Časové řady, typy trendových funkcí a odhady trendů

Jiří Neubauer

Katedra kvantitativních metod FVL UO Brno
kancelář 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Stochastický proces

Posloupnost náhodných veličin $\{Y_t, t = 0, \pm 1, \pm 2 \dots\}$ se nazývá **stochastický proces**. Pomocí něho budeme modelovat pozorované časové řady.
Střední hodnota stochastického procesu $\{Y_t\}$ je funkce μ_t daná vztahem

$$\mu_t = E(Y_t), \quad t = 0, \pm 1, \pm 2 \dots .$$

Autokovarianční funkce je definována jako

$$\gamma_{t,s} = C(Y_t, Y_s), \quad t, s = 0, \pm 1, \pm 2 \dots ,$$

kde $C(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E[Y_t, Y_s] - \mu_t \mu_s$. **Autokorelační funkce** je dána vztahem

$$\rho_{t,s} = \frac{C(Y_t, Y_s)}{\sqrt{D(Y_t) D(Y_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}.$$

Stacionarita

Jedním z důležitých vlastností stochastických procesů je **stacionarita**, což znamená, že pravděpodobnostní rozdělení, které řídí chování stochastického procesu je v čase neměnné, proces je ve „statistickém ekvilibriu“.

O procesu $\{Y_t\}$ řekneme, že je **striktně stacionární**, jestliže simultánní rozdělení $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ je stejné jako simultánní rozdělení $Y_{t_1-k}, Y_{t_2-k}, \ldots, Y_{t_n-k}$ pro všechna t a všechna možná zpoždění k.

Jestliže funkce $\gamma_{s,t}$ závisí na svých argumentech pouze prostřednictvím jejich rozdílů $k = s - t$, pak říkáme, že proces je **kovariančně stacionární**. Autokovarianční funkcí takového procesu budeme rozumět funkci jedné proměnné $\gamma_k = \gamma_{s-t} = \gamma_{s,t}$. Je-li navíc střední hodnota procesu μ_t konstantní pro všechna t ($\mu_t = \mu$), proces $\{Y_t\}$ označujeme za **slabě stacionární**. V dalším budeme místo slabě stacionární proces psát jen krátce proces stacionární.

Stacionarita – autokovarianční a autokorelační funkce

Autokovarianční funkce γ_k stacionárního stochastického procesu je definována jako

$$\gamma_k = C(Y_t, Y_{t-k}) = E[(Y_t - \mu)(Y_{t-k} - \mu)],$$

a **autokorelační funkce (ACF)** ρ_k je dána vztahem

$$\rho_k = \frac{C(Y_t, Y_{t-k})}{\sqrt{D(Y_t)D(Y_{t-k})}} = \frac{\gamma_k}{\gamma_0}.$$

Parciální autokorelační funkce

Korelace mezi dvěma náhodnými veličinami je často způsobena tím, že obě veličiny jsou korelovány s veličinou třetí. Parciální autokorelace podávají informaci o korelaci veličin Y_t a Y_{t-k} očištěnou o vliv veličin ležících mezi nimi.
Parciální autokorelaci se zpožděním k stacionárního procesu $\{Y_t\}$ vyjadřuje parciální regresní koeficient ϕ_{kk} v autoregresi k-tého řádu

$$Y_t = \phi_{k1} Y_{t-1} + \phi_{k2} Y_{t-2} + \cdots + \phi_{kk} Y_{t-k} + \epsilon_t,$$

kde e_t je veličina nekorelovaná s $Y_{t-j}, j \geq 1$. Je to funkce zpoždění k a nazývá se **parciální autokorelační funkce (PACF)** ρ_{kk}.

Parciální autokorelační funkce

Předpokládejme, že stacionární proces $\{Y_t\}$ má nulovou střední hodnotu. Po vynásobení obou stran předchozí rovnice veličinou Y_{t-j} má střední hodnota této rovnice tvar

$$\gamma_j = \phi_{k1}\gamma_{j-1} + \phi_{k2}\gamma_{j-2} + \cdots + \phi_{kk}\gamma_{j-k},$$

takže platí

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \cdots + \phi_{kk}\rho_{j-k}.$$

