

Aegride analüüs

Praktikum nr. 1, 2016

Käsitletavad teemad:

- Aegria andmete hankimine ning mitmesugustes formaatides andmete analüüsikeskkondadesse sisselugemine
- Aegria graafiline kujutamine
- Globaalse trendi sobitamine ja eemaldamine

Aegrida on teatud sagedusega (nt kord minutis, kord päevas, kord kuus, kord kvartalis, kord aastas) mõõdetud andmete kogum. Aegria andmestikus peaks olema nii mõõdetud andmed kui info selle kohta, kuidas igale mõõtmisele vastavat ajamomenti kindlaks teha. Osades andmestikes on kirjas kommentaar esimese vaatluse aja ja sageduse kohta ning ülejäänud andmestik koosneb ainult mõõtmistulemustest. Osades andmestikes on ajainfo ühes tulbas (nt kujul 2010-09-01) ning vastav vaatlustulemus teises tulbas. Sageli on aga ajainfo mitme välja vahel laiali (nt tulpades aasta, kuu, päev) ning mõnikord esitatakse andmestik ka tabeli kujul (kus read vastavad näiteks aastatele ning veerud kuudele). Erinevused võivad olla veel kasutatavates väljaeraldajates (tühik, tabulaator, koma, semikoolon jms). Seega on tähtis andmestiku sisuga kõigepealt visuaalselt tutvuda ning vaatlusest saadud info põhjal osata andmeid õigel kujul analüüsikeskkonda sisse lugeda.

Käesolevas praktikumis vaatleme andmete sisselugemist tekstifailidest, kuna muul kujul olevaid andmeid on enamasti teksti kujul (nt csv failina) küllalt lihtne salvestada.

Analüüsikeskkonnana kasutame R tarkvara, mis on vabavara. Selle saab igaüks oma arvutisse paigaldada lehelt <http://ftp.eenet.ee/pub/cran/>.

Ennem andmete sisselugemist on neid hea vaadata lihtsa tekstiredaktoriga, mis ei püüa neid kuidagi ilusti ja kasutajasõbralikult näidata (nagu näites Excel või Word vms), vaid näitavad faili sisu täpselt sellisena, nagu see on. Selleks sobivad hästi näiteks programmid Notepad++ (saadaval aadressilt <https://notepad-plus-plus.org/>), Editpad Lite ja mitmesugused programmeerimiseks mõeldud tekstiredaktorid.

Ül. 1 Andmestike hankimine. Internetis on saadaval suurel hulgal aegride andmestikke. Hangime ühe andmestiku: Kinnisvara ostu/müügitehingute summad Eestis (alustades lehelt <http://www.stat.ee/>, pärast mitmeid valikuid andmestik kv015, formaadiks valida tabelialdusega pealkirjata tekst). Aja kokkuhoiu mõttes on ülejäänud andmestikud (USA töötuse määr lehe <http://www.bls.gov/> vahendusel, Cisco aktsia ajaloolised sulgemishinnad aadressilt <http://finance.yahoo.com> ja tarbijahinnaindeks statistikaameti lehelt) saadavad kursuse veebilehelt <http://moodle.ut.ee/course/view.php?id=1742>. Kopeerige andmed oma kodukataloogi (või selle mingisse alamkataloogi) ning tutvuge (näiteks notepad vahendusel) nende failide sisuga.

Aegride sisselugemiseks R keskkonda on mitmeid võimalusi, kuid siin vaatleme ühte küllalt universaalset tegevuste jada, mis sobib kõigi kursusel vaadeldavate andmetike korral.

