

Aegridade analüüs

Praktikum nr. 11, 2016

Erinevad parameetrite leidmise meetodid. Mudelite võrdlemine AIC abil. ARIMA tüüpi mudelite sobitamine praktikast pärinevatele aegridadele. Prognoosimine teisendatud reale sobitatud mudeli korral

Mudelite parameetreid saab hinnata mitut moodi, millest kõige levinumateks moodusteks on suurima tõepära meetod (nõuab eeldust häirituste jaotuse kohta) ning ühesammuliste prognoosivigade minimiseerimine. R-is saab kasutatava meetodi ette anda arima käsusuvandiga `method`, võimalikud väärtused on `CSS` (tinglik ruutude summa), `ML` (suurima tõepära meetod) ning `CSS-ML` (esialgsed parameetrite hinnangud leitakse tinglike ruutude summa abil ning seejärel täpsustatakse suurima tõepära meetodil).

Kuigi sageli eelistatakse praktikas vähima parameetrite arvuga sobivat mudelit, on sageli suur kiusatus võtta kasutusele veidi suurema arvu parameetritega, kuid mingis mõttes paremini sobiv mudel. Et võrrelda mudeleid omavahel, on üheks üldkasutatavaks abivahendiks Akaike Informatsioonikriteerium (AIC) või selle vähestele vaatlustele sobitatud variant AICc. Selle kasutamiseks tuleb teha eeldus häirituste jaotuse kohta ning kriteeriumi väärtus arvutatakse vaatluste logaritmilise tõepära $\ln L$ ja parameetrite arvu k järgi valemiga

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

Saab näidata, et piisavalt suure vaatluste arvu korral lähendab vaadeldavatest mudelistest "tegelikku" mudelit kõige paremini vähima AIC väärtusega mudel. Tähtis on aga aru saada, et võrdlemiseks peavad olema mudelid sobitatud täpselt samadele andmetele, näiteks ei sobi võrrelda AIC baasil erinev arv kordi diferentsitud reale sobitatud mudeleid või erinevalt teisendatud ridadele sobitatud mudeleid.

Seni oleme enamasti sobitanud mudeleid andmetele, mis olid eelnevalt mingi ARIMA tüüpi mudeli abil genereeritud. Praktikast pärinevate ridade korral võib juhtuda, et tegelikult ükski mudel ei sobi päris hästi ning osa valikuid tuleb teha ka lähtudes konkreetse rea sisulisest tähendusest.

Lisaks vaatleme veel põgusalt prognoosimist juhul, kui me oleme sobitanud mudeli teisendatud rea jaoks. Olgu esialgne rida y_t , $t = 1, \dots, n$ ning olgu $z_t = f(y_t)$, kus f on mingi pööratav funktsioon (näiteks logaritm positiivse rea z_t korral). Olgu z_{t+1} prognoos ajal t kujul $\hat{z}_{t+1}$, siis

$$z_{t+1} = \hat{z}_{t+1} + A_{t+1},$$

kus A_{t+1} on tsentreeritud ning me teame ka tema standardhälvet. Sel juhul suuruse y_{t+1} nihketa hinnanguks ei ole enamasti $f^{-1}(\hat{z}_{t+1})$, vaid seda väärtust tuleb korrigeerida. Konkreetne korrigeerimine sõltub funktsioonist f ja A_{t+1} jaotusest. Meie vaatleme ainult logaritmimise juhtu; siis on nihketa prognoosiks

$$\hat{y}_{t+1} = e^{\hat{z}_{t+1}} E(e^{A_{t+1}})$$

ning $E(e^{A_{t+1}})$ on paljude jaotuste puhul teada. Näiteks normaaljaotuse eeldusel saame nihketa prognoosiks

$$\hat{y}_{t+1} = e^{\hat{z}_{t+1} + \frac{\sigma_A^2}{2}}.$$

Suuruste A_t standardhälve on aga ühesammuliste prognooside vigade standardhälve ning selle jaoks on hinnang tarkvara poolt sobitamise käigus enamasti juba leitud.

Praktikumi ülesandeks on leida parimad ARIMA tüüpi mudelid aegridadele andmestikust prax11.csv ning kõigi esimeses ja teises praktikumis vaadeldud praktikast pärinevate andmestike jaoks. Kui varieeruvus muutub ajas, siis uurida, kas eelnev logaritmine ja seejärel mudeli sobitamine annab parema tulemuse, kui otse esialgsete andmete modelleerimine.