M5201 – 3. cvičení : *Polynomická regrese*

1 Aplikace polynomického trendu na časové řady

Příklad 1

Máme k dispozici měsíční údaje o počtu nezaměstnaných v Austrálii. Data pokrývají období od února 1978 do srpna 1995.

Data jsou uložena v souboru `unemployed.dat`. Na prvním řádku je uveden nadpis, na druhém řádku je hlavička a pod ní jsou ve dvou sloupcích data.

Nejprve načteme nadpis. K tomu použijeme funkci `scan`, kde argumentem `nlines=1` určíme, že se má načítat pouze první řádek a všechny další ignorovat, a argumentem `what=character()` říkáme, že načítané položky jsou textové řetězce.

```
> fileDat <- paste(data.library, "unemployed.dat", sep = "")
> nadpis <- scan(fileDat, what=character(), nlines=1)
```

Nově vytvořená proměnná `nadpis` je vektor, jehož složkami jsou jednotlivá slova.

```
> nadpis
```

```
 [1] "Monthly"    "number"     "of"         "unemployed" "persons"
 [6] "in"         "Australia." "Feb"        "1978"       "-"
[11] "Aug"        "1995"
```

Abychom slova spojili do jednoho textového řetězce (ten můžeme využít třeba k popisování grafů), použijeme funkci `paste()`. Jako první argument zadáme, co chceme spojovat, tj. vektor `nadpis`, do druhého argumentu `collapse` zadáme spojovací textový řetězec, který se použije k oddělení slov z vektoru `nadpis`, v našem případě to bude mezera.

```
> (TXT <- paste(nadpis, collapse = " "))
```

```
[1] "Monthly number of unemployed persons in Australia. Feb 1978 - Aug 1995"
```

Nyní načteme vlastní data do datového rámce. Použijeme k tomu funkci `read.table()`. Protože data začínají až na druhém řádku, první řádek musíme při načítání přeskočit, proto zadáme argument `skip = 1`. Dále nezapomeneme uvést argument `header = TRUE`, protože na prvním načítaném řádku jsou názvy proměnných. Poté si prohlédneme strukturu vytvořeného datového rámce a jeho prvních šest řádků.

```
> unempl <- read.table(fileDat, header = T, skip = 1)
> str(unempl)
```

```
'data.frame':        211 obs. of  2 variables:
 $ Month     : Factor w/ 211 levels "1978-02","1978-03",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Unemployed: int  428800 424800 403400 398400 393500 380500 398300 387300 370400 372800 ...
```

```
> head(unempl)
```

```
    Month Unemployed
1 1978-02     428800
2 1978-03     424800
3 1978-04     403400
4 1978-05     398400
5 1978-06     393500
6 1978-07     380500
```

Z dat načtených do datového rámce vytvoříme časovou řadu a vykreslíme ji. Protože první údaj není z ledna roku 1978, ale až z února, zadáme čas začátku časové řady ve tvaru `c(rok, měsíc)`, tj. `c(1978, 2)`.

```
> unemplTS<-ts(unempl[,2], start = c(1978, 2), frequency = 12)
> par(mar = c(2,2,1,0) + 0.5)
> plot(unemplTS, main = TXT, cex.main = 0.85)
```

Obrázek 1: Počet nezaměstnaných v Austrálii od února 1978 do srpna 1995.

Nejprve zkusíme odhadnout model s lineárním trendem

$$Y_i = \beta_0 + \beta_1 t_i + \varepsilon_i.$$

Jako obvykle použijeme funkci `lm()`. Informace o odhadnutých parametrech získáme pomocí funkce `summary()`.

```
> Time <- time(unemplTS)
> model1 <- lm(unemplTS ~ Time)
> summary(model1)
```

```
Call:
lm(formula = unemplTS ~ Time)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-264506  -55088   -5523   71548  239487

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -56294351    2878294  -19.56   <2e-16 ***
Time            28648       1449   19.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 106800 on 209 degrees of freedom
Multiple R-squared: 0.6517,        Adjusted R-squared:  0.65
F-statistic: 391.1 on 1 and 209 DF,  p-value: < 2.2e-16
```

Příkaz `summary()` zobrazí výpis informací o modelu. Nejprve vypisuje zadaný tvar modelu, poté informace o reziduích – minimální a maximální reziduum, medián a horní a dolní kvartil.

Důležitá je další tabulka obsahující informace o odhadnutých regresních koeficientech. V ní jsou uvedeny v prvním sloupci odhady koeficientů, ve druhém sloupci jsou směrodatné odchylky těchto odhadů, ve třetím sloupci jsou uvedeny t-statistiky pro testování nulové hypotézy, že daný regresní koeficient se rovná nule, a v posledním sloupci jsou p-hodnoty tohoto testu. Pokud je p-hodnota menší než zvolená hladina významnosti testu, nulovou hypotézu zamítáme (a říkáme, že daný koeficient je významný).

Na závěr jsou uvedeny informace a statistiky týkající se souhrnně celého modelu.

V našem případě hodnota *p–value* označená jako `Pr(>|t|)` ukazuje, že lineární trend je významný (koeficient β_1 se statisticky významně liší od nuly).

V dalším kroku do grafu zakreslíme regresní přímku a kolem ní intervaly spolehlivosti pro odhad střední hodnoty predikovaného počtu nezaměstnaných. Nejprve vygenerujeme síť hodnot proměnné *Time*, kterou uložíme do vektoru `gridt`, a pomocí funkce `predict()` pro každý její bod spočítáme vyrovnanou hodnotu počtu nezaměstnaných osob (tj. číslo získané z odhadnutého modelu po dosazení dané složky `gridt`). Argumentem `se.fit = T` dosáhneme toho, že se pro každou vyrovnanou hodnotu spočítá také její směrodatná odchylka.

