

Lineaarsed mudelid

9. praktikum

Proгноosiintervall

Tahetakse luua perearstidele tööriist, mille abil nad saaksid jälgida väikelaste kasvamist – kas lapsele lisandub kasvu nii, nagu normaalsel lapsel peaks lisanduma. Kasvukõverate loomiseks on aja jooksul mõõdetud uskumatult paljude erinevas vanuses väikelaste pikkust. Loeme sisse neid mõõtmiseid sisaldava andmestiku:

```
andmed=read.csv2("http://www.ms.ut.ee/mart/linmud2018/lapsed2.csv",
  header=T)
head(andmed)
```

Soovime saada mitte ainult teatud vanuses laste keskmist pikkust kirjeldavat kõverat (see, mida lineaarsed mudelid tavaliselt modelleerivad) vaid ka piire, kuhu vahele jääks 50%, 80%, 95%, 99.5% normaalsetest lastest (proгноosiintervale). Mõistlik on teha eraldi graafikud poistele kui tüdrukutele (vaata konspekti lõppu lisatud näidist).

Üritame esmalt luua mudeli, mis kirjeldaks poisslaste kasvamist ajas.

```
m1=lm(pikkus~vanus, data=andmed[sugu=="M",])
```

Piilume ka andmeid:

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".")
```

või

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch=20,
  col=rgb(1, 0, 0, 0.01), cex=1.5)
```

ja lisame joonisele oma mudeli prognosijooni koos 95%-usaldusintervalliga

```
plot(vanus[sugu=="M"], pikkus[sugu=="M"], pch= ".", col="gray70")
x=seq(0,2.5, length=100)
y=predict(m1, data.frame(vanus=x), interval="prediction")
lines(x,y[,1], col=2, lwd=3)
lines(x,y[,2], col=2, lwd=2, lty=2)
lines(x,y[,3], col=2, lwd=2, lty=2)
```

Peaksime ka sellelt jooniselt märkama, et meil esinevad teatavad probleemid mudeliga. Seos vanuse ja pikkuse vahel pole lihtsalt lineaarne.

Proovi mudelit parandada, kasutades kõrgema astme polünoomi...

```
m2=lm(pikkus~vanus+I(vanus^2)+I(vanus^3)+I(vanus^4)+I(vanus^5)+I(vanus^6)
      +I(vanus^7)+I(vanus^8)+I(vanus^9)+I(vanus^10)+I(vanus^11)
      +I(vanus^12), data=andmed[sugu=="M",])
```

ja vali sobiv mudel, näiteks kasutades ühe alltoodud käsu tulemusi:

```
anova(m2)
drop1(m2, test="F")
```

Kumma käsu väljundit on praegu mõistlikum mudeli valikuks kasutada?

Vali välja sobiva suurusega polünoom ja kanna saadud mudeli prognoosid graafikule:

```
m3=lm(pikkus~....)
```

Kontrolli, kas väljavalitud mudeli eeldustega võiks kõik olla korras? Vaata diagnostilisi graafikuid käsu

```
plot(m3)
```

abil. Üks neist graafikutest (*Normal Q-Q plot*) on mõeldud normaaljaotuse eelduse kontrollimiseks. Kui vaatlused on normaaljaotusega (y -tunnuse tinglik jaotus tingimusel et x -tunnuste väärtused on fikseeritud on normaaljaotus) siis peaks vaatluseid iseloomustavad punktid sellel graafikul paiknema enam-vähem sirge peal. Mida näed? Kuidas on lood normaaljaotuse eeldusega?

Proovi joonisele kanda ka Olive mitteparameetrilisi prognoosiintervalle (nende arvutamisel võib olla sul kasu sellest, kui vaatad `predict`-käsu `se.fit`-parameetrit, vaatad mida teeb funktsioon `quantile` ning piilud `attributes(summary(m2))`-käsu abil millise nime taga on peidus jääkide standardhälbe hinnang...):

$$[\hat{Y}_{uus} + a_n \hat{q}_{\alpha/2}; \hat{Y}_{uus} + a_n \hat{q}_{1-\alpha/2}],$$

kus

$$a_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n}{n-p}} \cdot \sqrt{1 + \mathbf{x}_{uus}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_{uus}}$$

ja $\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}$ on hinnangud regressioonmudeli jääkide $(\alpha/2)$ ja $(1 - \alpha/2)$ -kvantiilile.

Ülesanne 1

Kanna poiste kasvukõverat iseloomustavale joonisele 0,95- ja 0,99- prognoosiintervallid (nii parameetrilised kui ka mitteparameetrilised) ja saada saadud graafik koos oma kommentaaridega (võrdle parameetrilisi ja mitteparameetrilisi prognoosiintervalle) õppejõule.

Ülesanne 2

Teadusartiklist loeme välja järgmist:

- * Hinnatud regressioonimudel: $Y = 10 + 2x + e$;
- * $MSE = 10$;
- * Regressioonimudeli hindamiseks kasutatud x tunnuse väärtused: kõik täisarvud 1..8, iga x -i väärtust on kasutatud kahe mõõtmise tegemiseks;

Milline tuleb prognoosiintervall (normaaljaotuse eeldusel) $x = 7,5$ korral?

