Časové řady

Alena Černíková

alena.cernikova@ujep.cz

18. února 2025

Podmínky zápočtu

- **docházka**
- **semestrální práce**
  zadaná bude cca v půlce semestru
  bude zahrnovat aplikaci většího množství postupů
  důraz bude kladen na komentáře a interpretace
  součástí bude obhajoba minimálně přede mnou a dr. Škvorem

Obsah kurzu

- Úvod
- Dekompozice časových řad – trend
    - klouzavé průměry
    - exponenciální vyrovnávání
- Dekompozice časových řad – sezónnost
- Náhodná složka
- Indexace
- R/S analýza a Hurstův exponent
- Box-Jenkinsova metodologie – ARIMA modely
- Předpovědi
- Regresní závislost v časových řadách

Časová řada

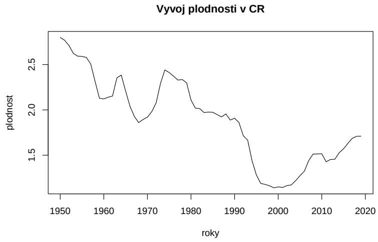
Co je časová řada

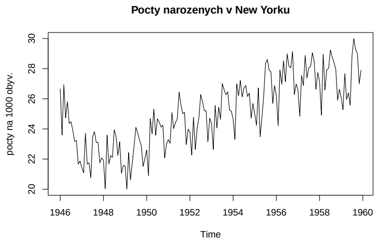
- soubor pozorování, realizací jedné náhodné veličiny y_t, kde t je čas
- čas může být diskrétní $t \in \{1, 2, \dots, n\}$, nebo i spojitý $t \in < 0, 1 >$
- důležité je, že pozorování nejsou nezávislá, ale jsou **závislá v čase**
- platí zde jiná pravidla při inferenci

Příklad.

- *Tlak pacienta měřený vždy v 7 hodin ráno.*
- *Měsíční průměrný kurz eura dle ČNB.*
- *Venkovní teplota měřená na stejném místě každou hodinu.*

Grafické zobrazení





Časová řada

Intervaly mezi dvěma sousedními měřeními

- předpokládá se, že jednotlivá pozorování mají mezi sebou stejné časové intervaly
- situace, v nichž se dělají opakovaná měření s různými časovými rozestupy vedou na jiné modely (longitudinal models) a nebudeme se jimi na tomto kurzu zabývat

Cíl analýzy časových řad

- **pochopení** mechanismu, který časovou řadu generuje
    - dekompozice časové řady – hledání trendu, sezónnosti, cykličnosti
    - Box-Jenkinsova metodologie – hledání časové závislosti mezi pozorováními
- konstrukce předpovědí vycházející jak ze závislosti na minulých hodnotách, tak z trendů
    - bodové a intervalové předpovědi
    - budeme brát pouze kvantitativní předpovědi získané exaktními výpočty (expertní předpovědi necháme expertům)
- zkoumání závislosti mezi časovými řadami

Typy časových řad

Časové řady mohou být diskrétní nebo spojité

- Příklady **diskrétních** řad
  - binární proces – hod mincí: řadu hodnot *panna, orel*
  - náhodná procházka – částice vycházející z bodu nula a v každém kroku se s pstí p posune doprava a s pstí q se posune doleva: řada bodů, ve kterých se částice nachází v čase t
  - Poissonův proces – př. počet hovorů na telefonní ústředně v čase t
    - počet hovorů za jednu časovou jednotku má Poissonovo rozdělení s intenzitou λ, počty za jednotlivé jednotky jsou vzájemně nezávislé
    - délka doby mezi dvěma hovory má exponenciální rozdělení s parametrem λ a jednotlivé doby mezi hovory jsou vzájemně nezávislé
    - získáváme rostoucí posloupnost počtu hovorů až do času t
- Kurz se bude věnovat převážně **spojitým** časovým řadám

Na co si dát pozor

Při sběru dat je potřeba ohlídat možné problémy

- Jaké zvolit body/okamžiky, kdy se bude časová řada měřit
- **Problém s kalendářem** – vzniká především u měsíčních ekonomických časových řad
  - různé měsíce mají různé počty dní
  - různé měsíce mají různé počty pracovních dní a víkendů
  - speciální pozornost je třeba věnovat pohyblivým svátkům
  - je potřeba volit vhodnou standardizaci, aby byly hodnoty srovnatelné
- **Délka časové řady** – je třeba zvolit dostatečně dlouhou řadu, aby v ní byl vidět aktuální trend, ale zase ne moc dlouhou, aby v ní byly prehistorická data, která mi aktuální trend zakryjí

Metody analýzy časových řad

- **Indexace** – doc. Sixta
  – výpočet speciálních indexů popisujících časovou řadu
- **Dekompozice** – já + doc. Vozár
  – analýza systematických složek řady
  – hledání trendu, sezónnosti, cykličnosti
- **Box-Jenkinsova metodologie**
  – analýza náhodné složky
  – hledání korelační struktury řady
- **Spektrální analýza**
  – myšlenka je, že časová řada je směs složek/sinusoid s různými frekvencemi
  – převádí se z časové složky na frekvenční složku
  – speciální nástroje: periodogram, spektrální hustota
- **R/S analýza a Hurstův exponent** – dr. Škvor

Dekompozice časových řad

Předpokládáme, že řada se řídí jedním z následujících dvou modelů

- **Aditivní model**

$$Y_t = Tr_t + C_t + S_t + \varepsilon_t$$

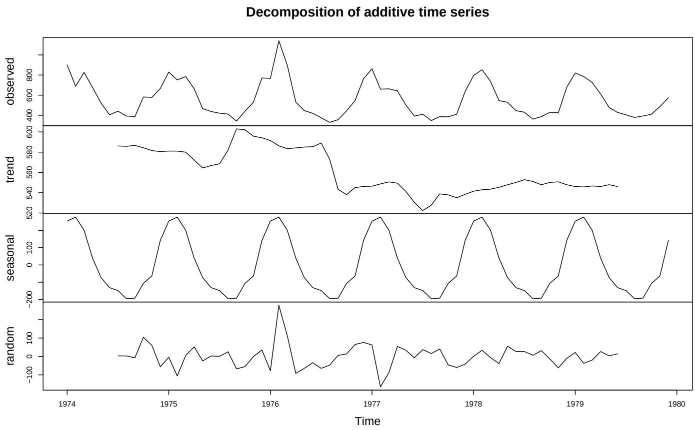
- **Multiplikativní model**

$$Y_t = Tr_t \times C_t \times S_t \times \varepsilon_t$$

kde

- Y_t – časová řada
- Tr_t – dlouhodobý/časový trend
- C_t – cyklická složka – dlouhodobé cykly, které většinou neumíme odhadnout, proto se tato složka odhaduje společně s trendem
- S_t – sezónní složka
- ε_t – náhodná složka