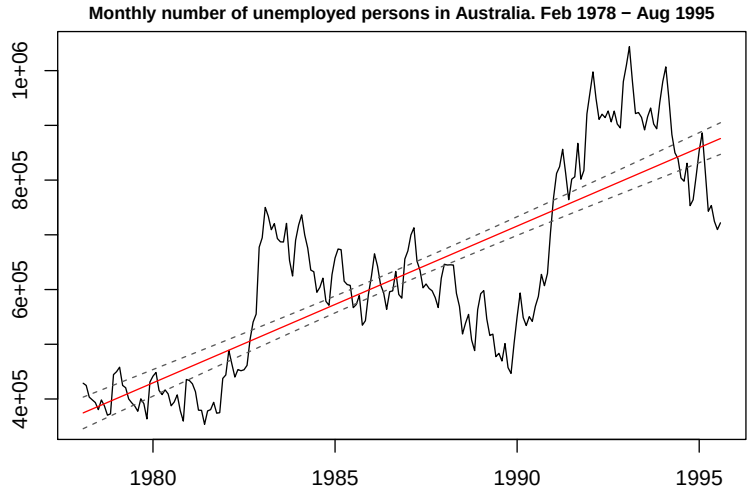
```
> n <- length(Time)
> gridt <- seq(Time[1],Time[n], length.out = 300)
> pred1 <- predict(model1, data.frame(Time = gridt), se.fit = T)
> str(pred1)


List of 4
 $ fit           : Named num [1:300] 374587 376264 377941 379617 381294 ...
  ..- attr(*, "names")= chr [1:300] "1" "2" "3" "4" ...
 $ se.fit        : Named num [1:300] 14654 14581 14508 14435 14362 ...
  ..- attr(*, "names")= chr [1:300] "1" "2" "3" "4" ...
 $ df            : int 209
 $ residual.scale: num 106812
```

Dále zkonstruujeme intervaly spolehlivosti kolem střední hodnoty predikce. Pro intervaly spolehlivosti zvolíme hladinu $\alpha = 0.05$. K tomu zjistíme kvantil Studentova t-rozdělení $t_{1-\frac{\alpha}{2}}(n-2)$, který budeme potřebovat k určení šířky intervalu spolehlivosti.

```
> alpha <- 0.05
> CV <- qt(1-alpha/2,pred1$df)
```

Teď už máme vše potřebné k vykreslení grafu.

```
> par(mar=c(2,2,1,0)+0.5)
> plot(unemplTS, main = TXT, cex.main = 0.85)
> matlines(gridt, cbind(pred1$fit, pred1$fit - CV*pred1$se.fit, pred1$fit + CV*pred1$se.fit),
         lty = c(1,2,2), col = c("red", "gray35", "gray35"))
```



Obrázek 2: Lineární trend pro počet nezaměstnaných v Austrálii v období od února 1978 do srpna 1995.

Z grafu je vidět, že regresní přímka zdaleka nekopíruje mnohem komplikovanější průběh časové řady. K tomu bychom potřebovali polynom vyššího stupně. Proto zkusíme daty proložit například polynom čtvrtého stupně

$$Y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \beta_4 t_i^4 + \varepsilon_i.$$

```
> Time <-time(unemplTS)
> model4 <- lm(unemplTS ~ Time + I(Time^2) + I(Time^3) + I(Time^4))
> summary(model4)


Call:
lm(formula = unemplTS ~ Time + I(Time^2) + I(Time^3) + I(Time^4))

Residuals:
    Min      1Q  Median      3Q     Max
-258748  -61158   -7673   70793  236290

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.300e+09  1.259e+09   1.032    0.303
Time        -1.336e+06  1.268e+06  -1.054    0.293
I(Time^2)    3.435e+02  3.190e+02   1.077    0.283
I(Time^3)           NA         NA      NA       NA
I(Time^4)           NA         NA      NA       NA

Residual standard error: 106800 on 208 degrees of freedom
Multiple R-squared: 0.6536,        Adjusted R-squared: 0.6503
F-statistic: 196.3 on 2 and 208 DF,  p-value: < 2.2e-16
```

Vidíme, že s odhadem parametrů polynomu čtvrtého stupně jsou již spojeny numerické problémy (viz hodnoty `NA` v řádku u `I(Time^3)` a `I(Time^4)`. Proto provedeme centrování času (od hodnot proměnné `Time` odečteme průměr) a výpočet provedeme znovu pro ekvivalentní model s centrovaným časem.

```
> delta <- 0.5*(start (unemplTS)[1] + end (unemplTS)[1])
> TimeC <- Time - delta
> model4 <-lm(unemplTS ~ TimeC + I(TimeC^2)+ I(TimeC^3)+ I(TimeC^4))
> summary(model4)


Call:
lm(formula = unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4))

Residuals:
    Min      1Q  Median      3Q     Max
-244393  -82082    7976   77540  210508

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 581059.68   13304.70  43.673  < 2e-16 ***
TimeC        20023.15    3557.15   5.629 5.88e-08 ***
I(TimeC^2)    3831.72    1073.23   3.570 0.000444 ***
I(TimeC^3)     203.18      72.34   2.809 0.005454 **
I(TimeC^4)     -55.17      15.64  -3.528 0.000517 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 103400 on 206 degrees of freedom
Multiple R-squared: 0.6785,        Adjusted R-squared: 0.6723
F-statistic: 108.7 on 4 and 206 DF,  p-value: < 2.2e-16
```

Vidíme, že koeficienty u všech mocnin jsou statisticky významné.