1. Andmete sisselugemine. Selleks sobiv käsk on
`andmed=read.table(failinimi,...)`,
kus kolme punkti asemel võib anda loetelu mitmetest suvanditest. Mõned olulisemad:
`skip=x` - ignoreerida x rida faili algusest. Vaikimisi 0
`sep="c"` - andmeväljade eraldajaks on sümbol c. Vaikimisi tühikud või tabulaatorid
`header=T` - esimene sisseloetav rida sisaldab veergude nimesid
`dec=","` - kümnenderaldajaks on koma. Vaikimisi on punkt
`na.strings="s"` - millega on puuduvad väärtused tähistatud. Vaikimisi on NA.
2. Aegria väärtuste väljaeraldamine. Kui aegrida paikneb ühes veerus, siis sobivad käsud kujul `rida=andmed$nimi` (kus nimi on vastava veeru nimi) või `rida=andmed[,nr]`, kus nr on veeru järjekorranumber. Kui vaatlused on esitatud tabeli kujul nii, et iga rida vastab näiteks aastale ja veerud kuudele või kvartalitele, siis saab õigel kujul olevad arvud kätte käsuga `rida=c(t(andmed[,a:b]))`. Siin a on veeru number, kus on esimene vaatlus ning b on veeru number, kus on viimane vaatlus selles reas. Seletuseks: `andmed[,a:b]` eraldab andmestikust näidatud veergudega määratud osa, `t()` transposeerib selle ning `c()` väljastab tulemuse ühe vektorina, kuhu argumendiks oleva maatriksi väärtused lähevad veergude kaupa järjest.
3. Aegria lõpus olevate puuduvate väärtuste eemaldamine. Sageli on andmestiku lõpus mõned puuduvad väärtused lihtsalt selle tõttu, et need vastavad tulevikukuupäevadele ja seega ei ole tegemist andmetega, mis sisuliselt puudu on (nagu näiteks mingi vahelejäädud mõõtmistulemus minevikus). Need tuleks enne aegria uurimist kõrvaldada. Üks võimalus on kasutada käsku kujul `rida=rida[1:(length(rida)-x)]`, kus x on rea lõpust eemaldatavate väärtuste arv. Kui puuduvad väärtused on ainult real lõpus, siis võib eemalda need ka käsuga `rida=rida[!is.na(rida)]`, mis eemaldab kõik puuduvad väärtused.
4. Andmete paigutamine ajas kasvavalt. Osades andmetikes on kõige uuemad andmed eespool; aegria jaoks oleks vaja, et mida edasi, seda uuemad andmed. Kui on vaja järjekorda muuta, siis seda saab teha käsuga
`rida=rida[length(rida):1]`
5. Mõõtmistulemustest aegria tegemine. Selleks, et oleks väärtuste rida aegreana käsitleda, tuleb lisada ka ajainformatsioon selle kohta, kui sageli tulemusi mõõdeti. Aegria tegemine käib kujul `aegrida=ts(rida,...)`, kus kolme punkti asemel võib olla loetelu suvanditest. Mõned tähtsamad:
`frequency=n` - mitu korda ajaühikus mõõdetakse. Vaikimisi 1, kvartaalsete andmete puhul 4, igakuiste andmete puhul 12.

`start=x` - esimese mõõtmise aeg. Kui mõõtmisi on üks kord ajaühikus või kui esimene väärtus on esimese ajaühiku esimene mõõtmine, siis `x` peaks olema esimese ajaühiku (nt aasta) number; Kui andmestik ei alga perioodi algusest (nt esimene tulemus on 2000 aasta mais), siis tuleb `x` asemel kirjutada paar ajaühikust ja mõõtmise numbrist kujul `c(aasta,nr)`.

6. Aegrea graafiku tekitamine käib käsuga `plot(aegrida)`. Kui on soovi kuvada ainult osa aegreast, siis seda saab teha käsku `window()` kasutades kujul `plot(window(aegrida,start=algus,end=lopp))`, kus `algus` ja `lopp` võivad olla lihtsalt perioodi (tavaliselt aasta) numbrid või paarid perioodi numbrist ja mõõtmise järjekorranumbrist.

Ül. 2 Lugege andmestikud R keskkonda sisse ja kujutage graafiliselt neli esimeses ülesandes hangitud aegrida. Salvestage andmed hilisemaks kasutamiseks käsuga kujul

```
save(rida1,rida2,file="prax1andmed.RData"))
```

Hiljem saab neid laadida käsuga `load("prax1andmed.RData")`.

Globaalse trendi sobitamine

Tutvume lihtsamate globaalse trendi leidmise võtetega. Olgu meil antud vaatlused $z_1, z_2, \dots, z_n$. Edasises vaatlеме parameetrite suhtes lineaarse trendi leidmist kujul

$$\text{trend}(t) = c_0 + c_1 \cdot f_1(t) + c_2 \cdot f_2(t) + \dots + c_m \cdot f_m(t),$$

kus f_i , $i = 1, \dots, m$ on mingid fikseeritud funktsioonid või teiste aegridade väärtused. Samuti vaatleme mõningaid parameetrite suhtes mittelineaarselt esituvate trendikõverate leidmise võimalusi.

Lineaarse trendi parameetrite leidmiseks on R keskkonnas mitmeid võimalusi. Kui meie eesmärgiks on ainult vastava trendikõvera sobitamine, siis on kõige õigemaks käsuks `lm()` (*linear models*). Samas on aegridade analüüsimisel vastava trendi leidmine ainult üks paljudest tegevustest, mida on aktsepteeritava mudeli saamiseks vaja teha ning seetõttu on mõistlikum ka vaadeldaval kujul trendi leida selliste vahendite abil, mis võimaldavad lihtsa vaevaga mudelit hiljem täiendada. Seetõttu vaatleme parameetrite suhtes lineaarse trendi leidmist käsuga `arima()`, mis võimaldab meil hilisemates praktikumides pärast trendi sobitamist prognoosivigasid uurida ja neile vajadusel täiendavaid aegridade mudeleid sobitada.

