění jména proměnných v datovém rámci

```
> detach(dataNiger)
```

4 Úkoly

1. Načtěte datový soubor `japanese-vehicle.txt` obsahující údaje o množství motorových vozidel vyrobených v Japonsku v letech 1947–1989.

   – Data vykreslete do grafu.

   – Odhadněte parametry následujících čtyř modelů:

Model 1:	$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Model 2:	$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$
Model 3:	$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$
Model 4:	$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i,$

   kde Y_i je počet vyrobených vozidel a x_i je čas. Pokud to bude vhodné, pracujte s centrovaným časem.

   – Pro každý model do grafu vykreslete data, odhadnutou regresní křivku, interval spolehlivosti kolem střední hodnoty a predikční interval spolehlivosti.

- Na závěr vykreslete grafy pro všechny modely do jednoho obrázku.

2. Do datového souboru `US_population.txt` byly zaznamenány údaje o počtu obyvatel v USA v desetiletých intervalech v rozmezí let 1790–1980.

   - Data načtěte.
   - Do grafu vykreslete
     (a) původní hodnoty časové řady
     (b) časovou řadu, v níž jsou počty obyvatel zlogaritmovány.
   - Odhadněte parametry v modelech:

$$\begin{array}{ll} \text{Model 1:} & Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \\ \text{Model 2:} & \ln Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \end{array}$$

     kde Y_i je počet obyvatel a x_i je rok.
   - Pro každý model do grafu vykreslete data, odhadnutou regresní křivku, interval spolehlivosti kolem střední hodnoty a predikční interval spolehlivosti. U modelu s logaritmem vykreslete jak graf se zlogaritmovanými hodnotami, tak s hodnotami na původní škále.
   - Na závěr vykreslete grafy pro oba modely do jednoho obrázku (v původním měřítku bez logaritmování).