Biomeetria

8. praktikum

Logistiline regressioon

Üritame prognoosida binaarse (kahe võimaliku väärtusega) tunnust teiste tunnuste abil

Loe sisse tudengite andmestik:

```
print(load(url("http://www.ms.ut.ee/mart/biomeetria2015/andmed.RData")))
```

Binaarseks tunnuseks, mille väärtuseid üritame teiste tunnuste abil prognoosida on tudengi sugu (1-naine; 2 - mees). Kuna logistiline regressioon ootab tunnust mille väärtused on vahemikus [0…1] siis peame esmalt muutma oma tunnuse kodeeringut. Kodeerime tunnuse ümber selliselt, et mehed saaksid väärtuse 1 ja kõik ülejäänud (ehk naised) saaksid ümberkodeeritud tunnuse väärtuseks 0:

```
sugu2=1*(sugu==2)
```

Ümberkodeerimise kontrolliks võid võrrelda esialgse tunnuse ja ümberkodeeritud tunnuse sagedustabeleid:

```
table(sugu)
table(sugu2)
```

Näeme, et valimis oli 512 naistudengit ja 148 meestudengit ehk valides arstiteaduskonnast juhuslikult ühe tudengi on meestudengi saamise šansid 148/512:

```
> 148/512
[1] 0.2890625
```

Proovime kas logistilise regressiooni abil jõuame samasuguse tulemuseni. Hindame logistilise regressiooni mudeli kus meheks olemise tõenäosuse (või šansi) prognoosimiseks pole kasutatud ühegi teise tunnuse abi:

```
> mudel0=glm(sugu2~1, family=binomial())
> summary(mudel0)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.24111    0.09333   -13.3   <2e-16 ***

    Null deviance: 702.54  on 659  degrees of freedom
Residual deviance: 702.54  on 659  degrees of freedom
AIC: 704.54
```

Näeme, et logistilise regressiooni mudel hindab arstiteaduskonna tudengite seast ühe tudengi juhuslikul väljanoppimisel meestudengi saamise šansiks 0.289:

```
> exp(-1.24111)
[1] 0.2890632
```

Erinevus viimastes komakohtades meie poolt arvutatud šansi ja logistilise regressioonmudeli poolt leitud šansi vahel tuleb sellest, et oleme kasutanud

vabaliikme ümmardatud väärtust (summary-käsk trükib välja vaid hinnatud parameetrite mõned esimesed tüvenumbrid…). Soovi korral võid arvutust korrata täpse vabaliikme väärtusega saamaks täpselt sedasama šanssi milleni jõudsime ise lihtsa arvutustehte abil:

```
sanss=exp(coef(mudel0))
sanss
```

Hinnatud šansi põhjal on võimalik leida hinnangut meestudengi saamise tõenäosusele:

```
p=sanss/(1+sanss)
p
```

Sama tõenäosust võiksime saada leides kas tunnuse sugu2 keskmise või vaadates meeste osakaalu sagedustabelis:

```
mean(sugu2)
prop.table(table(sugu2))
```

Antud meestudengi saamise tõenäosust (meestudengite osakaalu) võinuksime arvutada ka predict-käsu abil:

```
> predict(mudel0, data.frame(suvaline=1), type="response")
    1
0.2242424
```

Liigume sammukese keerulisemate mudelite poole. Oletame, et meil on võimalik tudengi soo prognoosimiseks kasutada lisainformatsiooni – näiteks seda, kui palju tudeng nädalas õlu joob. Hindame logistilise regressiooni mudeli, kus sõltumatu tunnusena (*independent variable, predictor variable*, *explaratory variable*) on sees tunnus *olu*:

```
> mudel1=glm(sugu2~factor(olu), family=binomial())
> summary(mudel1)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.2659     0.2101 -10.786  < 2e-16 ***
factor(olu)<1     0.6244     0.2681   2.329 0.019867 *
factor(olu)1-5    2.3965     0.2963   8.088 6.07e-16 ***
factor(olu)5-12   3.8754     0.5330   7.270 3.59e-13 ***
factor(olu)>13    4.0577     1.1004   3.688 0.000226 ***

    Null deviance: 702.54  on 659  degrees of freedom
Residual deviance: 560.73  on 655  degrees of freedom
AIC: 570.73
```

Hinnatud mudeli põhjal võime näiteks järeldada, et õlu mittetarbival tudengil on šansid olla meestudeng 0,1 (ligikaudu 1 meestudeng 10 naistudengi kohta):

```
> exp(-2.2659)
[1] 0.1037366
```