Dekompozice časových řad

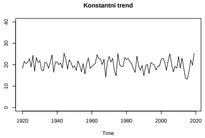


Časový trend

Ukázka základních časových trendů

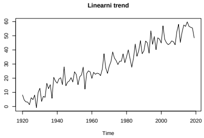
- **Konstantní trend**

$$Tr_t = \beta_0$$



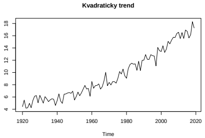
Lineární trend

$$Tr_t = \beta_0 + \beta_1 t$$



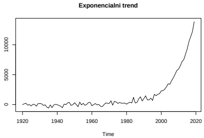
- **Kvadratický trend**

$$Tr_t = \beta_0 + \beta_1 t + \beta_2 t^2$$



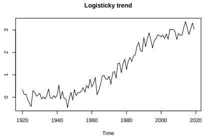
Exponenciální trend

$$Tr_t = \alpha \beta^t$$



Základní časové trendy

Logistický trend	Gompertzova křivka
$Tr_t = \frac{\gamma}{1+\alpha\beta^t}$	$Tr_t = \exp(\gamma + \alpha\beta^t)$
$\alpha > 0, 0 < \beta < 1, \gamma > 0$	$\alpha < -1, 0 < \beta < 1$



- Logistický trend je symetrický kolem inflexního bodu $-\log\alpha/\log\beta$
- Gompertzova křivka není symetrická kolem inflexního bodu $-\log(-\alpha)/\log\beta$

Informativní testy na typ trendu

Trend	test
Lineární	první difference $\delta y_t = y_{t+1} - y_t$ je konstrantní
Kvadratický	druhá difference $\delta^2 y_t = \delta y_{t+1} - \delta y_t$ je konstantní
Exponenciální	podíl y_{t+1}/y_t je konstantní
Logistický	podíl $(1/y_{t+2} - 1/y_{t+1})/(1/y_{t+1} - 1/y_t)$ je konstantní
Gompertz	podíl $(\log(y_{t+2}) - \log(t_{t+1}))/(\log(y_{t+1}) - log(y_t))$ je konstantní

Informační kritéria

Hodnocení kvality modelu Informační kritéria jsou založena na

- věrohodnosti modelu (L - *likelihood*)
  ukazatel, jak dobře model kopíruje data
- počtu parametrů použitých v modelu (k)

Věrohodnost se penalizuje složitostí modelu.

- Chceme model, který dobře kopíruje data a je co možná nejjednodušší
- Čím menší hodnota kritéria, tím lepší je model

Informační kritéria

Hodnocení kvality modelu

- **Akaikeho informační kritérium (AIC)**:

$$AIC = 2k - 2\ln(L)$$

- **Upravené Akaikeho informační kritérium (AICc)** pro malé vzorky:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

- **Bayesovské informační kritérium (BIC)**:

$$BIC = \ln(n)k - 2\ln(L)$$

Metoda klouzavých průměrů

Předpokládejme řadu, která je tvořena pouze trendovou a náhodnou složkou

$$Y_t = Tr_t + \varepsilon_t$$

- trend mění v čase svůj charakter
- nelze v celé délce popsat jedním funkčním předpisem
- funkční předpis lze aplikovat pouze lokálně, v okolí nějakého bodu
- př. celá řada se neřídí lineárním trendem, ale její trend je lineární v určitých časových intervalech

V takové případě se využívá tzv. **adaptivní přístup** ke trendové složce, kdy se parametry prokládaného modelu mění v čase.

Metoda klouzavých průměrů

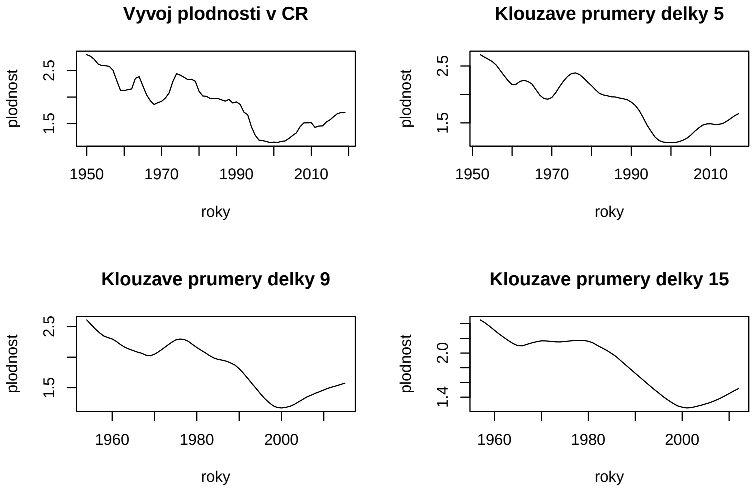
Nejjednodušší způsob, jak najít takový trend, tedy vyhladit časovou řadu, je použít klouzavé průměry. Vyhlazená hodnota v čase t se pak vypočte jako

$$\hat{Y}_t = \frac{\sum_{\tau=-m}^{m} Y_{t+\tau}}{2m+1}$$

- vyhlazuji pomocí hodnot na obě strany od času t
- klouzavé průměry délky $2m+1$ – liché číslo
- délka klouzavých průměrů závisí na požadované míře vyhlazení
- větší délka dá hladší křivku
- jednoduché klouzavé průměry uvařují lokálně lineární trend

Metoda klouzavých průměrů

Různé délky klouzavých průměrů



Metoda klouzavých vážených průměrů

Někdy nemusí vyhlazení lokálně lineárním trendem dostačovat a je třeba volit lokálně polynomiální trend vyššího řádu. Výsledkem jsou pak klouzavé vážené průměry.
Jako příklad uvažujme vyhlazení lokálně kubickým trendem

$$Tr(t)_\tau = \beta_0(t) + \beta_1(t)\tau + \beta_2(t)\tau^2 + \beta_3(t)\tau^3$$

kde $\tau = -m, \ldots, 0, \ldots m$ podle uvažované délky klouzavých průměrů.

Metoda klouzavých vážených průměrů

Koeficienty $\beta_0, \beta_1, \beta_2, \beta_3$ se hledají obdobně jako u lineárního trendu, tedy pomocí metody nejmenších čtverců.
Minimalizujeme

$$\sum_{\tau=-m}^{m} (Y_{t+\tau} - \beta_0 - \beta_1\tau - \beta_2\tau^2 - \beta_3\tau^3)^2$$

Parciálními derivacemi podle jednotlivých koeficientů získáme soustavu čtyř rovnic

$$\sum_{\tau=-m}^{m} Y_{t+\tau}\tau^j - b_0 \sum_{\tau=-m}^{m} \tau^j - b_1 \sum_{\tau=-m}^{m} \tau^{j+1} - b_2 \sum_{\tau=-m}^{m} \tau^{j+2} - b_3 \sum_{\tau=-m}^{m} \tau^{j+3} = 0$$

pro $j = 0, 1, 2, 3$. Hodnotu pro vyhlazení řady v čase t, tedy $\hat{Y}_t$, získáme jako odhad koeficientu b_0.

