Tilastotieteen harjoitustyö 2021

TILM3558

Marko Järvinen
518467
mabeja@utu.fi

1. Numeeristen vastemuuttujien mallitus

Satunnaisotos (700 otosta) valittu käyttämällä opiskelijanumeroa siemenlukuna

`elinolo.sav` (Tilastokeskuksen elinolotutkimuksen aineisto, N=2199)

```
library(foreign)
ht1.dat <- read.spss("elinolo2020.sav", to.data.frame = TRUE)
attach(ht1.dat)

set.seed(518467)
oma.otos1 <- ht1.dat[sample(nrow(ht1.dat), 700),]
attach(oma.otos1)
```

1.1. Varianssianalyysi

> *Tutki, onko sukupuolella ja asumisahtaudella yhteyttä asunnon pinta-alaan.*

Käytetään kaksisuuntaista varianssianalyysiä, koska kyseessä on yksi numeerinen selitettävä muuttuja ja kaksi kategorista selittävää muuttujaa.

Normaalijakaumaoletus

Jaetaan aineisto sukupuolen mukaan kahteen osaan:

```
library(dplyr)
miehet.dat <- select(filter(oma.otos1, supu=="mies" ), c(supu, ahtas, pala))
naiset.dat <- select(filter(oma.otos1, supu=="nainen" ), c(supu, ahtas, pala))
```

Tehdään Shapiro-Wilk -testit:

```
with(miehet.dat, tapply(pala, list(ahtas), shapiro.test))
with(naiset.dat, tapply(pala, list(ahtas), shapiro.test))
```

Testien tulosteiden perusteella kaikkien luokkien p-arvot eivät ylitä rajaa 0,5, joten normaalijakaumaoletus ei ole voimassa. Parametrista testiä voidaan kuitenkin käyttää

havaintojen suuresta määrästä johtuen. Myöskin havainnoista luotujen laatikko-jana-kuvioiden perusteella ne ovat normaalisti jakautuneita.

Tehdään kaksisuuntainen varianssianalyysi:

```
attach(oma.otos1)
library(car)
anova(lm(pala~supu*ahtas))
```

Kaksisuuntaisen varianssianalyysin perusteella sekä sukupuoli, että asumisahtaus ovat tilastollisesti merkitseviä selittäjiä, sillä niiden F-testin p-arvot ovat alle 0,001. Sen sijaan niiden yhdysvaikutuksen F-testin p-arvo on yli 0,001 eikä siksi ole tilastollisesti merkitsevä selittäjä.

1.2. Regressiomalli

> *Tutki, onko kotitalouden kuluttajayksiköiden lukumäärällä, asumismenoilla yhteensä ja alueella asumisajalla yhteyttä asunnon pinta-alaan.*

Tehdään regressioanalyysi, jossa selittäjinä ovat kotitalouden kuluttajayksiköiden lukumäärä, asumismenot ja alueella asumisaika sekä selitettävänä asunnon pinta-ala.

Tehdään sirontakuviot:

```
# sirontakuvio, kotitalouden kuluttajayksiköiden lukumäärä
plot(rkyks, pala)
abline(lm(pala~rkyks))

# sirontakuvio, asumismenot
plot(asmenot, pala)
abline(lm(pala~asmenot))

# sirontakuvio, alueella asumisaika
plot(alaika, pala)
abline(lm(pala~alaika))
```

Sirontakuvioista päätellen kotitalouden kuluttajayksiköiden lukumäärän, asumismenojen ja alueella asumisajan kasvaessa asunnon pinta-alakin kasvaa.

Lasketaan Pearsonin ja Spearmanin korrelaatiokertoimet:

```
cor.test(rkyks, pala, method="pearson")
cor.test(rkyks, pala, method="spearman", exact=FALSE)
cor.test(asmenot, pala, method="pearson")
cor.test(asmenot, pala, method="spearman", exact=FALSE)
cor.test(alaika, pala, method="pearson")
cor.test(alaika, pala, method="spearman", exact=FALSE)
```

Pearsonin korrelaatiokerroin kotitalouden kuluttajayksiköiden lukumäärän ja asunnon pinta-alalle on n. 0,534, joka osoittaa merkittävää suoraviivaista riippuvaisuutta. Spearmanin kerroin on vastaavasti n. 0,588, joka myös osoittaa merkittävää suoraviivaista riippuvuutta.

Pearsonin korrelaatiokerroin asumismenojen ja asunnon pinta-alalle on n. 0,118 ja Spearmanin kerroin on vastaavasti n. 0,072. Yhteyksiä näiden muuttujien välillä ei voida pitää suoraviivaisina.

Pearsonin korrelaatiokerroin alueella asumisajan ja asunnon pinta-alalle on n. 0,125 ja Spearmanin kerroin on vastaavasti n. 0,203. Yhteydet näiden muuttujien välillä ovat suoraviivaisia.