Pro $j = 1, 2, \ldots, k$ potom dostáváme

$$\begin{aligned}
\rho_1 &= \phi_{k1}\rho_0 + \phi_{k2}\rho_1 + \cdots + \phi_{kk}\rho_{k-1} \\
\rho_2 &= \phi_{k1}\rho_1 + \phi_{k2}\rho_0 + \cdots + \phi_{kk}\rho_{k-2} \\
&\cdots \\
\rho_k &= \phi_{k1}\rho_{k-1} + \phi_{k2}\rho_{k-2} + \cdots + \phi_{kk}\rho_0.
\end{aligned}$$

Tyto rovnice se nazývají **Yule-Walkerovy rovnice.**

Parciální autokorelační funkce

Řešením této soustavy (Cramerovým pravidlem) pro $k = 1, 2, \ldots$ postupně dostáváme

$$\rho_{11} = \phi_{11} = \rho_1,$$

$$\rho_{22} = \phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2},$$

$$\rho_{kk} = \phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}}$$

Odhady

Obecně jsou parametry μ, γ_0 a ρ_k neznámé, za předpokladu stacionarity použijeme odhady

$$\widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t,\ \widehat{\gamma}_0 = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \overline{Y})^2.$$

kde n je počet hodnot (délka) časové řady. Odhad ρ_k je dán výběrovou autokorelací

$$\widehat{\rho}_k = \frac{\sum_{t=k+1}^{n}(Y_t - \overline{Y_t})(Y_{t-k} - \overline{Y_t})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2},\ k = 1, 2, \ldots, n-1.$$

(V programu R lze spočítat pomocí funkce `acf`)

Odhady

Výběrovou parciální korelační funkci získáme nahrazením ρ_i jejím odhadem $\hat{\rho}_i$ v odpovídajícím vzorci. Byl však odvozen rekurzivní vztah, který výpočet zjednoduší

$$
\begin{aligned}
\widehat{\rho}_{11} &= \widehat{\rho}_1 \\
\widehat{\rho}_{kk} &= \frac{\hat{\rho}_k - \sum_{j=1}^{k-1} \widehat{\rho}_{k-1,j}\, \widehat{\rho}_{k-j}}{1 - \sum_{j=1}^{k-1} \widehat{\rho}_{k-1,j}\, \widehat{\rho}_j}, \\
\widehat{\rho}_{kj} &= \widehat{\rho}_{k-1,j} - \widehat{\rho}_{kk} \widehat{\rho}_{k-1,k-j}, \quad j = 1, 2, \ldots, k-1.
\end{aligned}
$$

(V programu R lze spočítat pomocí funkce `pacf`)

Proces bílého šumu – white noise

Důležitý stacionárním stochastickým procesem je tzv. proces **bílého šumu**. Jedná se o posloupnost nekorelovaných náhodných veličin se stejným rozdělením s nulovou střední hodnotou a konstantním rozptylem. Pro bílý $\{\epsilon_t\}$ platí

$$\rho_k = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases}$$

$$\rho_{kk} = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases}$$

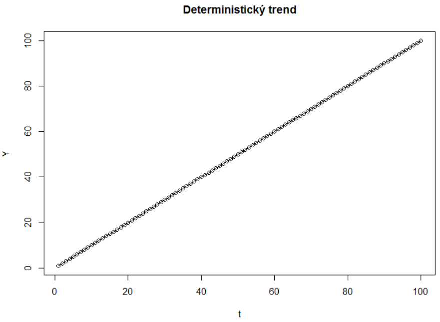
Gaussovský bílý šum – posloupnost nezávislých náhodných veličin s rozdělením $N(0, \sigma^2_{\epsilon_t})$.

Deterministický trend

Např. proces

$$Y_t = Y_0 + at, t = 1, \ldots n$$

obsahuje deterministický lineární trend. Y_0 označuje počáteční hodnotu.
Pro $n = 100, Y_0 = 0, a = 1$ proces zobrazený v grafu.



