**Aegread ja prognoosimine**

15. praktikum

Loengus näidati Holt-Winters’i meetodil leitud Tallinki reisijate arvu prognoosi. Esimeseks tänase praktikumi proovitööks ongi sama prognoosi „taasleidmine“.

Loeme sisse näiteandmed:

print(load(url("http://www.ms.ut.ee/mart/andmeteadus/reisijaid.RData")))

head(Tallink)

attach(Tallink)

12 vaatlust mahub ühte perioodi (12 kuud teevad kokku ühe aasta)

ja teisendame andmed aegreaks:

Aegrea algus: mai 2012

y=ts(reisijaid, start=c(2012,5), frequency=12)

y

plot(y)

Järgnevalt hindame Holt-Winters’i mudeli. Sessoonsuse modelleerimisel kasutame multiplikatiivset mudelit – kas oskad aimata, miks?

m1=HoltWinters(y, seasonal="mult")

m1

Hinnanguliselt püsib trend kogu aeg samasugune ega muutu ajas.

Hinannguline trend on 104 lisanduvat inimest igal kuul.

Smoothing parameters:

alpha: 0.230094

beta : 0

gamma: 0.5290757

Coefficients:

[,1]

a 7.936234e+05

b 1.039965e+02

s1 9.662629e-01

s2 1.003099e+00

s3 1.200376e+00

s4 1.507816e+00

s5 1.224919e+00

s6 8.811109e-01

s7 1.006581e+00

s8 8.483086e-01

s9 9.758298e-01

s10 7.237498e-01

s11 8.036465e-01

s12 9.190456e-01

Hinnang jooksvale tasemele (peale sessoonsuse mõju eemaldamist, viimane vaadeldud kuu)

Sessoonsus muutub suhteliselt kiiresti. Järgmise aasta sessoonsuse kirjeldamisel kasutatakse kaaluga 0.53 eelmise aasta sesoonsust ja varasemate aastate sesoonsuseid kokku arvestatakse kaaluga 0,47

Parameeter S1 ei näita tegelikult jaanuarikuu sesoonsust vaid käib hoopis aprillikuu kohta. Miks? Viimane vaatlus aegreas oli pärit märtsist 2018, ja sesoonsuse mõjud s1,… liiguvad sealt edasi (aprill; mai;…)

Kontrollime mudeli prognoosivõimet – kui hästi ta on varem suutnud tulevikku näha?

plot(m1)

Prognoosime leitud mudeli abil kaks aastat (24 kuud) tulevikku:

library(forecast)

prog = forecast(m1, h=24)

prog

plot(prog, xlim=c(2012, 2020), main="Tallinki reisijate arvu prognoos")

Soovi korral võid mängida kunstnikku ja leida kõige meelepärasem graafik prognooside esitamiseks:

prog = forecast(m1, h=24, level=0.95)

plot(prog, xlim=c(2012, 2020), main="Tallinki reisijate arvu prognoos")

prog = forecast(m1, h=24, level=c(0.5, 0.9, 0.95, 0.99))

plot(prog, xlim=c(2012, 2020), main="Tallinki reisijate arvu prognoos",

shadecols=c("gray95", "gray85", "gray80", "gray75") )

Ning võid lisada graafikule ka hiljem lisandunud reisijate arvud:

tegelik=c(789272, 852609, 989445, 1223901, 993078, 730631, 785583,

654240, 811261, 549278, 620006, 686488, 779113)

teg=ts(tegelik, start=c(2018, 4), frequency=12)

points(teg, col="red", pch=20, cex=2)

Võid vaadata ka, kuidas erinevad aegrea komponendid (keskmine ilma sessoonsuse mõjuta; trend antud ajahetkel; sesoonsus antud ajahetkel) on ajas käitunud:

plot(fitted(m1))

Võid võrdluseks vaadelda ka seda, kuidas mõni teine meetod üritab aegrida komponentideks jagada:

plot(stl(y, "per"))

**Tunnustevahelised seosed**

*ehk mis muudab naised kurjaks?*

Järgnev näide baseerub järgmisel allikal (kust leiab ka põhjalikumat taustainformatsiooni):

<https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices/Appendix-Timeseries-Regression.pdf>

loeme sisse näiteandmestiku:

install.packages("car")

library(car)

Hartnagel[1:4,]

Andmestiku lühikirjelduse saad käsuga

?Hartnagel

Ja võime andmetele ka pilgu peale heita:

attach(Hartnagel)

plot(fconvict ~ year, type="n",

ylab="Süüdimõistetuid 100,000 naise kohta", xlab="Aasta")

grid(lty=1)

points(year, fconvict, type="o", pch=16)

