Úvod do štatistického softvéru R

Tibor Žuffa, Jakub Benjamín Vrba

Čo je R?

- R je jazyk a prostredie pre štatistické výpočty a grafiku
- Jedná sa o objektovo-orientovaný skriptovací jazyk
- GNU projekt, ktorý je podobný jazyku S
- Ľahko rozšíriteľný o ďalšie metódy
- Voľne dostupný

Prostredie R

- Prostriedky pre efektívnu manipuláciu a ukladanie dát
- Sada operátorov pre výpočty nad vektormi a maticami
- Rozsiahle a integrované prostriedky na analýzu dát
- Grafické prostriedky pre analýzu a zobrazovanie dát

Ako začať?

- Link na github s príkladmi:
  https://github.com/tibor1/BigData
- Link na stiahnutie R:
  http://cran.r-project.org/
- Rozširovacie balíky:
  http://cran.rproject.org/web/packages/
- Rozhranie pre R - RStudio:
- *http://www.rstudio.com/ide/*

Čo nás dnes čaká?

- Dátové typy v R
- Práca s vektormi a maticami
- Grafické zobrazovanie dát
- Funkcie v R
- Analýza časových radov

Ukážky v RStudio

Dátové typy

- Numerická hodnota
- Boolean
- Reťazec
- Vektor
- Faktor
- Pole
- List
- Matica
- Tabuľka dát

Ukážka v RStudio

Základné matematické funkcie

Funkcia	Popis	Príklad
abs(*x*)	absolútna hodnota	
sqrt(*x*)	odmocnina	
ceiling(*x*)	zaokrúhlenie nahor	ceiling(3.475) = 4
floor(*x*)	zaokrúhlenie nadol	floor(3.475) = 3
trunc(*x*)	orezanie desatinnej časti	trunc(5.99) = 5
round(*x*, **digits=***n*)	zaokrúhlenie na určitý počet desatinných čísel	round(3.475, 2) = 3.48
signif(*x*, **digits=***n*)	Zaokrúhlenie na určitý počet číslic	signif(3.475, 2) = 3.5
cos(*x*), **sin**(*x*), **tan**(*x*), ...	goniometrické funkcie	
log(*x*)	prirodzený logaritmus	
log10(*x*)	Logaritmus so základom 10	
exp(*x*)	e^x	

Niektoré štatistické funkcie

Funkcia	Popis
mean(*x*)	priemer
sd(*x*)	smerodajná odchýlka
var(x)	variancia
median(*x*)	medián
quantile(*x*, *p*)	kvantil vektora x s pravdepodobnosťou p
range(*x*)	rozsah vektora x
sum(*x*)	suma hodnôt vektora x
min(*x*)	minimum
max(*x*)	maximum

Vyrovnávanie časových radov

- vylučovanie sezónnych a náhodných výkyvov v časovom rade
- po očistení môžeme posudzovať vývojovú tendenciu

Metódy:

- metóda kĺzavých priemerov
- exponenciálne vyrovnávanie

Metóda kĺzavých priemerov (MA)

- princíp vo výpočte priemerných hodnôt určitého počtu hodnôt
- vypočítanú priemernú hodnotu priraďujeme k prostrednému obdobiu (kĺzavej časti)
- dĺžka kĺzavej časti je nepárna h=2m+1
- miera očistenia závisí od zvolenej dĺžky
- vyjadríme vzorcom:

$$y'_t = \frac{y_{t-m} + y_{t-m+1} + \ldots + y_{t-1} + y_t + y_{t+1} + \ldots + y_{t+m-1} + y_{t+m}}{2m+1}$$

Zdroj: Ivan Martoš, Petra Vrablecová: Vyrovnávanie časových radov

Ukážka v RStudio

Exponenciálne vyrovnávanie

- adaptívny model – najnovšie pozorovania sú najdôležitejšie pri vytváraní prognózy
- krátkodobé prognózovanie
- nenáročnosť a nízke náklady

- Brownovo exponenciálne vyrovnávanie
- Holtovo exponenciálne vyrovnávanie

Brownovo exp. vyrovnávanie

- automatické váženie všetkých predchádzajúcich údajov - váha klesá exponenciálne s časom
- určujeme prognózu na jedno obdobie dopredu
- vhodné iba na čas. rady s konštantným trendom

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t$$

$\hat{y}_{t+1}$ *nová predpoveď*

α *vyrovnávacia konštanta* $0 < \alpha < 1$

y_t *nové pozorovanie*

$\hat{y}_t$ *predpoveď z obdobia* t

Ukážka v RStudio

Holtovo exp. vyrovnávanie

- tzv. dvojité exponenciálne vyrovnávanie
- uvažuje aj prítomnosť trendu

exponenciálne vyrovnaný rad: $L_t = \alpha y_t + (1-\alpha)(L_{t-1} + T_{t-1})$

odhad trendu: $T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$

predpoveď na p období: $\hat{y}_{t+p} = L_t + pT_t$

L_t *nová vyrovnaná hodnota*

α *vyrovnávacia konštanta* $0 < \alpha < 1$

y_t *nové pozorovanie*

β *vyrovnávacia konštanta pre odhad trendu* $0 < \beta < 1$

T_t *odhad trendu*

p *počet období predpovede*

$\hat{y}_{t+p}$ *predpoveď premennej*

Ukážka v RStudio

Dekompozícia časového radu

- aditívny model – zložky sú nezávislé

$$Y_t = T_t + S_t + C_t + R_t$$

- multiplikatívny model – zložky sú zavislé

$$Y_t = T_t * S_t * C_t * R_t$$

zdroj:Róbert Černý: Analýza zložiek časových radov

Ukážka v RStudio

Modely B-J metodológie

- Autoregresný model(AR)
- Pohyblivé priemery(MA)
- AR+MA (ARMA)
- Integrovaný ARMA (ARIMA(p,d,q))

Ukážka v RStudio

Ďakujeme za pozornosť