Stochastický trend

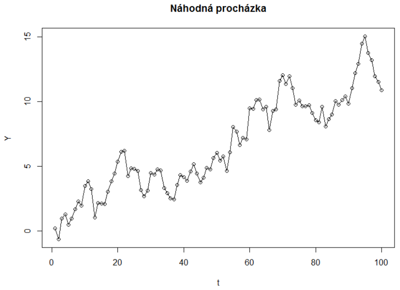
Např. proces („náhodná procházka“ nebo „random walk“)

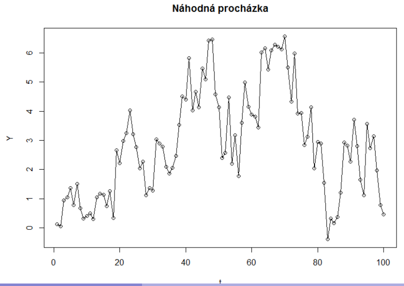
$$Y_t = Y_{t-1} + \epsilon_t, t = 1, \ldots n,$$

kde $\epsilon_t \sim \mathrm{WN}(0, \sigma^2)$ lze psát ve tvaru

$$\begin{aligned} Y_t &= Y_{t-1} + \epsilon_t = (Y_{t-2} + \epsilon_{t-1}) + \epsilon_t = \\ &= (Y_{t-3} + \epsilon_{t-2}) + \epsilon_{t-1} + \epsilon_t = \cdots = \\ &= Y_0 + \epsilon_1 + \cdots + \epsilon_t = Y_0 + \sum_{i=1}^{t} \epsilon_i \end{aligned}$$

Y_0 značí počáteční hodnotu. Dvě z možných realizací procesu (simulací) pro $n = 100, Y_0 = 0, \epsilon_t \sim \mathrm{WN}(0, 1)$ jsou zobrazeny v grafech.





Stochastický trend

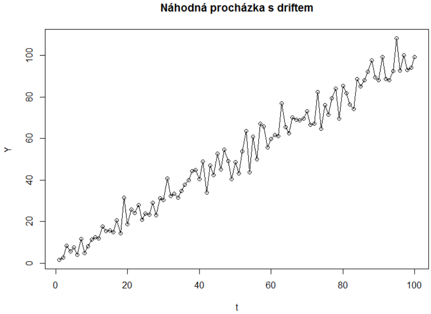
Např. proces („náhodná procházka“ s driftem)

$$Y_t = Y_{t-1} + a + \epsilon_t, t = 1, \dots n,$$

kde $\epsilon_t \sim \mathrm{WN}(0, \sigma^2)$ lze psát ve tvaru

$$\begin{aligned} Y_t &= Y_{t-1} + a + \epsilon_t = (Y_{t-2} + a + \epsilon_{t-1}) \\ &+ a + \epsilon_t = (Y_{t-3} + a + \epsilon_{t-2}) \\ &+ 2a + \epsilon_{t-1} + \epsilon_t = \dots = \\ &= Y_0 + at + \sum_{i=1}^{t} \epsilon_i \end{aligned}$$

Y_0 značí počáteční hodnotu. Jedna z možných realizací procesu (simulace) pro $n = 100, Y_0 = 0, \epsilon_t \sim \mathrm{WN}(0, 1)$ je zobrazena v grafu.



Dekompozice časových řad

Základem klasické analýzy časové řady Y_t je její rozklad na trend T_t, sezónní složku S_t a složku reziduální (zbytkovou, náhodnou) e_t. V **aditivním modelu** má dekompozice tvar

$$Y_t = T_t + S_t + \epsilon_t,$$

v **multiplikativním modelu** potom tvar

$$Y_t = T_t \cdot S_t \cdot \epsilon_t.$$

Obvyklou metodou, jak získat trend je využití lineárních filtrů

$$T_t = \sum_{i=-\infty}^{\infty} \lambda_i Y_{t+i}.$$

Klouzavé průměry

Jednoduchým příkladem lineárních filtrů jsou klouzavé průměry délky $2m+1$ s konstantními váhami

$$\hat{T}_t = \frac{1}{2m+1} \sum_{i=-m}^{m} Y_{t+i}.$$

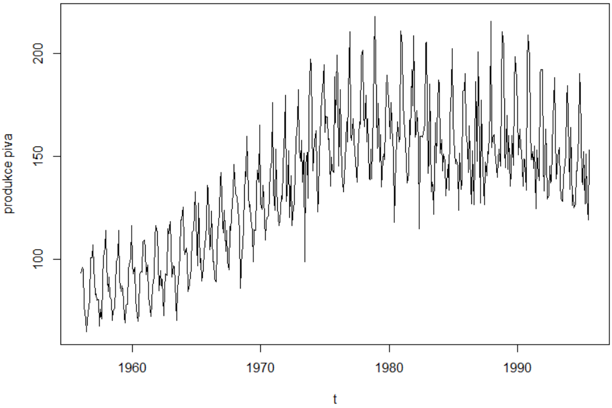
Vyrovnanou hodnotu časové řady v čase t získáme jako průměr hodnot $Y_{t-m}, \dots, Y_{t-1}, Y_t, Y_{t+1} \dots, Y_{t+m}$. Například pro $m = 1$ dostáváme klouzavý průměr délky 3