Vlastnosti odhadu $\hat{Y}_t$

- vážený klouzavý průměr hodnot $Y_{t+\tau}$, kde $\tau = -m, ldots, 0, \ldots, m$
- součet vah klouzavého průměru je roven jedné
- váhy jsdou symetrické kolem prostřední hodnoty
- je-li r sudé číslo, pak klouzavé průměry řádu r a $r+1$ se stejnou délkou jsou totožné
- takto vyhlazuji hodnoty řady od času m do $n = m$, kde n je délka řady (tedy prvních m a podleních m hodnot řady zůstává nevyhlazeno)

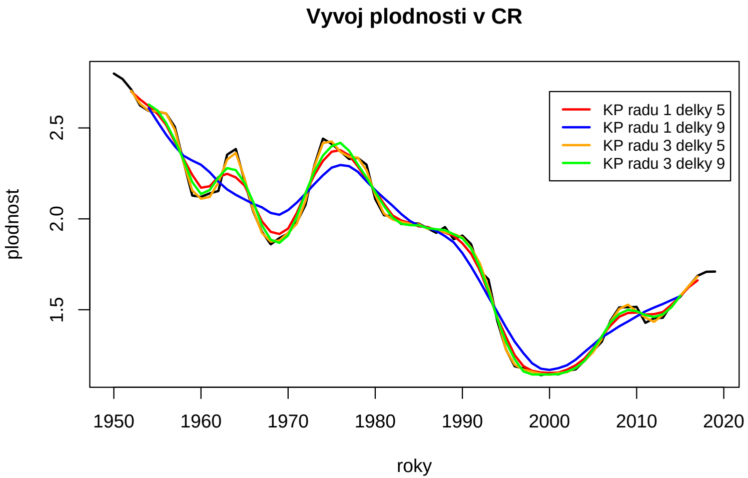
Váhy pro klouzavé průměry

Tabulka vah klouzavých průměrů pro různé délky a řády vyhlazení

Délka	Řád 2 a 3	Řád 4 a 5
3	(0,1,0)	(0,1,0)
5	$\frac{1}{35}(-3, 12, 17, \ldots)$	(0,0,1,. . .)
7	$\frac{1}{21}(-2, 3, 6, 7, \ldots)$	$\frac{1}{231}(5, -30, 75, 131, \ldots)$
9	$\frac{1}{231}(-21, 14, 39, 54, 59, , \ldots)$	$\frac{1}{429}(15, -55, 135, 179, \ldots)$
11	$\frac{1}{429}(-36, 9, 44, 69, 84, 89, \ldots)$	$\frac{1}{429}(18, -45, -10, 60, 120, 143, \ldots)$

Metoda klouzavých průměrů

Různé délky a řády klouzavých průměrů



Vyhlazení krajních hodnot

- "středové" hodnoty řady se vyhlazují pomocí sousedních hodnot na obě strany od daného času
- krajní hodnoty se vyhlazují pomocí krajních $2m + 1$ hodnot
- vyhlazuje se pomocí odhadovaného polynomiálního trendu
- jsou potřeba i ostatní koeficienty (nejen b_0, ale i např. $b1, b2, b3$)
- ve výsledku pak dostaneme jen jiné váhy krajních $2m + 1$ hodnot, které již nejsou symetrické
- obdobně lze počítat i predikce – každá další předpovídaná hodnota vyžaduje jiné váhy posledních $2m + 1$ hodnot řady

Volba řádu klouzavých průměrů

- Objektivní pravidlo je založené a diferencích
  - každá diference sníží řád původního polynomu o 1
  - pokud se r-tá diference jako první jeví konstantně, pak volím řád klouzavých průměrů r
- Subjektivní pravidlo
  - preference jednoduchého řešení, tj. klouzavé průměry co nejnižšího řádu
- Praktické pravidlo
  - v praxi se jiný než lineární řád nepoužívá
  - nejčastěji se používají nevážené klouzavé průměry nebo váhy, které odpovídají subjektivní důležitosti okolních hodnot
  - časté jsou např. váhy $\frac{1}{10}(1, 2, 4, 2, 1)$ nebo $\frac{1}{12}(1, 2, 2, 2, 2, 2, 1)$

Exponenciální vyrovnání

- Stejně jako metoda klouzavých průměrů, patří i exponenciální vyrovnání mezi adaptivní přístupy.
- Hlavní rozdíl oproti klouzavým průměrům je, že pro výpočet vyrovnané hodnoty v čase t nepoužívám hodnoty na obě strany od času t, ale pouze hodnoty předcházející času t.
- Používají se všechny minulé dostupné hodnoty s vahami exponenciálně klesajícími do minulosti.

Exponenciální vyrovnání

Předpokládejme, že časovou řadu tvoří pouze trend a náhodná složka

$$Y_t = Tr_t + \varepsilon_t$$

Mohu vyrovnávat

- lokálně konstantní trend $Tr_t = \beta_0$
  jednoduché exponenciální vyrovnávání
- lokálně lineární trend $Tr_t = \beta_0 + \beta_1 t$
  dvojité exponenciální vyrovnávání
- lokálně kvadratický trend $Tr_t = \beta_0 + \beta_1 t + \beta_2 t^2$
  trojité exponenciální vyrovnávání
- trendy vyššího řádu už se neuvažují

Exponenciální vyrovnání

Jednoduché exponenciální vyrovnávání

Odhad vyrovnávací hodnoty $b_0(t)$ získám minimalizací výrazu

$$\sum_{i=1}^{\infty}(Y_{t-i} - \beta_0)^2\alpha^i$$

kde α je vyrovnávací konstanta

- obecně musí platit $0 < \alpha < 1$
- praxe ukázala, že stačí uvažovat $0.7 < \alpha < 1$

Exponenciální vyrovnání

Jednoduché exponenciální vyrovnávání

Derivací výše uvedeného výrazu podle β_0 a položení této derivace rovno nule, získám odhad parametru $b_0(t)$ a tedy i vyrovnanou hodnotu v čase t.