Ehk alkoholi mittetarbiva tudengi tõenäosus olla meestudeng on ligikaudu 0,09:

```
> exp(-2.2659)/(1+exp(-2.2659))
[1] 0.09398676
```

Samas näiteks 13 või enam pudelit nädalas tarbival tudengil on meestudengiks olemise šansid exp(4.0577)≈58 korda suuremad:

```
> sanss=exp(-2.2659+4.0577)
> sanss
[1] 6.000243
```

Ehk selliste tudengite seas on hinnanguliselt 6 meestudengit ühe naistudengi kohta. Teisendame leitud šansi ka õllelembeste tudengite seast meestudengi leidmise tõenäosuseks:

```
> p=sanss/(1+sanss)
> p
[1] 0.8571478
```

Ehk ligikaudu 85,7% enam kui 13 pudelit päevas tarbivatest arstiteaduskonna tudengitest on (hinnanguliselt) meestudengid.

Proovi samu tõenäosuseid leida ka predict-käsu abil. Kasutades predict-käsku täida ära järgmine tabel:

Nädala jooksul tarbitud õllepudelite arv	hinnanguline meestudengite osakaal
ei joo	…….
<1	…….
1-5	…….
5-12	…….
>13	…….

Võrdle saadud tulemusi kas tunnuse sugu2 keskmistega või meestudengite osakaaludega:

```
by(sugu2, olu, mean)
prop.table(table(sugu2[olu=="ei joo"]))
prop.table(table(sugu2[olu=="<1"]))
….
```

Kuidas on võrreldavad logistilise regressioonmudeli abil saadud „prognoosid“ ja vaatlusandmete põhjal leitud meestudengite osakaalud?

Proovime lisaks õlletarbimisele kasutada tudengi soo prognoosimisel ka tema pikkust. Hindame logistilise regressioonmudeli kus sõltuvate tunnustena on sees nii *pikkus* kui *olu*:

```
> mudel2=glm(sugu2~factor(olu)+pikkus, family=binomial())
> summary(mudel2)


Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -67.04729    6.34317 -10.570  < 2e-16 ***
factor(olu)<1     0.32900    0.38599   0.852 0.394019
factor(olu)1-5    1.79275    0.46553   3.851 0.000118 ***
factor(olu)5-12   3.88738    0.73421   5.295 1.19e-07 ***
factor(olu)>13    3.48836    1.52730   2.284 0.022371 *
pikkus            0.37251    0.03603  10.340  < 2e-16 ***

    Null deviance: 702.54  on 659  degrees of freedom
Residual deviance: 258.54  on 654  degrees of freedom
AIC: 270.54
```

Näeme, et 1 cm võrra pikemal tudengil on exp(0.37251)=1,45… korda suuremad šansid olla mees. Kontrolliks võime ka proovida mudelisse lisada pikkuse ruutu:

```
> mudel3=glm(sugu2~factor(olu)+pikkus+I(pikkus^2), family=binomial())
> summary(mudel3)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.118e+02  1.547e+02  -0.723 0.469951
factor(olu)<1    3.307e-01  3.845e-01   0.860 0.389715
factor(olu)1-5   1.792e+00  4.654e-01   3.850 0.000118 ***
factor(olu)5-12  3.909e+00  7.459e-01   5.240  1.6e-07 ***
factor(olu)>13   3.506e+00  1.550e+00   2.262 0.023675 *
pikkus           8.810e-01  1.755e+00   0.502 0.615688
I(pikkus^2)     -1.443e-03  4.975e-03  -0.290 0.771777

    Null deviance: 702.54  on 659  degrees of freedom
Residual deviance: 258.45  on 653  degrees of freedom
AIC: 272.45
```

Pikkuse ruude lisamine mudelisse ei osutunud heaks ideeks: mudeli AIC väärtus kasvas, ruutliikme ees olev kordaja pole statistiliselt oluliselt nullist erinev. Muuseas, keerukamat ja lihtsamat mudelit saab võrrelda ka anova-käsu abil:

```
anova(mudel2, mudel3, test="Chisq")
```

Mis praegusel juhul näitab samamoodi, et lihtsam mudel võiks töötada sama hästi kui keerukam (ja seega eelistame pigem lihtsamat mudelit).

Iseloomustame väljavalitud mudeli (mudel2) prognoose ka graafiliselt:

```
xx=seq(150,210, length=1000)
y1=predict(mudel2, data.frame(pikkus=xx, olu=">13"), type="response")
y2=predict(mudel2, data.frame(pikkus=xx, olu="ei joo"),
         type="response")

plot(xx, y1, type="l", col="slateblue", lwd=2, xlab="pikkus",
ylab="Meestudengite osakaal")
lines(xx, y2, col="skyblue", lwd=2)

legend(190, 0.4, c(">13", "ei joo"), lwd=2,
         col=c("slateblue", "skyblue"), title="Õlletarbimine")
```

Lisa graafikule kõverik ka 1-5 pudelit nädalas joovatele tudengitele (ning korrigeeri vastavalt ka joonise legendi). Kas saad aru, mida graafikul on kujutatud?

