Logistiline regressioon

praktikum

Logistilise regressiooni abil saab prognoosida binaarset (kahe võimaliku väärtusega) tunnust teiste tunnuste abil.

Tänases praktikumis kasutame Tartu Ülikooli tudengite andmestikku:

```
load(url("http://www-1.ms.ut.ee/mart/andmeteadus/tudengid.RData"))
tudengid[1:3,]
```

Üritame koostada mudelit, mis ennustaks tudengi sugu (1 -naine; 2 -mees). Logistilise regressiooni käsk R-is tahab aga, et uuritava tunnuse väärtused oleksid 0 või 1 (prognoositakse väärtuse 1 saamise tõenäosust või šanssi). Seega peame esmalt, enne logistilise regressioonmudeli hindamist, muutma oma tunnuse kodeeringut. Kodeerime tunnuse ümber selliselt, et mehed saaksid väärtuse 1 ja kõik ülejäänud (ehk naised) saaksid ümberkodeeritud tunnuse väärtuseks 0:

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

Ümberkodeerimise kontrolliks võid võrrelda esialgse tunnuse ja ümberkodeeritud tunnuse sagedustabeleid:

```
attach(tudengid)
table(sugu, 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 (logistiline mudel on sisuliselt mudel šansi logaritmi arvutamiseks):

```
> 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 ***
```

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 soovi korral võimalik leida hinnangut meestudengi kohtamise tõenäosusele arstiteaduskonnas:

```
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 (ära unusta lisaparameetrit `type="resp"`):

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

                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 ***
```

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

```
> sanss0=exp(-2.2659); sanss0
[1] 0.1037366
```

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

```
> sanss0/(1+sanss0)
[1] 0.09398676
```

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

```
> sanss_13=exp(-2.2659+4.0577)
> sanss_13
[1] 6.000243
```

Ehk selliste tudengite seas on hinnanguliselt 6 meestudengit ühe naistudengi kohta. Meheks olemise šansid on ohtralt õlut tarbivate tudengite seas exp(4.0577)=57,8 korda suuremad võrreldes õlut mittetarbivate tudengitega:

```
> sanss_13/sanss0
[1] 57.84112
```

Teisendame leitud šansi ka õllelembeste tudengite seast meestudengi leidmise tõenäosuseks:

```
> p=sanss_13/(1+sanss_13)
> p
[1] 0.8571478
```

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

Sama tõenäosust saab leida ka `predict`-käsu abil.

Tõenäosus olla mees (mudeli1 arvates) kui jood 13 või enam pudelit õlut nädalas:

```
predict(mudel1, data.frame(olu=">13"), type="resp")
```

Tõenäosus olla mees kui jood 0..1 pudelit õlut nädalas:

```
predict(mudel1, data.frame(olu="<1"), type="resp")
```

Võrdle saadud tulemusi kas tunnuse sugu2 keskmistega või meestudengite osakaaludega:

```
by(sugu2, olu, mean)
table(sugu2, olu)
prop.table(table(sugu2, olu),2)
```

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 – kui ta tarbiks sama palju õlut kui lühem tudeng.

Praegune mudel eeldab, et pideva tunnuse (*pikkus*) väärtuste kasvades saavad meid huvitava sündmuse toimumise šansid vaid kasvada (kui pideva tunnuse ees olev kordaja on positiivne, nagu siinses näites) või vaid kahaneda (kui pideva tunnuse ees olev kordaja on negatiivne). Vahel võivad aga meid huvitava sündmuse šansid olla kõige suuremad näiteks siis, kui pideva tunnuse väärtused on mingis kindlas vahemikus. Kuidas muuta logistilise regressiooni mudelit rikkamaks selliselt, et hinnatud mudel võimaldaks vajadusel ka sündmuse toimumise šanssidel hiljem kahaneda? Üheks lahenduseks oleks lisada mudelisse ka pideva tunnuse ruutliige. Proovime ka siin 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 ruudu lisamine mudelisse ei osutunud heaks ideeks: mudeli AIC väärtus kasvas, ka ruutliikme ees olev kordaja pole statistiliselt oluline (võib olla null). Muuseas, keerukamat ja lihtsamat logistilise regressioonanalüüsi mudelit saab võrrelda ka `anova`-käsu abil (tõepärasuhte test):

```
anova(mudel2, mudel3)
```

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="resp")
y2=predict(mudel2, data.frame(pikkus=xx, olu="<1"), type="resp")

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

legend("topleft", c("13...", " 0"), lwd=2,
        col=c("slateblue", "skyblue"), title="Õlletarbimine")
```

Lisa graafikule kõverik ka 1..5 pudelit nädalas (`olu="1-5"`) joovatele tudengitele (ning korrigeeri vastavalt ka joonise legendi).

Kas oskad saadud graafikut interpreteerida?

Kui hea on meie mudel?