$$\hat{T}_t = \frac{1}{3}(Y_{t-1} + Y_t + Y_{t+1}).$$

V programu R lze klouzavé průměry určit pomocí funkce `filter` nebo funkce `ma` z balíčku `forecast`.

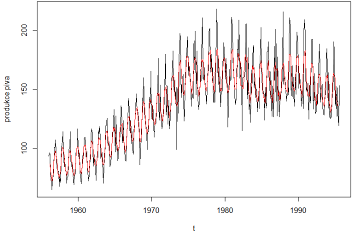
Klouzavé průměry

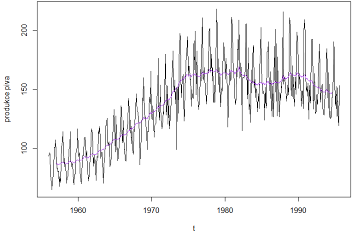
Graf zobrazuje obsahuje měsíční produkci piva v Austrálii od ledna 1956 do srpna 1995

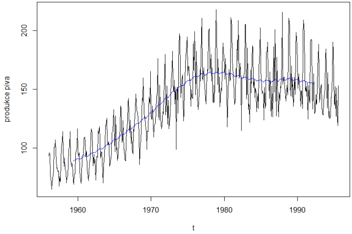


Klouzavé průměry

Grafy zobrazují klouzavé průměry délky 5 ($a = 2$), 25 ($a = 12$), 81 ($a = 20$).







Klouzavé průměry

Délka filtru ovlivňuje stupeň vyhlazení. Čím větší je délka klouzavého průměru, tím větší je vyhlazení časové řady. Délku klouzavého průměru obvykle volíme tak, aby odpovídala periodě sezónních nebo cyklických fluktuací. Při zpracování ekonomických časových řad se často zpracováváme čtvrtletní případně měsíční údaje, jež často obsahují sezónní složku opakující po sudém počtu pozorování (po čtyřech příp. dvanácti hodnotách). Pro tyto případy lze použít tzv. *centrované klouzavé průměry*. Pro případ čtvrtletních měření určíme vyrovnanou hodnotu ze vztahu

$$\hat{T}_t = \frac{1}{2}\left[\frac{1}{4}(Y_{t-2} + Y_{t-1} + Y_t + Y_{t+1}) + \frac{1}{4}(Y_{t-1} + Y_t + Y_{t+1} + Y_{t+2})\right] =$$
$$= \frac{1}{8}\left(Y_{t-2} + 2Y_{t-1} + 2Y_t + 2Y_{t+1} + Y_{t+2}\right).$$

Jedná je o klouzavý průměr délky 5 s váhami $\frac{1}{8}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}$. Analogicky pro měsíční použijeme klouzavé průměry délky 13 typu

$$\hat{T}_t = \frac{1}{24}\left(Y_{t-6} + 2Y_{t-5} + \cdots + 2Y_{t+5} + Y_{t+6}\right).$$

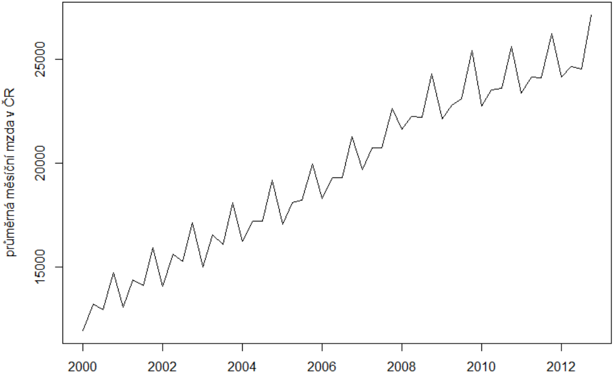
Dekompozice časových řad

Klouzavé průměry (v R je možné je počítat pomocí funkce `filter`) jsou základem klasické dekompozice, kterou v programu R provádí funkce `decompose`. Poněkud sofistikovanější metodu dekompozice nabízí funkce `stl`.

