Log-lineaarsed mudelid

praktikum

Tänases praktikumis uurime diabeedi saamise riski. Andmed on pärit Tervisestatistika ja terviseuuringute andmebaasist (https://statistika.tai.ee/pxweb/et/Andmebaas/Andmebaas__02Haigestumus__01Esmashaigestumus/EH02.px/):

```
load(url("https://www-1.ms.ut.ee/mart/kvalitatiivsed/diabeet.RData"))
Diabeet[1:4,]
```

Antud andmestikus on kirjas 2020.a. uute diabeedijuhtude arvud maakondade (ja vanusegruppide) kaupa:

juhte – diabeedi esmasdiagnooside arv
grupp – vanusegrupp (noor: 0-34a; vana: 35a vanad või vanemad)
maakond – millise maakonna kohta andmed käivad
inimesi – antud vanuses inimeste arv selles maakonnas 2020.aastal.

Võtame andmed kasutusse:

```
attach(Diabeet)
```

Hindame esmalt sellise mudeli, kus kõigil inimestel – sõltumata nende soost, vanusest, usust või elukohast – oleks võrdne võimalus saada diabeedidiagnoosi:

```
m0=glm(juhte~1+offset(log(inimesi)), family=poisson())
summary(m0)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.69562    0.01498  -380.3   <2e-16
```

Saadud mudel näeb välja järgmine:

$$\log(\mathrm{E}(juhte)) = -5{,}7 + \log(inimesi)$$

ehk

$$\mathrm{E}(juhte) = \exp(-5{,}7 + \log(inimesi)) = 0{,}0033\ldots \cdot inimesi\ .$$

Saadud mudelit võime kasutada ka juhtude arvu prognoosimisel. Kui meil oleks 100 000 inimest, siis ootaksime antud mudeli arvates nägevat niimitut uut haigusjuhtu:

```
predict(m0, newdata=data.frame(inimesi=100000), type="response")
```

```
       1
336.0669
```

Praegu on meil aga iga maakonna kohta kaks rida (noorte ja vanade kohta), vaata näiteks:

```
Diabeet[maakond=="Harju maakond",]

   juhte grupp       maakond inimesi
1     91  noor Harju maakond  255121
16  1593  vana Harju maakond  349908
```

Mis juhtuks siis, kui meil oleks kõigest üks number maakonna kohta, st kui meil oleks üks rida Harju maakonna jaoks kus juhtude arvuks oleks 91+1593 ja seal maakonnas elavate inimeste arvuks oleks 161886+197873? Proovime järgi!

```
andmed2=aggregate(cbind(juhte, inimesi)~maakond, FUN=sum)
m0a=glm(juhte~1+offset(log(inimesi)), family=poisson(), data=andmed2)
summary(m0a)

predict(m0a, newdata=data.frame(inimesi=100000),type="response")
```

Kas koondandmete pealt hinnatud mudel/prognoos langevad varasemate tulemustega kokku või mitte?

Uurime järgnevalt, kas kõikides maakondades on inimestel võrdsed võimalused saada diabeeti. Hindame mudeli, kus diabeedi esmasdiagnooside arv võib lisaks inimeste arvust ka maakonnast sõltuda:

```
m1=glm(juhte~factor(maakond)+offset(log(inimesi)), family=poisson())
summary(m1)
```

saadud mudel näeb matemaatilises kirjapildis välja järgmine:

$$\log(E(\textit{juhte})) = -5.88 - 0{,}422 \cdot I(\textit{maakond}=\text{“Hiiumaa“}) + \\ + 0{,}67 \cdot I(\textit{maakond}=\text{“Ida-Virumaa“}) + \ldots + \log(\textit{inimesi})$$

Seega Harjumaa jaoks oleks prognoositav juhtude arv leitav kui

$$E(\textit{juhte}) = \exp(-5.88+\log(\textit{inimesi})) = \exp(-5.88) \cdot \textit{inimesi} = 0{,}0028 \cdot \textit{inimesi}$$

ja Hiiumaa oodatav juhtude arv aastas oleks mudeli järgi:

$$\begin{aligned} E(\textit{juhte}) &= \exp(-5.88-0{,}422+\log(\textit{inimesi})) \\ &= \exp(-5.88)\exp(-0{,}422) \cdot \textit{inimesi} \\ &= 0{,}0028 \cdot 0{,}656 \cdot \textit{inimesi} \end{aligned}$$

ehk siis inimese kohta 0,66 korda „rohkem“ (36% vähem juhte inimese kohta).

