Biomeetria bioloogidele

Kordamisküsimused ja näidisülesanded

Kordamisküsimused

Tudeng, kes soovib kursuse edukalt ja positiivse hindega läbida peaks:

- Mõistma, mis on statistiline seos;
- Lineaarse regressioonsirge kaks võimalikku interpretatsiooni (sirge mis minimiseerib prognoosivigade ruutude summa; normaaljaotuse keskväärtuse muutumist kirjeldav sirge mille korral on nähtud valimi saamise tõenäosus kõige suurem);
- Lineaarse regressioonanalüüsi tulemuste interpreteerimine: mudeli parameetrite (vabaliikme ja sirge tõusu) tähendus; mida näitab usaldusintervall, mida prognoosiintervall; mida näitab ja milleks saab kasutada determinatsioonikordajat;
- Lineaarse regressioonmudeli eeldused. Mudeli eelduste kontrollimine – kuidas on võimalik avastada vale kujuga mudelit, kuidas saab leida üles erindit. Peaks tundma ühte-kahte võimalust kuidas saab lähendada tundmatut (mittelineaarset) seost tunnuste vahel.
- Dispersioonanalüüsi mudel. Parameetrite interpretatsioon.
- Mudeli parameetrite interpretatsioon mudelis, kus on enam kui üks tunnus. Multikollineaarsus.
- Kahe tunnuse koosmõju, koosmõju interpretatsioon.
- Mudeli valikust. Paremini prognoosiva mudeli otsimine; kuidas otsida põhjuslikke mõjusid kirjeldavat mudelit.
- Poissoni regressioon. Mudeli parameetrite interpreteerimine, eeldused. Mudeli *offset*-milleks kasutatakse, kuidas kasutatakse.
- Mis on üle- või alahajuvus, kuidas võib ülehajuvus (või alahajuvus) tekkida. Kuidas on võimalik ülehajuvuse probleemi tuvastada uurides näiteks Poissoni regressioonmudeli väljundit. Mudelid üle- või alahajuvusega andmetele (kvaasi-Poissoni mudel, negatiivsel binoomjaotusel baseeruv mudel). Milles seisneb nende mudelite erinevus Poissoni regressioonist? Kvaasi-poissoni ja negatiivsel binoomjaotusel baseeruvate mudelite parameetrite interpreteerimine.
- Logistiline regressioon. Parameetrite interpreteerimine (nii faktortunnuse kui ka pideva tunnuse korral). Mõisted šanss ja šansside suhe. Kuidas logistilise regressiooni väljundi põhjal leida sündmuse toimumise šansse ja sündmuse toimumise tõenäosust.

Eksamil on lubatud kasutada kirjalikke mateeriale (väljatrükitud slaide, oma loengukonspekti, praktikumides väljajagatud materiale, raamatuid vms). Ei ole lubatud kasutada sülearvuteid, rüperaale, nutiseadmeid ja telefone – nende seadmete kasutamise korral on õppejõul liiga raske kontrollida välise abi puudumist.

Soovitatav on kaasa võtta selline taskuarvuti, mille abil saab leida kas logaritmi või eksponentfunktsiooni väärtuseid.

Näidisülesanded

Ülesanne 1

Vaata alltoodud kolme hinnatud mudelit. Kõigi mudelite korral vasta järgmistele küsimustele:

a) Milline tuleb hinnang uuritava tunnuse y keskväärtusele siis, kui x=1?
b) Kuidas muutub uuritava tunnuse keskväärtus siis kui sõltumatu tunnuse väärtus muutub ühe ühiku võrra?

Mudel 1:

```
> summary(lm(y~x))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.00000    0.04414   22.33   <2e-16 ***
x            1.00000    0.04696   19.87   <2e-16 ***
```

Mudel 2:

```
> summary(glm(y~x, family=poisson()))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0000     0.0348   29.67   <2e-16 ***
x             1.0000     0.0153   64.77   <2e-16 ***
(Dispersion parameter for poisson family taken to be 1)
```

Mudel 3:

```
> summary(glm(y~x, family=binomial()))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.7956     0.2295   3.467 0.000527 ***
x             1.0956     0.1991   5.501 3.77e-08 ***

(Dispersion parameter for binomial family taken to be 1)
```

Ülesanne 2

Selgita iga testitud efekti tähendust. Kirjelda oma sõnadega, milline on siis seos sõltumatute tunnuste ja sõltuva tunnuse vahel!

Lisainformatsioon: faktortunnusel maakond on kolm taset (A, B, C).

```
> m1=lm(Y~maakond+vanus+vanus*maakond)
> summary(m1)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     14.9818    17.8697   0.838    0.407
maakondB         3.8838    17.8871   0.217    0.829
maakondC         4.3760    17.8871   0.245    0.808
vanus            0.5486     1.6999   0.323    0.749
maakondB:vanus  19.5471     1.7012  11.490 6.27e-14 ***
maakondC:vanus   9.4684     1.7012   5.566 2.24e-06 ***

Residual standard error: 1.7 on 38 degrees of freedom
Multiple R-squared: 0.9998,     Adjusted R-squared: 0.9998
F-statistic: 3.704e+04 on 5 and 38 DF,  p-value: < 2.2e-16
```

Ülesanne 3.

Uuritakse puude kõrgust (*Y*-tunnuseks on puu kõrgus meetrites). Paku välja kaks tunnust, millel võiks olla mõju puu kõrgusele. Paku välja kaks mudelit, üks milles on sees väljapakutud tunnuste vaheline koosmõju ja teine milles koosmõju pole. Seleta, milles seisneb nende kahe mudeli vaheline erinevus. Kas sinu arvates on vaja puu kõrguse modelleerimisel kasutada ka väljapakutud tunnuste vahelist koosmõju või mitte, põhjenda oma otsust.

Ülesanne 4

Uuritakse autoõnnetuste arvu sõltuvust ilmastikutingimustest ja muudest asjassepuutuvatest tunnustest. Sõltuvuste kirjeldamiseks kasutatakse Poissoni regressiooni.

Kirjelda kuidas andmeid võiks koguda ja mida antud andmete kogumise viisi puhul võiks kasutada offset-muutujana. Millisel viisil kogutud andmete puhul aga poleks offset-muutuja kasutamine üldse vajalik?