$$\hat{Y}_t = (1-\alpha)\sum_{i=0}^{\infty} \alpha^i Y_{t-i}$$

Získám tedy vážený součet minulých hodnot s vahami

$$(1-\alpha), (1-\alpha)\alpha, (1-\alpha)\alpha^2, \ldots$$

exponenciálně klesajícími do minulosti.
Je možné psát i rekurentní vzorec

$$\hat{Y}_t = (1-\alpha)Y_t + \alpha^i \hat{Y}_{t-1}$$

Exponenciální vyrovnání

Jednoduché exponenciální vyrovnávání

Volba vyrovnávací konstanty α

- čím menší α, tím tím rychleji metoda reaguje na změnu v řadě
- větší α zesílí vyrovnávací schopnost metody
- znám-li optimální délku *M* klouzavého průměru pro danou řadu, mohu volit
  $\alpha = \frac{M-1}{M+1}$
- jinak (častěji) volím numericky: spočítám si vyrovnání řady pro růuzné hodnoty $\alpha = 0.7, 0.72, \ldots, 0.98$, pro každé ohodnotím, jak vyrovnání souhlasí s mými daty (např. pomocí *SSE – Sum of Squared Errors*) a volím α s nejnižším *SSE*
- softwary už volí automaticky

Exponenciální vyrovnání

Jednoduché exponenciální vyrovnávání

Předpovědi

- budoucí hodnoty předpovídám poslední vyrovnanou hodnotou $\hat{Y}_n$
- předpovědní interval spolehlivosti má tvar

$$\left(\hat{Y}_{n+\tau}(n) - zd_\tau\Delta(n),\ \hat{Y}_{n+\tau}(n) + zd_\tau\Delta(n)\right)$$

  kde z je příslušný kvantil standartního normálního rozdělení, $d_t = 1.25$ a

$$\Delta(n) = \sum_{t=1}^{n} \frac{|Y_t - \hat{Y}_t(t-1)|}{n}$$

- výhodou těchto předpovědí a předpovědních intervalů je jijich snadná úprava, pokud do řady přibude další pozorování

Exponenciální vyrovnání

Jednoduché exponenciální vyrovnávání

Adaptivní řídící proces

- je možné uvažovat i vyrovnávací konstantu α měnící se v čase
- definuje se indikáror poruchy

$$I(\alpha, t) = \left| \frac{Y(\alpha, t)}{D(\alpha, t)} \right|$$

  kde

$$Y(\alpha, t) = \sum_{i=1}^{t} e_i(\alpha), \qquad D(\alpha, t) = \sum_{i=1}^{t} \frac{|e_i(\alpha)|}{t}$$

  a e_i je chyba předpovědi
- zvedne-li se hodnota $I(\alpha, t)$ nad kontrolní mez K, pak je to signál ke změně α či dokonce celé metody

Exponenciální vyrovnání

Dvojité exponenciální vyrovnávání

Odhad vyrovnávací hodnoty $b_0(t) + b_1(t)t$ získám minimalizací výrazu

$$\sum_{i=1}^{\infty}(Y_{t-i} - (\beta_0 + \beta_1(-i)))^2\alpha^i$$

kde α je opět vyrovnávací konstanta a platí pro ni to, co u jednoduchého exponenciálního vyrovnání

- obecně musí platit $0 < \alpha < 1$
- praxe ukázala, že stačí uvažovat $0.7 < \alpha < 1$
- jen optimální volba při znalosti optimální délky klouzavého průměru M je $\alpha = \sqrt{\frac{M-1}{M+1}}$

Exponenciální vyrovnání

Trojité exponenciální vyrovnávání

V praxi se využívá minimálně, ale "ještě to jde". Odhad vyrovnávací hodnoty $b_0(t) + b_1(t)t + b_2(t)t^2$ získám minimalizací výrazu

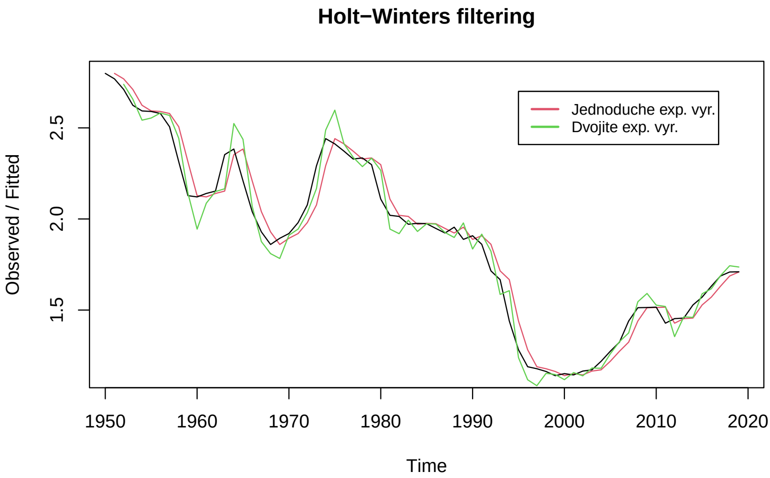
$$\sum_{i=1}^{\infty}(Y_{t-i} - (\beta_0 + \beta_1(-i) + \beta_2(-i)^2))^2\alpha^i$$

pro vyrovnávací konstantu α opět platí

- obecně musí platit $0 < \alpha < 1$
- praxe ukázala, že stačí uvažovat $0.7 < \alpha < 1$
- ale používá se trochu jiná počáteční volba hodnot při hledání optimálního α

Exponenciální vyrovnání

Grafická ukázka



Holt-Wintersova metoda

Máme-li sezónní data, je možné odhadovat jejich chování **Holt-Wintersovu metodu**

- zobecnění metody exponenciálního vyrovnání
- adaptivní metoda
- odhaduje sezónní složku, nevyhlazuje ji
- metoda je aplikovatelná jak na aditivní, tak na multiplikativní model

Předpokládejme lokálně lineární trend

$$Tr_t(t) = \beta_0(t) + \beta_1(t)t$$

a sezónní složku $s_t(t)$. Označme $a_0(t) = b_0(t) + b_1(t)$, kde $b_0(t)$ a $b_1(t)$ jsou odhady regresních koeficientů lokálně lineárního trendu a L počet sezón.