Mudel näitab küll erinevusi maakondade vahel, aga kas need erinevused on ka statistiliselt olulised? Vaatame, kas haigestumus inimese kohta tegelikult ikka maakondade vahel on erinev:

```
anova(m0, m1, test="LRT")

Model 1: juhte ~ 1 + offset(log(inimesi))
Model 2: juhte ~ factor(maakond) + offset(log(inimesi))
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        29       3151
2        15       2876 14   275.01 < 2.2e-16 ***
```

Kuna mudelite erinevus on statistiliselt oluline, siis peame järeldama – sama inimeste arvu korral saab mõnes maakonnas rohkem inimesi diabeedi kui teistes maakondades.

Kus tuleks elada, et saada diabeeti? Vaatame, kui palju diabeedijuhte ootaksime 100 000 elaniku kohta erinevates maakondades:

```
nimed=names(table(maakond))
data.frame(nimed,
  prog=predict(m1, newdata=data.frame(maakond=nimed,
      inimesi=100000), type="response"))
```

Mida need tulemused näitavad? Mis võiks vaadeldud erinevusi põhjustada?

Kuid Tartlasi vaevalt huvitab see, mis kuskil kaugetes kolgastes toimub. Meid huvitab rohkem meie armas Kagu-Eesti. Seega teostame sarnase analüüsi vaid siinse piirkonna jaoks:

```
KaguEesti = maakond %in% c("Võru maakond","Põlva maakond",
      "Valga maakond", "Tartu maakond","Viljandi maakond")

m2=glm(juhte~offset(log(inimesi)), family=poisson(),
                                data=Diabeet[KaguEesti,])

exp(coef(m2))*100000
exp(confint(m2))*100000
```

Näeme, et Kagu-Eestis on diabeeti haigestumus 310 juhtu / 100 000 inimese kohta (95%-usaldusintervall 290...331). NB! See pole prognoosiintervall!

Ülesanne 1

Kas Kagu-Eesti maakondade vahel esineb erinevusi diabeeti haigestumuses? Hinda Kagu-Eesti maakondade jaoks mudel **m3**, kus haigusjuhtude arv inimese kohta võib maakondades erinev olla. Kas vajame tunnust maakond? Kas haigusjuhtude arvud inimese kohta on maakondades erinevad või võivad olla samasugused? Milline tuleb vastavat oletust kontrolliva statistilise testi p-väärtus?

Kas inimese vanusel on tähtsust? Kas noored ja vanad jäävad sama sageli diabeeti?

```
m4=glm(juhte~factor(maakond)+factor(grupp)+offset(log(inimesi)),
                    family=poisson(), data=Diabeet[KaguEesti,])

anova(m3, m4, test="LRT")
drop1(m4, test="LRT")

summary(m4)
```

Kuidas interpreteerida „vanade“ parameetrit 2,26340?

Ülesanne 2

Hinda sarnane mudel kogu Eesti jaoks. Leia aasta jooksul oodatavad uued diabeedijuhtude arvud kõigi maakondade jaoks 100 000 inimese kohta nii noorte kui ka vanade jaoks:

Maakond	noored	vanad
Harju maakond	36.52622	454.6378
Hiiu maakond	21.53984	...
Ida-Viru maakond	...	...
Jõgeva maakond	...	...
Järva maakond	...	...

Kuidas võiks kontrollida, kas noorte-vanade erinevus on kõigis maakondades samasugune?