Dekompozici časové řady lze také provádět pomocí lineární regrese (funkce `lm` – viz regresní analýza). Mimo trendu (lineárního, kvadratického atd.) je často vhodné do regresního modelu přidat buď sezónní složky, nebo periodické funkce s vhodnými periodami.

Použití regresní analýzy

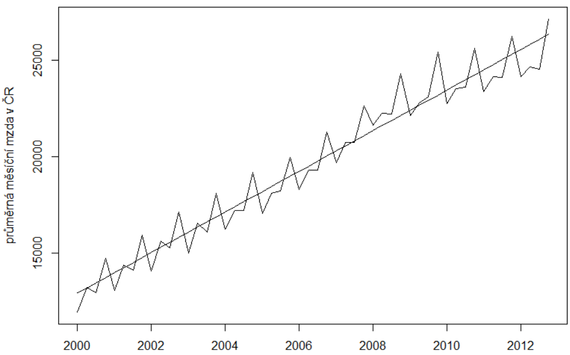
Na obrázku je znázorněn vývoj hrubé měsíční mzdy v ČR v období 2000–2012, jedná se o čtvrtletní data.



Použití regresní analýzy

Trend odhadneme pomocí přímkové regrese, pro numerickou stabilitu výpočtu provedeme transformaci časové proměnné $t = rok - 1999$, takže $t = 1 \ldots, 13$.

	Odhad	Sm. chyba	t-test	p-hodnota
konstanta	11875,9388	286,5841	41,44	0,0000
t	1051,1374	34,6343	30,35	0,0000



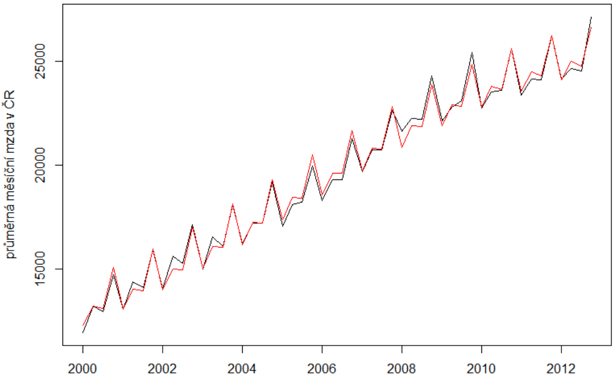
Použití regresní analýzy

Periodickou složku odhadneme pomocí „dummy“ proměnných q_1, q_2, q_3, q_4. Trend potom pomocí polynomu 3. stupně. konstantu do modelu nezahrneme, vznikne vlastně součtem „dummy“ proměnných q_1, q_2, q_3, q_4.

$q_1 = (1, 0, 0, 0, 1, 0, 0, 0, \ldots, 1, 0, 0, 0)'$
$q_2 = (0, 1, 0, 0, 0, 1, 0, 0, \ldots, 0, 1, 0, 0)'$
$q_3 = (0, 0, 1, 0, 0, 0, 1, 0, \ldots, 0, 0, 1, 0)'$
$q_4 = (0, 0, 0, 1, 0, 0, 0, 1, \ldots, 0, 0, 0, 1)'$

	Odhad	Sm. chyba	t-test	p-hodnota
t	484,4349	152,8450	3,17	0,0027
t^2	111,4610	23,3280	4,78	0,0000
t^3	-5,7907	1,0430	-5,55	0,0000
q_1	11684,2530	282,8990	41,30	0,0000
q_2	12457,0559	287,3388	43,35	0,0000
q_3	12138,0542	291,5674	41,63	0,0000
q_4	13921,6368	295,6681	47,09	0,0000

Použití regresní analýzy



Použití regresní analýzy

Vyjdeme-li z uvedeného regresního modelu, dostaneme predikce da rok 2013 spolu s 95% intervaly spolehlivosti

	predikce	dolní	horní
2013, 1.čtvrtletí	24423,14	24008,05	24838,23
2013, 2.čtvrtletí	25237,73	24777,01	25698,44
2013, 3.čtvrtletí	24943,50	24431,21	25455,79
2013, 4.čtvrtletí	26734,30	26164,51	27304,09