Hea oleks mudeli prognoosivõimet ka kuidagi kirjeldada. Teeme seda ROC-kõvera abil ja leiame ka AUC (ROC-kõvera alune pindala) väärtuse. Selleks kasutame lisamooduli Epi abi, mis tuleb eelnevalt arvutisse paigaldada (kuna te vaevalt olete eelnevalt seda lisamoodulit kasutanud):

```
install.packages("Epi")
library(Epi)
```

joonistame ROC-kõvera:

```
ROC(form=sugu2~factor(olu)+pikkus)
```

Näeme graafikult, et meie mudeli jaoks tuli AUC-väärtuseks 0,962 (mis on päris hea). Võrdluseks: kui perearst diagnoosib meestel eesnäärmevähki vereproovist mõõdetud PSA-väärtuse abil, siis vastava testi AUC-väärtuseks on väidetud olevat kõigest 0,7.

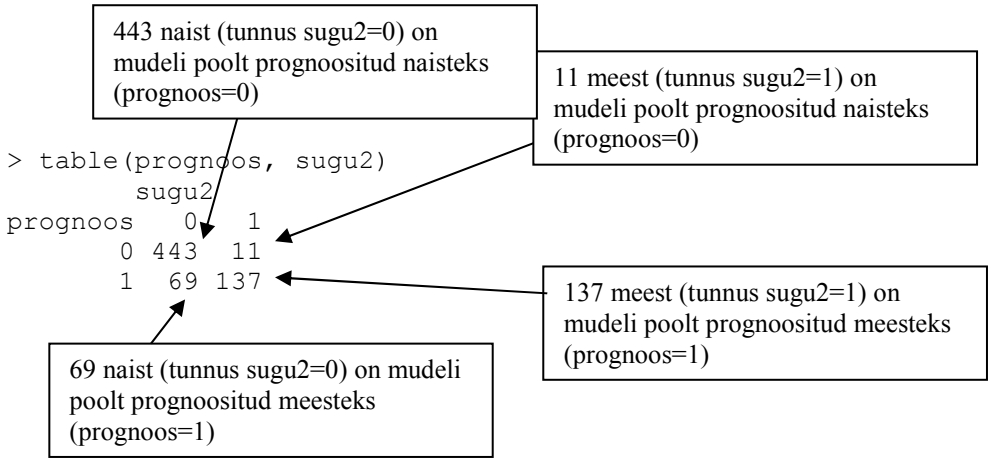
Vaatame kui hästi suudavad meie mudeli prognoosid jagada tudengeid meestudengiteks ja naistudengiteks. Leiame esmalt kõigi tudengite jaoks tõenäosuse, et nad on meestudengid (meie mudeli arvates):

```
p=predict(mudel2, tudengid, type="response")
```

Kui kellegi meheks olemise tõenäosus on suurem kui 0,155 siis „prognoosime“ ta meheks:

```
prognoos=1*(p>0.155)
```

Võrdleme prognoose ja tegelikke väärtuseid:



Näeme, et valitud otsustuskriteeriumi kasutades saavutame tundlikkuse 0,926 (meeste osakaal, kes ka mudeli poolt saadud prognoosi kohaselt on mehed – ehk kui suurele osale 1-dele anname prognoosiks ühe):

```
> 137/(137+11)
[1] 0.9256757
```

ja spetsiifilisuse (naiste osakaal keda mudel arvab olevat naised; ehk kui suurele osale tegelikult väärtusega „0“ vaatlustele annab mudel prognoosimise käigus prognoosiks nulli) 0,86:

```
> 443/(443+69)
[1] 0.8652344
```

Ülesanne 1

Milline tuleks spetsiifilisus ja tundlikkus siis, kui oleksime prognoosinud tunnust sugu järgmisel moel: kui tõenäosus olla mees on suurem kui 0,5 siis prognoosime tudengi meheks, kui tõenäosus olla mees on väiksem kui 0,5 siis prognoosime tudengi naiseks.

Tundlikkus: ……..

Spetsiifilisus: ……

Ülesanne 2
Proovi prognoosida, kas inimene tegeleb aktiivselt spordiga või mitte. Kasuta prognoosimiseks tudengi sugu ja kaalu. Millise mudelini jõuad? Visualiseeri saadud mudelit joonise abil!