Oba dva modely (`model1` a `model4`) můžeme porovnat pomocí příkazu `anova()`, kde zadáme jako argumenty tyto dva modely. Tato funkce slouží obecně k porovnání dvou modelů, z nichž jeden je submodelem druhého (obsahuje navíc některé proměnné, které obecnější model nezahrnuje). V takové situaci bychom rádi zjistili, zda model s více proměnnými popisuje data významně lépe než jednodušší model.

```
> anova(model4, model1)


Analysis of Variance Table

Model 1: unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4)
Model 2: unemplTS ~ Time
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1    206 2.2008e+12
2    209 2.3844e+12 -3 -1.8359e+11 5.7282 0.0008765 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

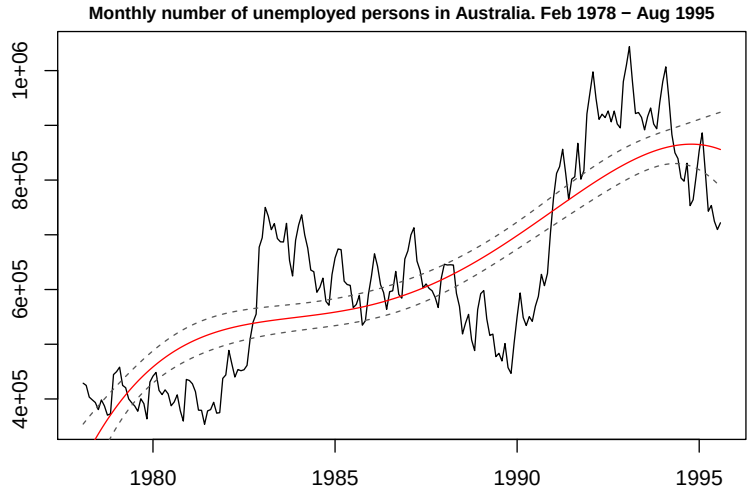
Ve výpisu jsou nejdůležitější poslední dva sloupce tabulky. Ve sloupci označeném `F` je uvedena F-statistika, která se používá k testování hypotézy, zda platí jednodušší model. Jinými

slovy se testuje, jestli složitější model přináší významné zlepšení ve vyrovnání dat, nebo je toto zlepšení pouze malé, takže jednodušší model vykazuje srovnatelné výsledky při nižším počtu odhadovaných parametrů. V posledním sloupci je uvedena p-hodnota tohoto testu (`Pr(>F)`).

V tomto případě je p-hodnota malá (nižší než třeba nejčastěji používaná hladina významnosti 0.05), takže odtud můžeme usoudit, že model s polynomem čtvrtého stupně je významně lepší než model s lineárním trendem.

Výsledky modelu s polynomem řádu čtyři opět zakreslíme do grafu. Postupovat budeme analogicky jako v případě lineárního trendu.

```
> gridtC <- gridt-delta
> pred4 <- predict(model4, data.frame(TimeC = gridtC), se.fit = T)
> alpha <- 0.05
> CV <- qt(1-alpha/2, pred4$df)
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(unemplTS, main = TXT, cex.main = 0.85)
> matlines(gridt, cbind(pred4$fit, pred4$fit - CV*pred4$se.fit, pred4$fit + CV*pred4$se.fit),
        lty = c(1, 2, 2), col = c("red", "gray35", "gray35"))
```



Obrázek 3: Počet nezaměstnaných v Austrálii v období od února 1978 do srpna 1995 s polynomickým trendem čtvrtého stupně.

Protože odhadnutý parametr u čtvrté mocniny se statisticky významně liší od nuly, zkusíme uvažovat model vyššího řádu, řekněme 6. řádu.

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


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

Residuals:
    Min      1Q  Median      3Q     Max
-155329  -56223   -3842   49448  173791
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.094e+05  1.048e+04  58.163  < 2e-16 ***
TimeC       -2.992e+04  4.225e+03  -7.082 2.26e-11 ***
I(TimeC^2)  -2.307e+03  1.621e+03  -1.423   0.1562
I(TimeC^3)   3.207e+03  2.220e+02  14.445  < 2e-16 ***
I(TimeC^4)   1.227e+02  5.741e+01   2.138   0.0337 *
I(TimeC^5)  -3.505e+01  2.628e+00 -13.337  < 2e-16 ***
I(TimeC^6)  -1.116e+00  5.428e-01  -2.056   0.0410 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 69870 on 204 degrees of freedom
Multiple R-squared: 0.8545,        Adjusted R-squared: 0.8503
F-statistic: 199.8 on 6 and 204 DF,  p-value: < 2.2e-16
```

Z tabulky vyplývá, že polynomický trend šestého stupně je statisticky významný. Ještě porovnáme modely s šestým a čtvrtým stupněm pomocí funkce `anova()`.