Soovime mõista mis võiks mõjutada naiskurjamite arvu ühiskonnas. Hindame alustuseks lihtsa mudeli süüdimõistetud naiste arvule. Mudelisse võime lisada näiteks tunnused *degrees* (haritumad naised sattuvad ehk harvem kurjale teele?); *tfr – total fertility rate* (lapse olemasolu ehk vähendab riskantset käitumist?); *partic* (Women's labor-force participation rate per 1000; tööd tehes ja sissetulekut omades pole ehk vaja panka röövima minna?) ; *mconvict* – süüdimõistetud mehi 100000 mehe kohta (mõni aasta lihtsalt sobib pahategudeks?)

> mudel1=lm(fconvict~tfr + partic + degrees + mconvict)

> summary(mudel1)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 127.639997 59.957044 2.129 0.0408 \*

tfr -0.046567 0.008033 -5.797 1.75e-06 \*\*\*

partic 0.253416 0.115132 2.201 0.0348 \*

degrees -0.212049 0.211454 -1.003 0.3232

mconvict 0.059105 0.045145 1.309 0.1995

Residual standard error: 19.19 on 33 degrees of freedom

Multiple R-squared: 0.6948, Adjusted R-squared: 0.6578

F-statistic: 18.78 on 4 and 33 DF, p-value: 3.905e-08

Mudel tuleb ilus, paraku on leitud p-väärtused ja standardvead eksitavad. Oleme analüüsi käigus unustanud, et uuritava tunnuse väärtused on korreleeritud – kui ühel aastal tehakse palju kuritegusid siis tõenäoliselt ka järgmisel aastal esineb palju kuritegevust. Proovime seda vaatluste korreleeritust arvesse võtta. Teeme esimese katse:

library("nlme")

mudel2 = gls(fconvict ~ tfr + partic + degrees + mconvict,

correlation=corAR1(form=~year), method="ML")

summary(m2)

Võime veenduda, et mudel2 sobib andmetega märksa paremini kui vaatluste sõltumatust eeldanud mudel1:

AIC(mudel1, mudel2)

Uurime vaatluste vahelisi korrelatsioone veidi põhjalikumalt. Kui kauge aja taha ikka ühe aasta mõju ulatub?

plot(pacf(residuals(mudel1)))

Joonise tõlgendus:

Kõrge tulp *Lag 1* kohal näitab seda, et kuritegevus järjestikustel aastatel on tugevalt korreleeritud (mudeli jääkide korrelatsioon suurem kui 0,8) – kui ühel aastal on keskmisest rohkem kuritegusid (või kuritegudes süüdimõistetud naisi) siis arvatavasti ka järgmisel aastal leiab aset liiga palju kuritegusid. See on ootuspärane ja sellega ka meie poolt hinnatud mudel2 arvestab.

Paraku näeme ka erakordselt madalat tulpa graafikul *Lag 2* kohal. Kui eelmine aasta (lag 1) oli kuritegude arvu poolest keskmine, aga üleelmisel aastal oli palju kuritegusid siis esineb tänavu keskmisest vähem kuritegusid (tuleviku prognoosimisel ei oma tähendust mitte ainult praegune seis, vaid ka trend – kas kuritegevus on langustrendis või tõusutrendis). Sellist veidrat pikaajalist mõju meie mudel2 praegu arvesse ei võta. Kaugemate sõltuvuste kirjeldamiseks peame keerukama mudeli kasutusse võtma:

Kõigepealt hindame keerukamat mudelit kasutades samasuguse mudeli kui mudel2 (eelmise aasta kuritegevus mõjutab tänast):

mudel3 = gls(fconvict ~ tfr + partic + degrees + mconvict,

correlation=corARMA(form=~year, p=1), method="ML")

AIC(mudel2, mudel3)

Ja nüüd hindame keerukama mudeli, mis lubab ka üleelmise aasta kuritegevuse tasemel rolli mängida (lisaks sellele, mis eelmisel aastal toimus):

mudel4 = gls(fconvict ~ tfr + partic + degrees + mconvict,

correlation=corARMA(form=~year, p=2), method="ML")

AIC(mudel2, mudel4)

Keerukam mudel osutus paremaks, jääme selle juurde ja vaatame tulemust:

summary(mudel4)

Kas tulemus on sinu jaoks ootuspärane?