Kui täpsed on logistilise regressiooni abil saadud prognoosid?
Sellele küsimusele vastamisel peame esmalt mõistma, et logistilise regressiooni mudel ei ütle meile iga uuritava kohta seda, kas ta on mees või mitte – vaid annab meile kõigest tõenäosuse (kui tõenäoliselt antud x-tunnuste väärtuste korral leiab aset meid huvitav sündmus). Kui hästi nende prognoositud tõenäosuste järgi on võimalik eristada tudengeid meesteks ja naisteks? Sellele küsimusele aitab vastata ROC-kõver:

```
library("pROC")
prognoos=predict(mudel2, newdata=tudengid,
      type="response")
roc_k6ver=roc(sugu2, prognoos)
plot(roc_k6ver, print.auc=TRUE,
print.auc.x=0.2,print.auc.y=0.05,
      print.thres=TRUE)
```

Näeme, et näiteks on võimalik saavutada tundlikkust 0,926 ja spetsiifilisust 0,867 kui loeme meesteks kõik need tudengid, kelle meheksolemise tõenäosus on suurem kui 0,167:

```
prognoosM=1*(prognoos>0.167)
addmargins(table(prognoosM, sugu2))
```

```
          sugu2
prognoosM   0   1 Sum
       0   444  11 455
       1    68 137 205
       Sum 512 148 660
```

Meestudengitest leidsime korrektselt üles 137 meest 148-st, seega tundlikkus = 0,926. Kokku 512-st naistudengist lahterdasime korrektselt 444 naiseks, seega spetsiifilisus on 444/512=0,867.

Veel üks täiendav joonis otsustuskriteeriumi visualiseerimiseks:

```
# Prognoositud tõenäosused olla mees tegelikult meeste ja
# tegelikult naiste jaoks:
par(mfrow=c(2,1))
hist(prognoos[sugu2==1], col="skyblue", main="Mehed")
hist(prognoos[sugu2==0], col="pink", main="Naised")
abline(v=0.167, xpd=NA)
```

Ülesanne 1

Prognoosi saadud mudeli abil omaenda sugu (pikkuse ja õlle tarbimise järgi). Milline tuleb sinu puhul meheksolemise tõenäosus? Kas see tuli suurem või väiksem kui valitud kriitiline piir? Kas meie mudeli arvates oled sa naine või mees?

Ülesanne 2

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: ……

Tudengite andmestiku saamiseks on küsitletud (~juhuslikult valitud) arstiteaduskonna tudengeid. Vahel kogutakse andmeid veidi teisiti. Näiteks võidakse küsitleda 100 naistudengit ja 100 meestudengit (märkus: usutavam mõne teise uuritava tunnuse puhul: küsitleti sadat autojuhti kes on eelmisel aastal avarii teinud ja sadat autojuhti kellega pole midagi juhtunud). Mis oleks siis olnud teisiti?

Mängime selle hüpoteetilise olukorra korra läbi. Nopime oma tudengite andmestikust välja 100 naist ja 100 meest (andmestikku valim2):

```
set.seed(1)
nr1 = 1:length(sugu2)
nr2 = c(sample(nr1[sugu2==0], 100), sample(nr1[sugu2==1], 100))
valim2=tudengid[nr2, ]
table(valim2$sugu2)
```

ja hindame logistilise regressiooni mudelid nii 100 meestudengi ja 100 naistudengi andmeid kasutades – ning kogu andmestiku pealt:

```
mudelA=glm(sugu2~pikkus, family=binomial(), data=tudengid)
mudelB=glm(sugu2~pikkus, family=binomial(), data=valim2)
```

Prognoosime 175cm pikkuse tudengi meheksolemise tõenäosust mõlema mudeli abil:

```
predict(mudelA, data.frame(pikkus=175), type="resp")
predict(mudelB, data.frame(pikkus=175), type="resp")
```

Kas prognoosid tulevad sarnased?

Mis võiks olla viltu?

Mis tüüpi uuringuga on tegemist, kui uurime sadat meestudengit ja sadat naistudengit?

Mida sellise uuringu puhul saab ja mida ei saa logistilise regressioonanalüüsi väljundist kasutada?

Kas tubli õlletarbija on pikk?

Loome tunnuse, mis näitab, kas tudeng kuulub 10% kõige pikemate tudengite sekka:

```
pikktudeng=1*(pikkus>quantile(pikkus, 0.9))
prop.table(table(pikktudeng))
```

Üritame prognoosida, kas tudeng kuulub kõige pikemate tudengite sekka või mitte:

```
m1=glm(pikktudeng~factor(olu), family=binomial())
```

Vaatame kas tarbitud õlle kogus aitab prognoosida seda, kas tudeng on pikk või mitte:

```
m0=glm(pikktudeng~1, family=binomial())
anova(m0, m1)
```

või, alternatiivina (LRT=Likelihood Ratio Test):

```
drop1(m1, test="LRT")
```

Veendusime eelnenu põhjal, et tarbitud õlle kogus aitab prognoosida seda, kas tudeng on pikk või mitte. Vaadates informatsiooni mudeli kohta

```
summary(m1)
```

Võime näha, et 13 või enam pudelit õlut nädalas ära joova tudengi šansid olla pikk on 17 korda suuremad kui õlut mittetarbival tudengil:

```
> exp(2.8557)
[1] 17.3866
```

NB! Antud mõju ei tohi tõlgendada kui põhjuslikku mõju – tegemist pole randomiseeritud katsega! Milline segav tunnus võib põhjustada seda, et rohkem õlut tarbivad tudengid on tõenäolisemalt ka pikad?

Vaata ka järgmist mudelit:

```
m2=glm(pikktudeng~factor(sugu)+factor(olu), family=binomial())
drop1(m2, test="LRT")
summary(m2)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -10.8824     1.4457  -7.528 5.17e-14 ***
sugu              5.2501     0.7435   7.061 1.65e-12 ***
factor(olu)<1     0.1556     0.4828   0.322    0.747
factor(olu)1-5    0.2435     0.4822   0.505    0.614
factor(olu)5-12  -0.1958     0.5716  -0.342    0.732
factor(olu)>13    0.3787     0.9043   0.419    0.675
```

Kuidas võiks selle mudeli korral tõlgendada tunnuse õlu väärtuse „>13“ mõju (0.3787)?