Holt-Wintersova metoda

Aditivní Holt-Wintersova metoda

Výše uvedené odhady se řídí následujícími vztahy

$$
\begin{aligned}
a_0(t) &= \alpha(Y_t - s_{t-L}(t-L)) + (1-\alpha)(a_0(t-1) + b_1(t-1)) \\
b_1(t) &= \beta(a_0(t) - a_0(t-1)) + (1-\beta)b_1(t-1) \\
s_t(t) &= \gamma(Y_t - a_0(t)) + (1-\gamma)s_{t-L}(t-L)
\end{aligned}
$$

Odhad hodnoty $Y_{t+\tau}$ pak můžeme počítat jako

$$\hat{Y}_{t+\tau}(t) = a_0(t) + b_1(t)\tau + s_{t+\tau-L}(t+\tau-L)$$

- za $s_{t+\tau-L}(t+\tau-L)$ používáme nejaktuálnější odhad sezónní složky, který máme k dispozici
- za počáteční hodnoty a_0, b_1 bereme odhady regresních koeficientů ze všech dat
- za počáteční odhad sezónní složky bereme průměr dané sezóny proti průměru všech dat

Holt-Wintersova metoda

Multiplikativní Holt-Wintersova metoda

Výše uvedené odhady se řídí následujícími vztahy

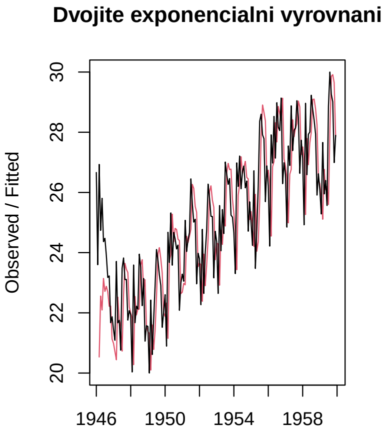
$$
\begin{aligned}
a_0(t) &= \alpha\left(\frac{Y_t}{s_{t-L}(t-L)}\right) + (1-\alpha)(a_0(t-1) + b_1(t-1)) \\
b_1(t) &= \beta(a_0(t) - a_0(t-1)) + (1-\beta)b_1(t-1) \\
s_t(t) &= \gamma\left(\frac{Y_t}{a_0(t)}\right) + (1-\gamma)s_{t-L}(t-L)
\end{aligned}
$$

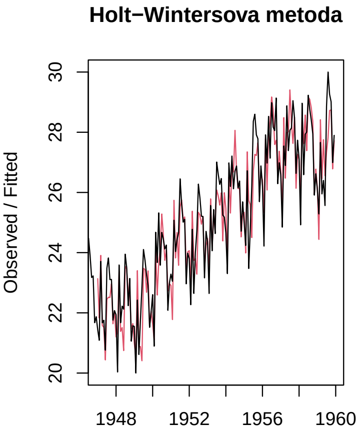
Odhad hodnoty $Y_{t+\tau}$ pak můžeme počítat jako

$$\hat{Y}_{t+\tau}(t) = (a_0(t) + b_1(t)\tau)s_{t+\tau-L}(t+\tau-L)$$

Holt-Wintersova metoda

Porovnání exponenciálního vyrovnání a Holt-Wintersovy metody





Náhodná složka

Předpokládáme, že náhodná složka je tvořena tzv. **bílým šumem.**

Definition

Nezávislé, stejně rozdělené náhodně veličiny s nulovou střední hodnotou a konstantním rozptylem.

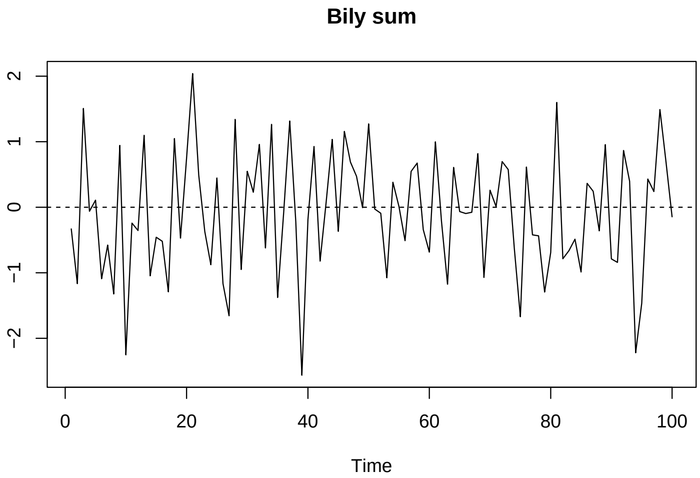
U náhodné složky je často třeba zjistit, zda je skutečně náhodná. Rozlišujeme dva případy

- máme časovou řadu, která na první pohled vypadá jako bílý šum
- máme časovou řadu, kde jsme odstínili trend a sezónní složku a o zbytku potřebujeme rozhodnout, zda tvoří bílý šum

V každém z případů se používají jiné testy náhodnosti.

Náhodná složka

Bílý šum



Náhodná složka

Testy náhodnosti pro řadu, co vypadá jako býlý šum

Testuje se H_0 : řada je náhodná vs. H_1 : řada není nahodná

- Test založený na **znaménkách diferencí**
  spočítám, v kolika bodech řada roste,
  je-li řada náhodná, mělo by to být cca v jedné polovině případů
- Test založený na **Kendallově koeficientu** τ
  Kendallův korelační koeficient pro řadu Y_t a indexy pozorování t

$$\tau = \frac{4v}{n(n-1)} - 1$$

kde v je počet případů, kde $Y_t > Y_s$ pro $t > s$

Náhodná složka

Testy náhodnosti pro řadu, co vypadá jako býlý šum

- Test založený na **Spearmanově koeficientu** ρ
  Spearmanův korelační koeficient pro řadu Y_t a indexy pozorování t

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{t=1}^{n} (t - r_t)^2$$

  kde r_t je pořadí t-té hodnoty Y_t
- **Wald-Wolfowitz Runs Test** (nebo též mediánový test) založen na hodnocení, zda je číslo nad(+)/pod(-) mediánem a počítá se, kolik úseků stejného znaménka v řadě máme

Náhodná složka

Testy náhodnosti pro residuální složku

Testuje se: H_0 : autokorelace je nulová vs. H_1 : autokorelace není nulová

- **Ljung-Box** neboli **Box-Pierce** test náhodnosti
  testuje velikost autokorelace v určených bodech

$$Q = n(n+2)\sum_{\tau=1}^{h} \frac{r_\tau^2}{n-\tau}$$

  kde r_τ je odhad autokorelace v čase τ (tedy korelace hodnot posunutých o τ kroků)

-

Box-Jenkinsonova metodologie

Pro vysvětlení Box-Jenkinsova metodologie analýzy časových řad definujme následující pojmy

- **stacionarita** – znamená *stálost* časové řady
- striktní stacionarita – všechny členy řady mají stejné rozdělení
- slabá stacionarita – všechny členy řady mají stejnou (předpokládejme že nulovou) střední hodnotu a stejný rozptyl σ_y^2

Základy modely budeme definovat pro stacionární (stačí slabě) časové řady.

