

Biomeetria 8. praktikum

Mitut sõltumatut tunnust sisaldav mudel.

Loeme sisse ja võtame kasutusele tudengite andmestiku

```
load(url("http://www.ms.ut.ee/mart/biomeetria2012/andmefail.RData"))
attach(kokku)
```

Tunneme huvi õllejoomise ja tudengi pikkuse vahelise seose vastu:

```
mudel1=lm(pikkus~factor(olu))
summary(mudel1)
drop1(mudel1, test="F")
```

Näeme, et õllejoomise ja inimese pikkuse vahel on olemas tõestatav (statistiliselt oluline) seos (p -väärtus $< 2.2e-16$). Mudeli parameetreid vaadates näeme, et enam õlut tarbivad inimesed kipuvad olema ka pikemad.

Lisame mudelile ka tunnuse sugu.

```
> mudel2=lm(pikkus~factor(olu)+factor(sugu))
> drop1(mudel2, test="F")
Single term deletions
```

```
Model:
pikkus ~ factor(olu) + factor(sugu)
              Df Sum of Sq  RSS   AIC  F value Pr(>F)
<none>                 23836 2379.2
factor(olu)      4      129.1 23965 2374.8    0.8855 0.4722
factor(sugu)     1     17262.9 41098 2736.8  473.6606 <2e-16 ***
```

Drop1-käsu tulemustest võime lugeda, et õlu (`factor(olu)`) pole pikkuse prognoosimisel enam vajalik (juhul kui teised prognoosimiseks kasutatavad muutujad jäävad mudelisse ehk kui teame inimese sugu, siis inimese õlletarbimine ei lisa enam pikkuse prognoosimiseks täiendavat informatsiooni).

Küll aitab aga inimese soo teadmine täpsustada ainult õlletarbimise pealt leitud pikkuse prognoosi (p -väärtus $< 2e-16$).

Informatsiooniks: vaata, millise drop1-käsu väljundi reaga kokkulangeva analüüsi saad järgmiste mudelite omavahelise võrdlemise abil:

```
mudelA=lm(pikkus~factor(sugu))
mudelB=lm(pikkus~factor(sugu)+factor(olu))

anova(mudelA, mudelB)
```

abil!

Hinnatud mudeli parameetrite tõlgendamine:

```
summary(mudel2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	168.02792	0.37514	447.904	<2e-16 ***
factor(olu)2	-0.19703	0.52584	-0.375	0.708
factor(olu)3	0.58049	0.78367	0.741	0.459
factor(olu)4	-1.76034	1.25775	-1.400	0.162
factor(olu)5	0.08445	2.36404	0.036	0.972
factor(sugu)2	14.11890	0.64873	21.764	<2e-16 ***

factor(olu)5 ees olev kordaja 0,08445 interpretatsioon: kui vaatame naisi, kes ei joo (olu=1) ja naisi, kes joovad olu suurtes kogustes (olu=5), siis nende kahe inimgrupi keskväärtuste erinevus on hinnanguliselt 0,08445 (aga erinevus võib olla ka 0, nagu näeme lisatud statistilise testi p-väärtusest). Sama erinevus on (antud mudeli arvates) ka olut mittetarbivatele meeste ja olut palju joovate meeste pikuste vahel.

Hinnatud mudeli põhjal võime leida ka näiteks prognoosi olut palju tarbivale (olu=5) mehele (sugu=2):

Prognoos = vabaliige + olletarbimise efekt + soo efekt

Prognoos = 168,02792 + 0,08445 + 14,1189 = 182,2313

Sama tulemuseni saame muidugi jõuda ka predict-käsku kasutades:

```
> predict(mudel2, newdata=data.frame(sugu=2, olu=5))
      1
182.2313
```

Sellises olukorras (õlletarbimise teadmisest pole tõestatavat kasu pikkuse prognoosimisel) eelistatakse enamasti lihtsamat mudelit – st. eemaldatakse mudelist õlletarbimist mõõtev tunnus.

Ülesanne 1

Uri summary(mudel2) käsu tulemusi ja proovi selgitada, mida näitab:

- Mudeli vabaliige (168.0279)
- Mudeli vabaliikme real toodud p-väärtus (<2e-16) – mida siin ikkagi on testitud?
- Real „factor(olu)2“ toodud hinnang (-0.19703) ja vastav p-väärtus (0.708) ?

Proovime (lihtsalt huvi pärast, sisulist põhjust pole) lisada mudelile ka soo ja õlletarbimise koosmõju:

```
model3 = lm(pikkus~factor(olu)+factor(sugu)+ factor(olu)*factor(sugu))
```

Uri ka järgmiste käskude abil saadud väljundit ja proovi vastata alltoodud küsimustele:

```
summary(mudel3)
drop1(mudel3, test="F")
anova(mudel2, model3)
```

Küsimused

- Mida näitab vabaliige (168.0058...)?

- Kuidas interpreteerida `real factor(sugu) 2` asuvat numbrit 14.354.. (sugu: 2-naine)?
- Mida võiks meile öelda `drop1`-käsu tulemus?

Koosmõjudega mudeli korral võib vahel kerkida üles ka küsimus, kas ikka üldse mõnda tunnust tarvis on pikkuse prognoosimiseks. Sellele küsimusele vastamiseks võime võrrelda kahte mudelit – ühte kus see tunnus on eemaldatud (pole ei peamõju ega koosmõju) ja mudelit, kus antud tunnus on sees millisel iganes vajalikul kujul:

```
anova(mudelA, mudel3)
```

Teema 2

Erinevalt õlletarbimisest on kaalu teadmises pikkuse prognoosimisel ikkagi kasu (isegi siis, kui teame juba inimese sugu):

```
model3=lm(pikkus~kaal+I(kaal^2)+factor(sugu))
drop1(model3, test="F")
summary(model3)
```

Joonistame ka vaatlused ja leitud regressioonmudeli. Joonistamisel valimi punktide värvid järgmiselt: kui tunnuse sugu väärtus on 1 (naine), siis võtame punkti värviks värvide vektorist 1. värvi, kui tunnuse sugu väärtuseks on aga 2 (mees), siis võtame värvide vektorist teise värvi („midnightblue“):

```
plot(kaal, pikkus, col=c("orange", "midnightblue")[sugu], pch=20)
```

Järgnevalt leiame pikkuste prognoosid inimestele, kelle kaalud on
40kg, 41kg, 42kg, ..., 120kg:

```
x=seq(40,120)
```

Leiame pikkuste prognoosid naistele (sugu=1)

```
y_naine=predict(model3, newdata=data.frame(kaal=x, sugu=1))
```

Leiame pikkuste prognoosid meestele (sugu=2)

```
y_mees=predict(model3, newdata=data.frame(kaal=x, sugu=2))
```

Lisame nii meeste kui naiste prognoosijooned joonisele

```
lines(x, y_naine, col="orange", lwd=2)
lines(x, y_mees, col="midnightblue", lwd=2)
```

Saadud joonisel paiknevad meeste regressioonijoon ja naiste regressioonijoon teineteisest alati sama kaugel (sama kaaluga mehed on sama kaaluga naistest alati 7.761cm pikemad. Antud regressioonimudel ei võimaldagi meeste-naiste erinevusel muutuda. Kasutatud regressioonimudel sunnib naiste-meeste erinevuse alati ühesuguseks.

Kui soovime lubada, et erinevate kaalude korral võiks meeste-naiste pikkuste erinevus ka muutuda, tuleb mudelisse lisada koosmõju (kaalu ja soo koosmõju). Koosmõju lubab soo efektil olla erinev iga tunnuse kaalu väärtuse korral (alternatiivne sõnastus: lubab meestel ja naistel erinevat kaalu-ees olevat kordajat). Kaalu ja soo koosmõju saame mudelisse lisada järgnevalt:

```
model4=lm(pikkus~kaal+I(kaal^2)+factor(sugu)+ factor(sugu)*kaal)
```

Drop1-käsku kasutade võime vaadata, kas koosmõju lisamine aitab mudeli prognoose parandada:

```
drop1(mudel4, test="F")
```

Single term deletions

Model:

```
pikkus ~ kaal + I(kaal^2) + factor(sugu) + factor(sugu) * kaal
              Df Sum of Sq  RSS   AIC F value    Pr(>F)
<none>                        15089 2066.9
I(kaal^2)          1   1121.83 16210 2112.0   48.402 8.481e-12 ***
kaal:factor(sugu)  1    458.52 15547 2084.6   19.783 1.020e-05 ***
```

Näeme, et kaalu ja soo koosmõju on statistiliselt oluline (1.020e-05), seega koosmõju lisamine mudelisse parandas oluliselt mudeli prognoosivõimet. Joonista ka uue mudeli jaoks regressioonsirged nii meeste kui naiste jaoks. Interpreteeri, mida näed.

Pane kirja, milline on antud mudeli korral naistudengite kaalu ja pikkuse vaheline seos ja milline on see seos meestudengite jaoks.

Selgita, kas leitud mudeli juures on midagi imelikku (ja kuidas võiks saadud mudelit ehk veelgi parandada).