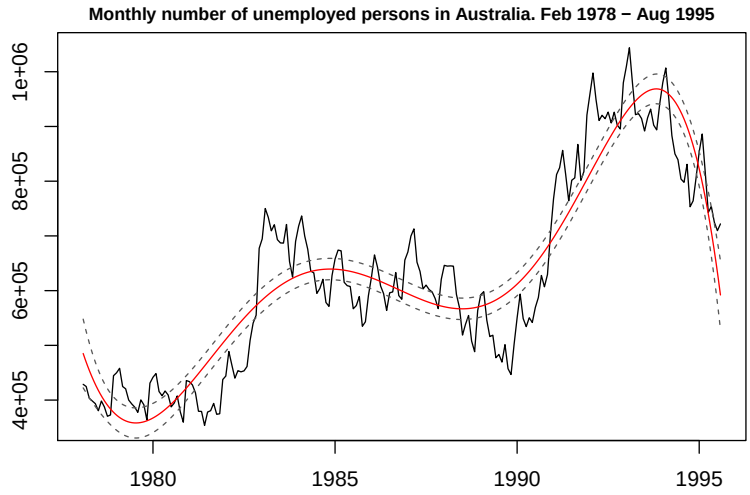
```
> anova(model6, model4)
```

```
Analysis of Variance Table

Model 1: unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4) + I(TimeC^5) +
    I(TimeC^6)
Model 2: unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4)
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1    204 9.9576e+11
2    206 2.2008e+12 -2 -1.2051e+12 123.44 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Protože p-hodnota srovnávacího testu označená jako `Pr(>F)` je menší než 0.05, je model s polynomem šestého stupně významně lepší než model s polynomem čtvrtého řádu.

Výsledky opět zakreslíme do grafu.

```
> gridtC <- gridt - delta
> pred6 <- predict(model6, data.frame(TimeC = gridtC), se.fit = T)
> alpha <- 0.05
> CV <- qt(1-alpha/2, pred6$df)
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(unemplTS, main = TXT, cex.main = 0.85)
> matlines(gridt, cbind(pred6$fit, pred6$fit - CV*pred6$se.fit, pred6$fit + CV*pred6$se.fit),
         lty = c(1, 2, 2), col = c("red", "gray35", "gray35"))
```



Obrázek 4: Počet nezaměstnaných v Austrálii v období od února 1978 do srpna 1995 s polynomickým trendem šestého řádu.

2 Odhady počtu regresních koeficientů

Mějme **regresní model** plné hodnosti (např. polynomický regresní model):

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

Odhad neznámých parametrů $\boldsymbol{\beta}$ *metodou nejmenších čtverců* je dán vztahem:

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

Pro **odhad počtu regresních koeficientů** se nabízejí dvě intuitivní metody vycházející z testování významnosti regresních koeficientů.

- **Od nejmenšího k nejvyššímu počtu regresorů**:

  Začneme od $k = 1$ a postupně zvyšujeme počet regresoru o jeden a zároveň testujeme hypotézu
  $\boxed{H_0 : \beta_k = 0 \text{ proti } H_1 : \beta_k \neq 0}$ pomocí statistiky $T_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{s_k^2 v_{kk}}} \sim t(n-k)$,
  kde $(\mathbf{X}'\mathbf{X})^{-1} = (v_{ij})_{i,j=1,\ldots,k}$ (viz. Anděl, 1993).

  Jestliže H_0 **zamítneme** $\Rightarrow$ **zvyšujeme** počet regresoru.

- **Od maximálního počtu regresorů dolů**:

  Nejprve zvolíme maximální stupeň polynomu $k = k_{max}$. Testujeme opět hypotézu
  $\boxed{H_0 : \beta_k = 0 \text{ proti } H_1 : \beta_k \neq 0}$
  pomocí statistiky T_k.

  Jestliže H_0 **nezamítneme** $\Rightarrow$ **snižujeme** počet regresorů.

Přestože jsou tyto postupy jednoduché, neukázaly se jako zcela spolehlivé. Proto vyložme ještě jiný přístup k tomuto problému, který byl nalezen poměrně nedávno.

Penalizační metoda odhadu počtu regresních koeficientů

Předpokládejme, že $\boxed{k_0}$ je skutečný počet regresních parametrů, který neznáme a chceme ho zjistit.
Lze ukázat, že platí pro $k < k_0$: $E(s_k^2) > \sigma^2$
pro $k \geq k_0$: $E(s_k^2) = \sigma^2$.

Pokud bychom vykreslili graf s_k^2 pro $k = 0, \ldots, k_{max}$, měl by graf přibližně klesat až do bodu k_0 a od něj dál by měl být zhruba vodorovný. Slova „zhruba“ a „přibližně“ jsou na místě, protože vztahy uvedené výše platí pouze pro střední hodnotu s_k^2. Tudíž zůstává otázkou, jak z grafu hodnot s_k^2 určit právě tu hodnotu k_0, od níž počínaje již graf dostává vodorovný charakter.

Ukázalo se, že místo s_k^2 je lepší řídit se kritériem $\boxed{A_k = s_k^2(1 + kw_n),}$ kde w_n je tzv. *penalizační funkce* a udává jakousi relativní cenu, kterou je třeba zaplatit za přidání každého dalšího parametru.

V praxi se osvědčilo volit $\boxed{w_n = \frac{1}{\sqrt[4]{n}}}$, tj. $A_k = s_k^2(1 + \frac{k}{\sqrt[4]{n}})$.

Za odhad správného počtu regresních koeficientů $\hat{k}$ se bere hodnota $k \in \{0, 1, \ldots, k_{max}\}$, pro kterou A_k **nabývá svého minima**, kde k_{max} je maximální počet parametru, které jsme ochotni uvažovat a o němž jsme si jisti, že $k_0 \leq k_{max}$.

Kromě tohoto existují ještě **další kritéria pro určení počtu regresních koeficientů**:

Akaikeovo informační kritérium	$AIC_k = \ln s_k^2 + \frac{2k}{n}$	nadhodnocuje k_0
Swarz (1978) a Rissanen (1978)	$SR_k = \ln s_k^2 + \frac{k \ln n}{n}$	
Hannan a Quinn (1979)	$HQ_k = \ln s_k^2 + \frac{2kc \ln \ln n}{n}$	$c > 1$, obvykle $c = 2$ nebo 3

Příklad 1 (pokračování)

Na základě kritérií uvedených výše určíme vhodný stupeň polynomu, který bychom použili v lineárním regresním modelu k popisu dat týkajících se počtu nezaměstnaných v Austrálii. V souboru `FunkceM5201.R` je naprogramována funkce `fndptr`, pomocí níž jsme schopni všechna kritéria vypočítat a navíc přehledně znázornit graficky.

Protože funkce není součástí R, musíme nejprve dotyčný soubor načíst.

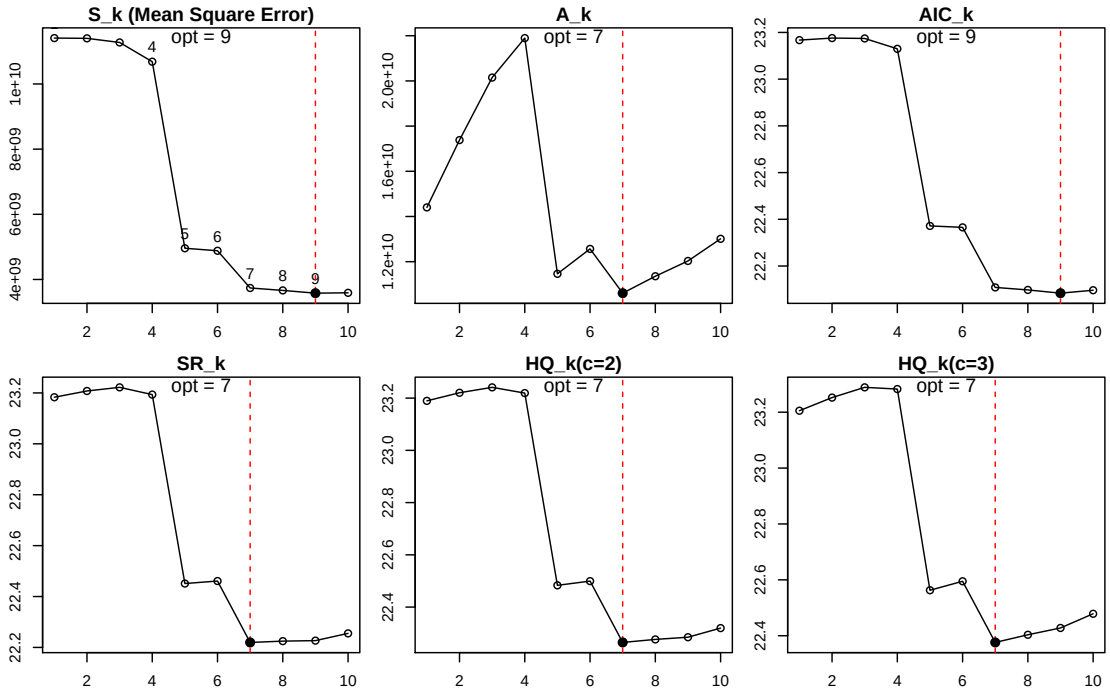
```
> fileSkript <- paste(data.library, "FunkceM5201.R", sep = "")
> source(fileSkript)
```

Funkci `fndptr` aplikujeme na zjištěné počty nezaměstnaných osob v Austrálii, kdy se budeme snažit určit nejvhodnější počet regresorů, tj. ideální stupeň polynomu. Funkce implicitně nastavuje jako minimální stupeň polynomu 1 a maximální stupeň 10. Maximální uvažovaný stupeň polynomu lze změnit pomocí argumentu `maxdgr`. (*Pozn.:* Pokud by nás zajímalo, jak je funkce napsána, stačí zadat příkaz `fndptr` do příkazového řádku a R funkci vypíše. Stejně to platí i pro ostatní funkce.)