Vaadeldaval kujul oleva trendi leidmiseks tuleb tekitada ühe regressorfunktsiooni ($m=1$ korral) vektor vaadeldavatele ajamomentidele vastavatest väärtustest ning mitme regressori korral maatriks, mille veergudes on funktsioonide f_i , $i = 1, \dots, m$ väärtused ajamomentidel $t = 1, 2, \dots, n$ näiteks käsuga `regressorid=cbind(F1, ..., Fm)`, kus vektorid F_i sisaldavad funktsioonid f_i väärtusi ning kasutada seejärel käsku kujul

```
trend=arima(z,xreg=regressorid)
```

Siin `cbind` käsus kolm punkti tähistavad kõigi vahepealsete liikmete väljakirjutamist ning nimetuse `trend` asemel võib kasutada suvalist enda antud nime. Aegreast trendi lahutamisel saadava rea saab kätte kujul `residuals(trend)` (või kujul `trend$residuals`), olemasolevatele väärtustele vastavaid trendi väärtuseid saab arvutada käsuga `rida-residuals(trend)`, tulevikuväärtuste arvutamisel on käsu kujuks

```
predict(trend,n.ahead=mitu,newxreg=cbind(F1_uus,...,Fm_uus))
```

Eelnevas `Fi_uus` vastab regressori $f_i(t)$ nende ajamomentide väärtustele, mille jaoks me prognoose tahame leida.

Otsitavate parameetrite suhtes mittelineaarse trendi sobitamisel saab aga kasutada käsku `nls()`. Käsu kasutamiseks lugege vastavat R abifaili (trükkides käsuraale `?nls`).

- Ül. 3 Sobitage lineaarne ja kuuptrend tarbijahinna indeksi andmetele. Selleks leidke selle rea vaatluste arv `n` käsuga `length()`, defineerige vektorid `F1=1:n`, `F2=F1**2` ja `F3=F1**3` ning seejärel leidke trendile vastavad mudelid.
- Ül. 4 Mitme aegrea kujutamine ühel joonisel: kõigepealt joonistada ühe aegrea graafik plot käsuga, kasutades lisaks suvandit `col="värvi_nimi"` ning seejärel lisada teiste aegride graafikud käsuga `lines(rida,col="värv")`. Moodustage lineaarse trendi ning ruuttrendi väärtustele vastavad aegread ning kujutage need koos tarbijahinna indeksiga ühel joonisel.
- Ül. 5 Sobitage tarbijahinna indeksile mittelineaarne trend kujul $c_1 \cdot e^{c_2 t}$. Lisage ka see eelnevale graafikule. Käsu `nls()` kasutamisel peab ette andma otsitava funktsiooni valemi kujul, mis praegusel juhul on `z~c1*exp(c2*t)` (kus `t` peab olema eelnevalt defineeritud ajamomentide vektor) ning samuti parameetritele realistlikud algväärtused, mille ümbrusest siis parimad leitakse. Praegusel juhul on tarbijahinna indeksi algväärtus veidi suurem sajast ning parameeter `c2` peaks kirjeldama indeksi igakuist kasvuprotsenti, mis peaks olema küllalt lähedane nullile (alla 1%). Seetõttu võib algväärtused ette anda suvandiga `start=list(c1=100,c2=0.01)`.
- Ül. 6 Trendi tulevikuväärtuste prognoosimine. Vaatleme kuuptrendi järgneva 5 väärtuse prognoosimist. Selleks tuleb regressorridade jaoks tekitada vektorid järgmisele viiele ajamomendile vastavate väärtustega siis kasutada `predict` käsku. Kui `n` on tarbijahinna indeksi vaatluste arv, siis kuuptrendi korral $f_1(t)$ järgnevad 5 väärtust saab defineerida kujul `F1_uus=(n+1):(n+5)` ja $f_2(t)$ ning $f_3(t)$ väärtused on ikka eelneva vektori väärtuste ruudud ja kuubid.
- Ül. 7 Oluline küsimus trendikõverate sobitamisel on nende kasutatavus prognoosimiseks. Vaadeldge prognoosivigade graafikuid kõigi sobitatud trendikõverate puhul ja katsuge aru saada, kas on mõtet tulevikku prognoosida ainult trendikõverat kasutades.