Cvičenie 4

Predikcia časových radov

Pri predikcii časových radov sa snažíme predpovedať hodnotu v čase *t* na základe predchádzajúceho priebehu hodnôt *t*-1, *t*-2, *t*-3.... Pre predpovedanie y_t môžeme priamo použiť aj jednoduchý model lineárnej regresie do ktorého dosadíme ako vstupné atribúty časovo oneskorené hodnoty:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_M y_{t-M}$$

kde *M* je veľkosť zvoleného časového okna, ktoré udáva koľko hodnôt z minulosti zahrnieme do modelu. Ide o tzv. autoregresný model (keďže hodnota priebehu v čase *t*, závisí od svojich predchádzajúcich hodnôt).

Niekedy môžeme mať viacero časových radov, ktoré sú navzájom závislé (napr. môžeme mať časový priebeh nákladov x_t a zisku y_t), tzn. pre predpovedanie hodnoty závislej premennej y_t môžeme zahrnúť do modelu nie len jej predchádzajúce hodnoty y_{t-1}, y_{t-2}, ... ale aj aktuálnu a predchádzajúce hodnoty nezávislej premennej x_t, x_{t-1}, x_{t-2}, ... Môžeme mať aj naraz viacero nezávislých premenných.

V ekonometrii nás často nemusia zaujímať priamo hodnoty y_t, ale relatívny rozdiel od predchádzajúcej hodnoty $y_t - y_{t-1}$ (rozdiel napr. predstavuje relatívny nárast resp. pokles zisku a pod.). Samotný model teda môže mať na ľavej strane rozdiel závislej premennej $y_t - y_{t-1}$ a na pravej strane priamo oneskorené hodnoty, alebo oneskorené rozdiely závislej aj nezávislej premennej (napr. $y_{t-1} - y_{t-2}$, alebo $x_{t-3} - x_{t-4}$).

Pri výpočte by sme mohli použiť priamo funkciu lm, ale bolo by potrebné dáta predspracovať a z časového radu by bolo potrebné vytvoriť trénovacie príklady posúvaním časového okna pre výstupnú hodnotu *t* = 1, 2, 3, ... Namiesto lm môžeme v R použiť balík dynlm a funkciu dynlm, ktorá umožňuje priamo zapísať model, ktorý obsahuje oneskorené hodnoty a rozdiely. Napr. model predpovedajúci rozdiel $y_t - y_{t-1} = \beta_0 + \beta_1(y_{t-1} - y_{t-2}) + \beta_2 x_{t-4}$ je možné zapísať v tvare d(y) ~ L(d(y),1) + L(x,4). Funkcia d() vypočíta rozdiel a L() časové oneskorenie.

Najprv si nainštalujeme balík dynlm a načítame si dátovú množinu USMacroG, ktorá obsahuje viacero časových radov makroekonomických ukazovateľov USA.

```
install.packages("dynlm")
library("dynlm")

data("USMacroG", package="AER")
summary(USMacroG)
```

Zobrazíme si v jednom grafe priebehy príjmov dpi a spotreby consumption:

```
ts.plot(USMacroG[, c("dpi", "consumption")], lty=c(3,1))
```

Aby sme zistili či je vhodné modelovať časový rad lineárnym modelom, môžeme si vypočítať autokorelačné koeficienty, tzn. koreláciu medzi hodnotami y_t a y_{t-1}, y_t a y_{t-2}, atď. Pre výpočet a zobrazenie môžeme použiť funkciu acf, napr. zobrazíme autokorelačné koeficienty pre časový rad dpi s maximálnym oneskorením 4:

```
dpi = USMacroG[,"dpi"]
autocorr = acf(dpi, lag.max=4)
```

Môžeme si zobraziť aj závislosť medzi y a y_{t-k} kde *k* je oneskorenie, napr. ak si chceme zobraziť závislosť medzi y_t a y_{t-1}, najprv si prevedieme časovú postupnosť na vektor funkciou as.numeric a potom si vyberieme prvky od [1:N-1] a [2:N] kde N je dĺžka vektora (počet hodnôt v časovom rade):

```
dpi = as.numeric(dpi)
plot(dpi[1:length(dpi)-1], dpi[2:length(dpi)])
cor(dpi[1:length(dpi)-1], dpi[2:length(dpi)])
```

Ak porovnáme hodnotu korelácie z posledného výpisu s výstupom autokorelačnej funkcie acf pre oneskorenie 1 (autocorr[1]) zistíme, že sa odlišujú pričom by malo ísť o koreláciu medzi rovnakými hodnotami. Rozdiel je spôsobený odlišnou normalizáciou pri výpočte korelačného koeficientu, acf hodnota je normovaná 1/N a cor 1/(N-1).