```
> fndptr(unemplTS)
```

```
   dgr significant         S_k         A_k    AIC_k     SR_k HQ_k(c=2)
1    1           1 11408758559 14402179383 23.16713 23.18301  23.18945
2    2           0 11400054622 17382328806 23.17584 23.20761  23.22048
3    3           0 11274406531 20148915474 23.17424 23.22189  23.24120
4    4           1 10683670853 21896362450 23.12990 23.19344  23.21918
5    5           1  4958065549 11462531651 22.37167 22.45110  22.48328
6    6           1  4881185691 12565514964 22.36553 22.46084  22.49945
7    7           1  3738919271 10606021527 22.10841 22.21961  22.26466
8    8           1  3664008220 11354884278 22.09765 22.22474  22.27622
9    9           1  3579454855 12032025028 22.08378 22.22675  22.28468
10  10           0  3590574729 13011495602 22.09636 22.25522  22.31958
   HQ_k(c=3)
1   23.20535
2   23.25228
3   23.28890
4   23.28278
5   22.56278
6   22.59485
7   22.37596
```

```
8   22.40342
9   22.42777
10  22.47858
```



Obrázek 5: Penalizační kritéria pro výběr vhodného stupně polynomu pro měsíční počty nezaměstnaných v Austrálii v letech 1978–1995.

Čtyři z šesti kritérií nám doporučují zvolit polynom 7. stupně, zbylá dvě kritéria navrhují 9. stupeň.

Proto nyní odhadneme model s polynomickým trendem 7. stupně.