Box-Jenkinsonova metodologie

Analýza těchto řad je založena na vzájemných vztazích mezi členy řady

- autokovarianční funkce – je to funkce vzdálenosti mezi dvěma členy řady

$$\gamma_k = \mathrm{cov}(\mathrm{Y_t}, \mathrm{Y_{t+k}}) = \mathrm{E}(\mathrm{Y_t} - \mu)(\mathrm{Y_{t+k}} - \mu)$$

  kde μ je střední hodnota členů řady

- **autokorelační funkce**

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma_y^2}$$

Box-Jenkinsonova metodologie

- parciální autokorelační funkce – jedná se o parciální korelační koeficient řad Y_t a Y_{t+k} při pevných hodnotách $Y_{t+1}, Y_{t+2}, \ldots, Y_{t+k-1}$

$$\rho_{kk} = \frac{|\mathbf{P}_k^*|}{|\mathbf{P}_k|}$$

kde

$$\mathbf{P}_k = \begin{bmatrix} 1 & \rho_1 & \ldots & \rho_{k-1} \\ \rho_1 & 1 & \ldots & \rho_{k-1} \\ \ldots & \ldots & \ldots & \ldots \\ \rho_{k-1} & \rho_{k-2} & \ldots & 1 \end{bmatrix} \quad \mathbf{P}_k^* = \begin{bmatrix} 1 & \rho_1 & \ldots & \rho_1 \\ \rho_1 & 1 & \ldots & \rho_2 \\ \ldots & \ldots & \ldots & \ldots \\ \rho_{k-1} & \rho_{k-2} & \ldots & \rho_k \end{bmatrix}$$

Box-Jenkinsova metodologie

Odhad autokorelační funkce a jeho vlastnosti.

- odhad střední hodnoty $\overline{Y} = \sum_{i=1}^{n} \frac{Y_t}{n}$
- odhad rozptylu řady $s_y^2 = c_0 = \sum_{t=1}^{n} \frac{(Y_t - \overline{Y})^2}{n}$
- odhad autokovarianční funkce $c_k = \sum_{t=1}^{n-k} \frac{(Y_t - \overline{Y})(Y_{t+k} - \overline{Y})}{n}$
- odhad autokorelační funkce $r_k = \frac{c_k}{c_0}$

Tento odhad kovarianční funkce a tedy ani autokorelační funkce není nestraný (to by se muselo dělit $n - k$), ale má malou střední chybu ($\mathrm{E}(c_k - \gamma_k)^2$). Výše uvedený odhad je asymptoticky nestranný.

Box-Jenkinsova metodologie

Někdy je třeba otestovat, od kterého bodu se autokorelační funkce dá považovat za nevýznamnou. K tomu se používá Bartletova aproximace

$$\sigma(r_k) = \sqrt{\mathrm{Var}(\mathrm{r_k})} \sim \sqrt{\frac{1}{n}\left(1 + 2\sum_{j=1}^{k_0} r_j^2\right)}$$

pro$k > k_0$. Pro zjednodušení se porovnává odhadovaná hodnota autokorelační funkce $|r_k|$ s číslem 2σ.

Box-Jenkinsova metodologie

Parciální autokorelační funkce se odhaduje rekurentně na základě následujících pravidel

- $r_{11} = r_1$
- $r_{kk} = \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_j}$
- kde $r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j}$

Test o nulovosti parciální autokorelační funkce je založen na Quenouilleově aproximaci

$$\sigma(r_{kk}) \sim \sqrt{\frac{1}{n}}$$

pro $k > k_0$. Pro jednoduchost se opět odhad parciální autokorelační funkce porovnává s hodnotou $2\sigma(r_{kk})$

Box-Jenkinsova metodologie

Lineární proces

Lineárním procesem nazýváme řadu tvaru

$$Y_t = \varepsilon_t + \psi_i \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \dots$$

kde ε_t je bílý šum s rozptylem σ_ε^2 a ψ_j jsou parametry.
Zaveďme operátor zpětného posunutí B následovně

$$BY_t = Y_{t-1}, \qquad B^j Y_t = Y_{t-j}$$

Jeho pomocí je možné zapsat lineární proces jako

$$Y_t = \psi(B)\varepsilon_t$$

kde

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \dots = 1 + \sum_{j=1}^{\infty} \psi_j B^j$$

Box-Jenkinsova metodologie

Pomocí operátoru zpětného posunu nastavíme podmínky na časové řady, aby byly popsatelné modely Box-Jenkinsovy metodologie.
Řada $\psi(B)$ konverguje pro $|B| \leq 1$, uvažujeme-li B jako číselnou proměnnou. Tato podmínka zároveň zaručí, že lineární proces je stacionární se střední hodnotou nula. Za určitých podmínek lze proces napsat jako

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \cdots + \varepsilon_t$$

Pak se lineární proces nazývá invertibilní.

Box-Jenkinsova metodologie

Proces klouzavých součtů (Moving Averages).

Proces klouzavých součtů řádu q se značí $MA(q)$ a má tvar

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

který se možné zapsat jako $Y_t = \theta(B)\varepsilon_t$, kde $\theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j$ je tzv.operátor klouzavých součtů.
Proces $MA(q)$ je

- stacionární pro libovolné θ_j
- má nulovou střední hodnotu
- jeho rozptyl je $\sigma_y^2 = \gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma_\varepsilon^2$

Box-Jenkinsova metodologie

Proces klouzavých součtů (Moving Averages).

- jeho autokorelační funkce je

$$\rho_k = \frac{\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2}$$

  pro $k = 1, \ldots, q$ a je $\rho_k = 0$ pro $k > q$
- autokorelační funkce má bod useknutí $k_0 = q$
- parciální autokorelační funkce nemá bod useknutí a je omezena geometricky klesající posloupností nebo sinusoidou s geometricky klesající amplitudou
- proces je invertibilní, pokud všechny kořeny polynomu $\theta(B)$ leží vně jednotkového kruhu

Box-Jenkinsova metodologie

Proces MA(1)

Nejjednodušším procesem klouzavých součtů je proces $MA(1)$.

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

s vlastnostmi

- proces $MA(1)$ je invertibilní, pokud $|\theta_1| < 1$
  pak lze tento proces napsat jako
  $Y_t = \theta_1 Y_{t-1} - \theta_1^2 Y_{t-2} - \theta_1^3 Y_{t-3} \cdots + \varepsilon_t$
- autokorelační funkce

$$\rho_1 = \frac{\theta_1}{1 + \theta_1^2}$$

  a $\rho_k = 0$ pro $k > 1$
- pro libovolný proces $MA(1)$ musí být $|\rho_1| < \frac{1}{2}$

Box-Jenkinsova metodologie

Proces MA(1)

- parametr θ_1 lze spočítat z autokorelace ρ_1

$$\theta_1 = \frac{1 - \sqrt{1 - 4\rho_1^2}}{2\rho_1}$$

  a pouze jeden z kořenů splňuje podmínku invertibility
- parciální autokorelační funkce je

$$\rho_{kk} = \frac{(-1)^{k-1}\theta_1^k(1 - \theta_1^2)}{1 - \theta_1^{2(k+1)}}$$

- platí $|\rho_{kk}| < |\theta_1|^k$

Box-Jenkinsova metodologie

Autoregresní proces (Autoregressive process)

Autoregresní proces řádu p se značí $AR(p)$ a má tvar

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t$$

tento proces můžeme pomocí zpětného operátoru zapsat jako

$$\varphi(B) Y_t = \varepsilon_t$$

kde $\varphi(B) = 1 - \sum_{j=1}^{p} \varphi_j B^j$ je tzv. autoregresní operátor.