Vytvoríme model ktorý bude predikovať spotrebu v čase *t* na základe aktuálnej hodnoty príjmov dpi a predchádzajúcej spotreby v čase *t*-1:

```
cons_lm1 <- dynlm(consumption ~ dpi + L(consumption),
data=USMacroG)
```

Výsledný objekt je podobného typu ako pri lineárnej regresii, tzn. funkcia summary napr. zobrazí hodnoty vypočítaných koeficientov a ich významnosť pre predikciu a základné štatistiky o rezíduách. Podobne môžete vypočítať kvadratickú chybu na trénovacích dátach funkciou deviance.

```
summary(cons_lm1)
deviance(cons_lm1)
```

Predikované hodnoty pre celý časový rad si môžeme vypočítať funkciou fitted:

```
fitted(cons_lm1)
```

Aby sme vyhodnotili presnosť a kvalitu modelu, môžeme si zobraziť časový priebeh a predikované hodnoty v jednom grafe. Tak isto je užitočné si zobraziť v čase aj hodnoty rezíduí, ktoré sa dajú vypočítať funkciou residuals:

```
ts.plot(USMacroG[,"consumption"], fitted(cons_lm1),
lty=c(1:2))
```

```
ts.plot(residuals(cons_lm1))
```

Úlohy na cvičení

1. Načítajte dáta USMacroG. Pre časový rad consumption vypočítajte a zobrazte pomocou funkcie acf autokorelačné koeficienty až do oneskorenia 10.
2. Pre časový rad consumption si vytvorte dátovú množinu zloženú z oneskorených atribútoch v čase *t*, *t*-1, *t*-2, *t*-3, *t*-4 a *t*-5. Pomocou funkcie cor vypočítajte korelačnú maticu pre všetky atribúty. Zistite s ktorým atribútom je najviac korelovaný atribút *t*-2, *t*-3 a *t*-4, a zobrazte závislosti medzi *t*-2, *t*-3 a *t*-4 a najviac korelovanými atribútmi.
3. Pomocou funkcie cor vypočítajte korelačný koeficient medzi hodnotami časového radu consumtion a všetkými ostatnými ukazovateľmi v dátovej množine USMacroG. Vyberte dva ukazovatele, ktoré sú s consumption najviac korelované. Vytvorte prediktívny model, ktorý predikuje consumption$_t$ na základe týchto dvoch atribútov oneskorených o t-1.
4. Zobrazte priebeh radu consumption a predikcie modelu z predchádzajúceho príkladu. Vypočítajte rezíduá a zobrazte priebeh rezíduí.
5. Vytvorte autokorelačné prediktívne modely consumption$_t$ = β_0 + $\beta_1$dpi$_t$ + $\beta_2$dpi$_{t-1}$ a consumption$_t$ = β_0 + $\beta_1$dpi$_t$ + $\beta_2$dpi$_{t-1}$ + $\beta_3$consumption$_{t-1}$ a consumption$_t$ = β_0 + $\beta_1$dpi$_t$ + $\beta_2$consumption$_{t-1}$. Na jednom grafe zobrazte priebeh časového radu consumption a predikcie všetkých troch modelov.
6. Vyhodnoťte presnosť modelov pomocou funkcie deviance. Zistite, ktorý model je najlepší a zistite, ktorý vstupný atribút je najdôležitejší pre tento model. Zobrazte X-Y závislosť medzi consumption a najdôležitejším atribútom.
7. Zobrazte časový priebeh rezíduí a histogram rezíduí pre model consumption$_t$ = β_0 + $\beta_1$dpi$_t$ + $\beta_2$dpi$_{t-1}$ + $\beta_3$consumption$_{t-1}$.
8. Načítajte dáta AirPassengers. Pomocou funkcie acf zobrazte autokorelačný graf. Všimnite si ako sa mení hodnota autokorelačných koeficientov pri časových radoch s cyklickou zložkou. Vytvorte autoregresný model $y_t = \beta_0 + \beta_1(y_{t-1} - y_{t-2}) + \beta_2 y_{t-4}$. Dekomponujte časový priebeh AirPassengers na aditívne zložky pomocou funkcie dec = decompose(AirPassengers). Vyberte z priebehu iba zložku trendu dec$trend a vytvorte autoregresný model $y_t = \beta_0 + \beta_1(y_{t-1} - y_{t-2}) + \beta_2 y_{t-4}$ pre časový priebeh trendu. Vypočítajte predikciu modelu funkciou fitted a pripočítajte k nej sezónnu zložku dec$seasonal. Výsledný časový rad zobrazte v jednom grafe spolu s priebehom AirPassengers a predikciou predchádzajúceho modelu.