```
> model7 <- lm(unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4) + I(TimeC^5) + I(TimeC^6) +
        I(TimeC^7))
> summary(model7)


Call:
lm(formula = unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4) +
    I(TimeC^5) + I(TimeC^6) + I(TimeC^7))

Residuals:
    Min      1Q  Median      3Q     Max
-135862  -44198   -7312   39573  168812

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.200e+05  9.267e+03  66.909  < 2e-16 ***
TimeC       -6.097e+04  5.376e+03 -11.342  < 2e-16 ***
I(TimeC^2)  -6.029e+03  1.494e+03  -4.036 7.70e-05 ***
I(TimeC^3)   6.850e+03  4.974e+02  13.773  < 2e-16 ***
I(TimeC^4)   3.001e+02  5.497e+01   5.460 1.38e-07 ***
I(TimeC^5)  -1.402e+02  1.341e+01 -10.454  < 2e-16 ***
```

```
I(TimeC^6)  -3.119e+00  5.376e-01  -5.801 2.50e-08 ***
I(TimeC^7)   8.582e-01  1.078e-01   7.958 1.22e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61150 on 203 degrees of freedom
Multiple R-squared: 0.8891,        Adjusted R-squared: 0.8853
F-statistic: 232.6 on 7 and 203 DF,  p-value: < 2.2e-16
```

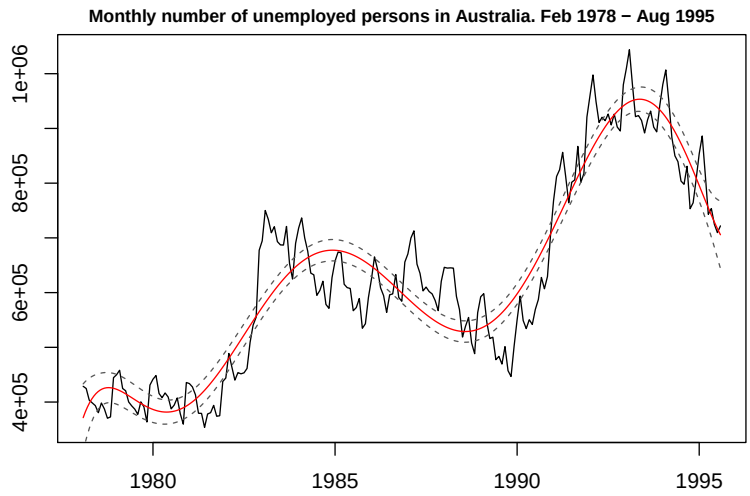
Vidíme, že regresní koeficient u sedmé mocniny se statisticky významně liší od nuly. Pomocí příkazu `anova()` srovnáme model s šestým a sedmým stupněm.

```
> anova(model7, model6)


Analysis of Variance Table

Model 1: unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4) + I(TimeC^5) +
    I(TimeC^6) + I(TimeC^7)
Model 2: unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4) + I(TimeC^5) +
    I(TimeC^6)
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1    203 7.5900e+11
2    204 9.9576e+11 -1 -2.3676e+11 63.324 1.216e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Protože p-hodnota srovnávacího testu označená jako `Pr(>F)` je menší než 0.05, je model s polynomem sedmého stupně významně lepší než model s polynomem šestého řádu.

Výsledky modelu se sedmým stupněm zakreslíme do grafu.

```
> pred7 <- predict(model7, data.frame(TimeC = gridtC), se.fit = T)
> alpha <- 0.05
> CV <- qt(1-alpha/2, pred7$df)
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(unemplTS, main = TXT, cex.main = 0.85)
> matlines(gridt, cbind(pred7$fit, pred7$fit - CV*pred7$se.fit, pred7$fit + CV*pred7$se.fit),
         lty = c(1, 2, 2), col = c("red", "gray35", "gray35"))
```



Obrázek 6: Počty nezaměstnaných osob v letech 1978-1995 s polynomickým trendem sedmého stupně.

Odhadneme ještě model 9. řádu a porovnáme jej s modelem 7. řádu.

```
> model9 <- lm(unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4) + I(TimeC^5) +
         I(TimeC^6) + I(TimeC^7) + I(TimeC^8) + I(TimeC^9))
> summary(model9)


Call:
lm(formula = unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4) +
    I(TimeC^5) + I(TimeC^6) + I(TimeC^7) + I(TimeC^8) + I(TimeC^9))

Residuals:
    Min      1Q  Median      3Q     Max
-123054  -48150   -4969   39129  167796

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.262e+05  1.021e+04  61.309  < 2e-16 ***
TimeC       -4.638e+04  7.205e+03  -6.438 8.73e-10 ***
I(TimeC^2)  -8.514e+03  2.696e+03  -3.158  0.00184 **
I(TimeC^3)   4.218e+03  1.045e+03   4.038 7.68e-05 ***
I(TimeC^4)   4.505e+02  1.766e+02   2.551  0.01147 *
I(TimeC^5)  -1.354e+01  4.846e+01  -0.279  0.78019
I(TimeC^6)  -5.929e+00  3.998e+00  -1.483  0.13961
I(TimeC^7)  -1.378e+00  8.786e-01  -1.569  0.11825
I(TimeC^8)   1.577e-02  2.902e-02   0.543  0.58756
I(TimeC^9)   1.309e-02  5.451e-03   2.402  0.01720 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59830 on 201 degrees of freedom
Multiple R-squared: 0.8949,        Adjusted R-squared: 0.8902
F-statistic: 190.2 on 9 and 201 DF,  p-value: < 2.2e-16
```