- proces $AR(p)$ je automaticky invertibilní
- proces $AR(p)$ je stacionární, pokud všechny kořeny polynomu $\varphi(B)$ leží vně jednotkového kruhu
- střední hodnota $AR(p)$ je nulová

Box-Jenkinsova metodologie

Autoregresní proces (Autoregressive process)

- pro jeho autokorelační funkci platí
$$\rho_k = \varphi_1 \rho_{k-1} + \varphi_2 \rho_{k-2} + \cdots + \varphi_p \rho_{k-p}$$
pro $k > 0$.
- autokorelační funkce je lineární kombinací klesajících geometrických posloupností a sinusoid s geometricky klesajícími amplitudami
- parametry $\varphi_1, \ldots, \varphi_p$ je možné z autokorelační funkce dopočítat pomocí Yule-Walkerových rovnic
- rozptyl
$$\sigma_y^2 = \gamma_0 = \frac{\sigma_\varepsilon^2}{1 - \varphi_1 \rho_1 - \ldots - \varphi_p \rho_p}$$
- parciální autokorelační funkce má bod useknutí $k_0 = p$ a je tedy $\rho_{kk} = 0$ pro $k > p$.

Box-Jenkinsova metodologie

Proces AR(1)

Nejjednodušším autoregresním procesem je $AR(1)$ tvaru

$$Y_t = \varphi_1 Y_{t-1} + \varepsilon_t$$

s vlastnostmi

- $AR(1)$ je stacionární, pokud $|\varphi_1| < 1$
- autokorelační funkce procesu $AR(1)$ je

  $$\rho_k = \varphi_1^k$$

  pro $k > 0$ a je to tedy geometrická posloupnost klesající k nule
- pro $k = 1$ máme $\rho_1 = \varphi_1$
- parciální autokorelační funkce procesu $AR(1)$ je

  $$\rho_{11} = \rho_1 = \varphi_1, \qquad \rho_{kk} = 0$$

  pro $k > 1$, tedy parciální autokorelační funkce má bod useknutí $k_0 = 1$

Box-Jenkinsova metodologie

Smíšený proces

Smíšený proces z částí *MA* řádu q a *AR* řádu p nazýváme *ARMA* řádu p a q a má tvar

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Pomocí operátoru zpětného posunutí je zápis jednodušší

$$\varphi(B) Y_t = \theta(B) \varepsilon_t$$

Vlastnosti procesu $ARMA(p, q)$

- podmínka stacionarity je stejná jako u $AR(p)$
- podmínka invertibility je stejná jako u $MA(q)$
- kořeny polynomů $\theta(B)$ a $\varphi(B)$ musí ležet vně jednotkového kruhu

Box-Jenkinsova metodologie

Smíšený proces

- střední hodnota $ARMA(p, q)$ je nulová
- pro autokorelační funkci ρ_k platí

$$\rho_k = \varphi_1 \rho_{k-1} + \varphi_2 \rho_{k-2} + \cdots + \varphi_p \rho_{k-p}$$

  pro $k > q$
- je-li $q \geq p$ pak od času $q - p + 1$ je autokorelační funkce tvořena geometricky klesající posloupností, nebo kombinací sinusoid s geometricky klesajícími amplitudami
- parciální autokorelační funkce ρ_{kk} je obdobně od času $p - q + 1$ omezena geometricky klesající posloupnoustí nebo sinusoidou s geometricky klesající amplitudou

Box-Jenkinsova metodologie

Proces ARMA(1,1)

Proces $ARMA(1,1)$ má tvar

$$Y_t = \varphi_1 Y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

s vlastnostmi

- podmínka stacionarity $|\varphi_1| < 1$
- podmínka invertibility $|\theta_1| < 1$
- rozptyl řady je

$$\sigma_y^2 = \gamma_0 = \frac{1 + \theta_1^2 + 2\varphi_1\theta_1}{1 - \varphi_1^2}\sigma_\varepsilon^2$$

- pro autokorelační funkci platí $\rho_k = \varphi_1 \rho_{k-1}$ pro $k > 1$
- autokorelační funkce v čase 1 je

$$\rho_1 = \frac{(1 + \varphi_1\theta_1)(\varphi_1 + \theta_1)}{1 + \theta_1^2 + 2\varphi_1\theta_1} 1 - \varphi_1^2$$

Box-Jenkinsova metodologie

Proces ARMA(1,1)

- parametry pak můžeme odhadnout pomocí

$$\varphi_1 = \frac{\rho_2}{\rho_1}, \qquad \theta_1 = \frac{b \pm \sqrt{b^2 - 4}}{2}$$

kde

$$b = \frac{1 - 2\rho_2 + \varphi_1^2}{\rho_1 - \varphi_1}$$

a jen jeden kořen odpovídá podmínce invertibility

- pro autokorelační funkci musí dále platit

$$2\rho_1^2 - |\rho_1| < \rho_2 < |\rho_1|$$

- parciální autokorelační funkce je omezena geometricky klesající posloupností

Box-Jenkinsova metodologie

Jak najít optimální model pro stacionární řadu

první nápovědu dají autokorelační a parciální autokorelační funkce

	ρ_k	ρ_{kk}
$AR(p)$	neexistuje bod useknutí	bod useknutí $k_0 = p$
$MA(q)$	bod useknutí $k_0 = q$	neexistuje bod useknutí
$ARMA(p, q)$	neexistuje bod useknutí po $q - p$ hodnotách omezena	neexistuje bod useknutí po $p - q$ hodnotách omezena

Box-Jenkinsova metodologie

Odhad parametrů modelu

- používá se metoda nejmenších čtverců (minimalituje se součet $\sum_{t=1}^{n}(\varepsilon_t(\varphi,\theta))^2$)
- jelikož funkce ε_t je v parametrech φ a θ nelineární, jedná se o nelineární součet čtverců a je třeba postupovat iterativně
- počítač necháme odhadnout několik základních modelů (většinou některé z $AR(1), AR(2), AR(3), MA(1), MA(2), MA(3), ARMA(1,1), ARMA(2,1), ARMA(1,2), ARMA(2,2)$)
- pomocí určitého kritéria zvolíme optimální model
- porovnáme volbu s průběhem autokorelační a parciální autokorelační funkce

Box-Jenkinsova metodologie

Kritéria pro výběr modelu

- Akaikeho informační kritérium

$$AIC = \ln \hat{\sigma}_{\varepsilon}^2 + 2\frac{k}{n}$$

kde k je počet parametrů modelu
nejznámější kritérium, které však upřednostňuje modely s vyšším počtem parametrů

- **Bayesovské informační kritérium**

$$BIC = \ln \hat{\sigma}_{\varepsilon}^2 + \frac{k \ln n}{n}$$

toto kritérium je nejvhodnější pro optimální volbu modelu

- Hannan-Quinn informační kritérium

$$HQ = \ln \hat{\sigma}_{\varepsilon}^2 + c\frac{k \ln \ln n}{n}$$

nejčastěji se volí $c = 1$
kritérium vymyšleno přímo pro B-J modely

Čím menší hodnota kritéria, tím lepší model.