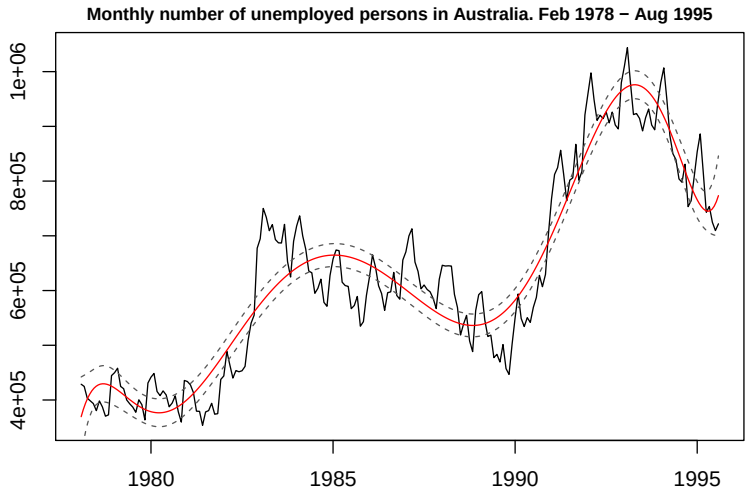
Koeficient u deváté mocniny je tedy statisticky významný.

```
> anova(model9, model7)


Analysis of Variance Table

Model 1: unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4) + I(TimeC^5) +
    I(TimeC^6) + I(TimeC^7) + I(TimeC^8) + I(TimeC^9)
Model 2: unemplTS ~ TimeC + I(TimeC^2) + I(TimeC^3) + I(TimeC^4) + I(TimeC^5) +
    I(TimeC^6) + I(TimeC^7)
  Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
1    201 7.1947e+11
2    203 7.5900e+11 -2 -3.953e+10 5.5218 0.004629 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P-hodnota srovnávacího testu označená jako `Pr(>F)` je opět menší než 0.05, tudíž je model s polynomem devátého stupně významně lepší než model s polynomem sedmého stupně.

Výsledky modelu s devátým stupněm opět zakreslíme do grafu.

```
> pred9 <- predict(model9, data.frame(TimeC = gridtC), se.fit = T)
> alpha <- 0.05
> CV <- qt(1-alpha/2, pred9$df)
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(unemplTS, main = TXT, cex.main = 0.85)
> matlines(gridt, cbind(pred9$fit, pred9$fit - CV*pred9$se.fit, pred9$fit + CV*pred9$se.fit),
        lty = c(1, 2, 2), col = c("red", "gray35", "gray35"))
```

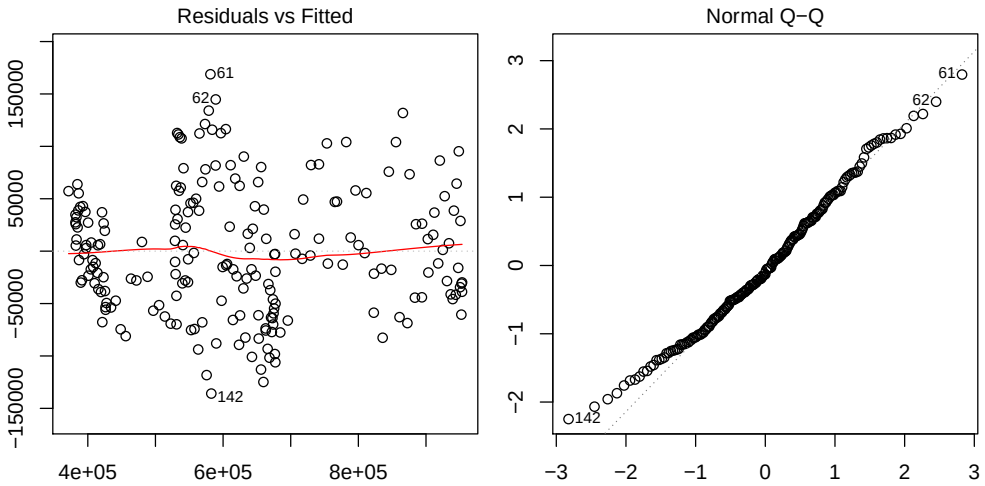


Obrázek 7: Počty nezaměstnaných osob v letech 1978-1995 s polynomickým trendem devátého stupně.

Kvalitu modelu lze posoudit pomocí analýzy reziduí, kterou provedeme pro modely s polynomy stupně sedm a devět.

Pomocí příkazu `plot`, ve kterém bude argumentem odhadnutý model, můžeme vykreslit celou řadu grafů. Vybereme si první a druhý (argument `which`).

```
> par(mfrow = c(1, 2), mar = c(2, 2, 1, 0) + 0.5)
> plot(model7, which = c(1, 2))
```

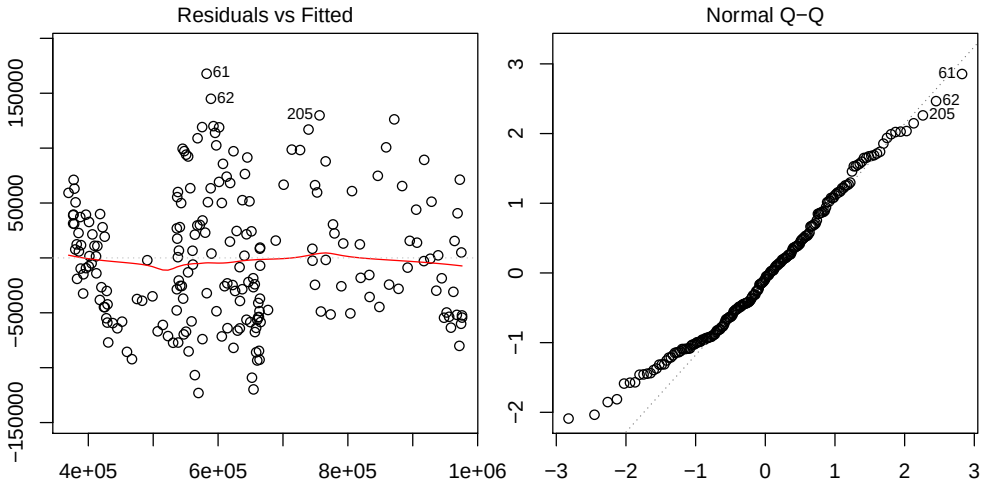


Obrázek 8: Vybrané grafy z analýzy reziduí pro polynomický trend sedmého řádu pro měsíční počty nezaměstnaných v Austrálii.

Na prvním grafu jsou na x-ové ose vyneseny vyrovnané hodnoty a proti nim na y-ové ose odpovídající rezidua. Pokud je model zvolen dobře, neměla by mezi vyrovnanými hodnotami a rezidui existovat žádná zjevná závislost a rezidua by měla kolísat kolem nuly.

Na druhém grafu je tzv. *Q-Q plot* (zkonstruovaný pro normální rozdělení). Jestliže mají rezidua přibližně normální rozdělení, měly by body na tomto grafu ležet přibližně na přímce.

```
> par(mfrow = c(1, 2), mar = c(2, 2, 1, 0) + 0.5)
> plot(model9, which = c(1, 2))
```

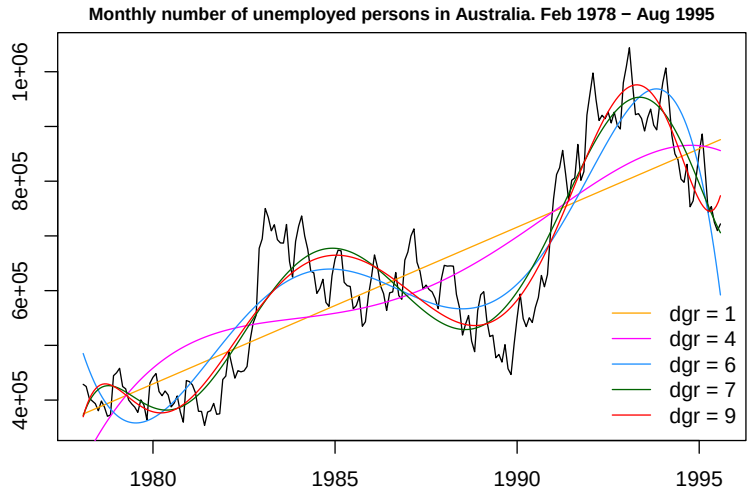


Obrázek 9: Vybrané grafy z analýzy reziduí pro polynomický trend devátého řádu pro měsíční počty nezaměstnaných v Austrálii.

Ani u jednoho modelu jsme při zběžné analýze reziduí nenarazili na nic, co by svědčilo o jeho nevhodnosti.

Na závěr všechny odhadnuté trendy zakreslíme do jediného grafu.

```
> par(mfrow = c(1, 1))
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(unemplTS, main = TXT, cex.main = 0.85)
> matlines(gridt, cbind(pred1$fit, pred4$fit, pred6$fit, pred7$fit, pred9$fit),
         lty = 1, col = c("orange", "magenta", "dodgerblue", "darkgreen", "red"))
>          legend("bottomright", legend = paste("dgr =", c(1, 4, 6, 7, 9)), bty = "n",
         lty = 1, col = c("orange", "magenta", "dodgerblue", "darkgreen", "red"))
```



Obrázek 10: Polynomické trendy řádu jedna, čtyři, šest, sedm a devět pro měsíční počty nezaměstnaných v Austrálii v letech 1978 až 1995.

3 Úkol

1. Načtěte datový soubor `births.dat`, který obsahuje roční údaje o počtu novorozenců na 10 000 žen ve věku 23 let v USA v letech 1917–1975. Zvolte vhodný stupeň polynomu pro lineární regresní model s polynomickým trendem. Použijte k tomu penalizační kritéria (funkce `fndptr`). Jestliže penalizační kritéria nedávají zcela jednoznačný výsledek, vyzkoušejte více regresních modelů.

2. Načtěte datový soubor `coffee_consumption.dat`. V něm jsou uvedeny údaje o roční spotřebě kávy v USA v letech 1910–1970.

   - Také pro tato data zvolte vhodný stupeň polynomu, zde ale použijte postup "od nejmenšího k nejvyššímu počtu regresorů" s postupným zvyšováním stupně polynomu a testováním statistické významnosti regresního koeficientu u nejvyšší mocniny modelu.
   - Poté vyzkoušejte opačný postup "od maximálního počtu regresorů dolů", jako maximální stupeň polynomu při tom zvolte 7. stupeň. Liší se vybraný stupeň polynomu?
   - Na závěr můžete zjistit, jaký stupeň polynomu doporučují použít penalizační kritéria.