Box-Jenkinsova metodologie

Ověření vybraného modelu

- střední chyba odhadnutých parametrů by měla být víc jak dvakrát menší než absolutní hodnota samotného odhadu
- odhadnutý rozptyl bílého šumu by neměl být příliš velký
- residua odhadnutého modelu by měla být nezávislá, jejich autokorelační funkce by tedy měla nabývat malých hodnot např. pro $AR(1)$ by $r_1(\varepsilon)$ měla být menší než $2\varphi_1^2/n$
- máme i test na hodnoty autokorelační funkce residuí (Ljung-Box test)

Box-Jenkinsova metodologie

Integrované ARMA modely

- tyto modely nevyžadují stacionaritu - umí si poradit s trendem
- odstínění trendu se zde realizuje přes diference
- počet nutných diferencí před odhadem ARMA modelu odpovídá komplikovanosti trendu
- většinou stačí diferencovat jednou, maximálně dvakrát

Integrovaný ARMA model se značí $ARIMA(p, d, q)$, kde

- p je řád autoregrese
- q je řád klouzavých průměrů
- d je nutný počet diferencí potřebný se stacionarizaci původní řady

Box-Jenkinsova metodologie

Model ARIMA(p,d,q)

Model zapisujeme jako

$$\varphi(B)w_t = \theta(B)\varepsilon_t$$

kde $w_t = \Delta^d Y_t$ je d-tá diference modelu a platí

$$\begin{aligned} \Delta Y_t &= Y_t - Y_{t-1} = (1-B)Y_t \\ \Delta^2 Y_t &= (1-B)^2 Y_t = (1-2B+B^2)Y_t = Y_t - 2Y_{t-1} + Y_{t-2} \end{aligned}$$

Model $ARIMA(p,d,q)$ pak můžeme zapsat jako

$$\varphi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t$$

Nevýhodou je, že řada diferencí má o d prvků méně než původní řada, kde d je řád diference

Box-Jenkinsova metodologie

Určení řádu diference pro procesy ARIMA

- využití "okometrické metody"
  diferencuji tak dlouho, až mi výsledná řada diferencí přijde jako stacionární
- s pomocí autokorelační funkce
  jestliže hodnoty autokorelační funkce diferencí klesají pomalu (přibližně lineárně a ne geometricky), diferencuji dál
- pomocí odhadu rozptylu řady diferencí
  vyberu tu diferenci, která má minimální odhad rozptylu řady
- pomocí počítače
  použiji nejnižší diferenci, pro niž mi počítač umí ARIMA model odhadnout

Box-Jenkinsova metodologie

Sezónní integrovaný ARMA model

I se sezónními řadami si umí BJ metodologie poradit. Řešením jsou tzv. **SARIMA modely**.

Myšlenka konstrukce těchto modelů

- pro každou sezónu (např. leden) se zkonstruuje samostatný model $ARIMA(P, D, Q)$
- modely pro jednotlivé sezóny by měly být přibližně stejné
- náhodné složky v těchto modelech η_t by měly být vzájemně korelované
- pro tyto náhodné složky zkonstruujeme další model $ARIMA(p, d, q)$
- spojením těchto modelů dostaneme

$$\varphi(B)\Phi(B^s)\Delta^d\Delta_s^D Y_t = \theta(B)\Theta(B^s)\varepsilon_t$$

kde s je počet sezón (pro měsíční data 12, pro čtvrtletní 4, atd.). Výsledkem je tzv. multiplikativní sezónní model řadu

Box-Jenkinsova metodologie

Identifikace modelu

- stěžejní je určení řádu diferencí
- diferencuje se přes délku sezóny i přes sousední hodnoty
- určení řádu difference záleží na autokorelační funkci
- pokud má autokorelační funkce vysokou hodnotu v délce sezóny, je nutné diferencovat přes sezónu
- pokud má autokorelační funkce pomalu klesající hodnoty, je třeba diferencovat přes sousední hodnoty
- je možné volit optimální model i podle minimalizace rozptylu řady

Box-Jenkinsova metodologie

Předpovědi

- předpovědi konstruujeme rekurentně
- namísto aktuálních hodnot bílého šumu bereme nulu, za starší hodnoty bílého šumu bereme chyby předchozích předpovědí
- předpovědní interval spolehlivosti

$$\hat{Y}_{t+k}(t) \pm 2\sigma\{e_{t+k}(t)\}$$

kde $e_t(t-1) = Y_t - \hat{Y}_t(t-1)$.

Závislost časových řad

Nejjednodušším způsobem, jak zkoumat závislost mezi dvěma časovými řadami je použít **kroskorelační funkci** definovanou jako

$$\rho_k^c = \frac{\mathrm{cov}(\mathrm{Y_t}, \mathrm{X_{t+k}})}{\sigma_y \sigma_x}$$

- v porovnání s běžným korelačním koeficientem uvažujeme i korelaci se zpožděním
- odhad této funkce je

$$r_k^c = \frac{\sum_{t=1}^{n-k}(Y_t - \overline{Y})(X_{t+k} - \overline{X})}{\sqrt{\sum_{t=1}^{n}(Y_t - \overline{Y})^2}\sqrt{\sum_{t=1}^{n}(X_t - \overline{X})^2}}$$

- jako hrubý odhad střední chyby odhadu r_k^c můžeme uvažovat (obdobně jako u autokorelační funkce)

$$\sigma({}^c r_k) = \sqrt{\mathrm{Var}(\mathrm{r_k^c})} \sim \sqrt{\frac{1}{n}\left(1 + 2\sum^{k_0}(r_j^c)^2\right)}$$

Závislost časových řad

Je možné uvažovat i regresní závislost mezi časovými řadami. Jedná se o tzv. **lineární dynamické modely**

- do těchto modelů vstupují jako regresory jak předcházející hodnoty vlastní řady, tak současné, případně zpožděné hodnoty řad jiných
- optimální zpoždění, s jakým budeme modelovat závislost řady Y_t na řadě X_t odhadneme např. z kroskorelační funkce
- modely se většinou odhadují pomocí běžné metody nejmenších čtverců
- správný model je ten, jehož residua mají nulovou autokorelační funkci pro $k > 0$