Kolmen selittäjän regressiomalli:

```
lm.pala <- lm(pala~rkyks+asmenot+alaika)
summary(lm.pala)
```

Mallin regressioyhtälöksi saadaan 15,06 + 26,97 · kotitalouden kuluttajayksiköiden lukumäärä + 0,00052 · asumismenot + 0,37 · alueella asumisaika = asunnon pinta-ala.

1.3. Toistomittausmalli

Satunnaisotos (600 otosta) valittu käyttämällä opiskelijanumeroa siemenlukuna

`Toistomittausaineisto2020.sav` (7 maasta kerätty aineisto potilaan ohjauksesta, N=1299)

```
library(foreign)
ht2.dat<-read.spss("Toistomittausaineisto2020.sav", to.data.frame=TRUE)
attach(ht2.dat)

set.seed(518467)
oma.otos2<-ht2.dat[sample(nrow(ht2.dat), 600), ]
attach(oma.otos2)
```

Tutkijalla on hypoteesi, että potilaan mielestä saatu ohjaus leikkauksen jälkeen toiminnallista seikoista (`Functional_M2`) on ollut vähäisempää kuin odotettu ennen leikkausta (`Functional_M1`). Eli keskiarvo toisessa mittauksessa on matalampi. Lisäksi kiinnostaa se, onko tuo ero mittausten välillä erilainen sukupuolittain.

> *Tutki saavatko nämä tutkimushypoteesit tukea mallittamalla aineisto toistettujen mittausten varianssianalyysillä.*

Molemmat sukupuolet

Suoritetaan toistettujen mittausten varianssianalyysi. Tarkistetaan ensin onko normaalijakaumaoletus voimassa Shapiro-Wilk -testeillä:

```
shapiro.test(Functional_M2)
shapiro.test(Functional_M1)
```

Sekä `Functional_M2`:n ja `Functional_M1`:n p-arvot ovat alle 0,001. Tästä voidaan todeta, että normaalijakaumaoletus ei ole voimassa.

Koska normaalijakaumaoletus ei ole voimassa, käytetään epäparametrista testiä.

Järjestetään havainnot Friedmanin testiä varten:

```
library(dplyr)
filtered_data <- na.omit(select(oma.otos2, patient, Functional_M1, Functional_M2, D2))
filtered_data <- with(filtered_data, filtered_data[order(patient), ])
filtered_data <- filtered_data[!duplicated(filtered_data$patient), ]

M1_patient <- list(select(filtered_data, patient))
M1_value <- list(select(filtered_data, Functional_M1))
M1_Sex <- list(select(filtered_data, D2))
M1 <- do.call(rbind.data.frame, Map('c', M1_patient, M1_value, M1_Sex))
M1['Functional'] = 'M1'
names(M1)[names(M1) == 'Functional_M1'] <- 'Functional_Value'

M2_patient <- list(select(filtered_data, patient))
M2_value <- list(select(filtered_data, Functional_M2))
M2_Sex <- list(select(filtered_data, D2))
M2 <- do.call(rbind.data.frame, Map('c', M2_patient, M2_value, M2_Sex))
M2['Functional'] = 'M2'
names(M2)[names(M2) == 'Functional_M2'] <- 'Functional_Value'

data <- rbind(M1, M2)
```

Tehdään Friedmanin testi:

```
attach(data)
friedman.test(Functional_Value ~ Functional | patient, data=data)
detach(data)
```

Friedmanin testin p-arvo on alle 0,001, joten muuttujien välillä on tilastollisesti merkitseviä eroja. Tehdään muuttujien välinen vertailu Wilcoxonin testillä:

```
attach(oma.otos2)
wilcox.test(Functional_M1, Functional_M2, paired = TRUE)
```

Testin tuloksen mukaan mittausten välinen ero on tilastollisesti merkitsevä, sillä testin p-arvo on alle 0,001.

Sukupuolet erikseen

Tehdään toistettujen mittausten varianssianalyysi, jossa luokitteleva tekijä on sukupuoli (`D2`).

```
attach(data)
summary(aov(Functional_Value ~ D2 * Functional + Error(patient / Functional), data=data))
detach(data)
```

Tulosteesta huomataan, että sukupuolien välillä ei ole tilastollisesti merkitseviä eroja (p=0,628). Sen sijaan mittausten väliset erot ovat tilastollisesti merkitseviä (p<0,001). Myöskään näiden välinen yhdysvaikutus ei ole tilastollisesti merkitsevää (p=0,642).

2. Kategoristen vastemuuttujien mallitus

Satunnaisotos (900 otosta) valittu käyttämällä opiskelijanumeroa siemenlukuna

`EK2011.sav` (eduskuntavaaliaineisto vuodelta 2011, N=1318)

```
library(foreign)
ht3.dat<-read.spss("EK2011.sav", to.data.frame=TRUE)
attach(ht3.dat)

set.seed(518467)
oma.otos3<-ht3.dat[sample(nrow(ht3.dat), 900), ]
attach(oma.otos3)
```

2.1. Muuttujien riippuvuusrakenne

1. **Tarkastellaan muuttujia sukupuoli (`d2`), työttömyys viimeisen 12 kuukauden aikana (`d32`) ja oman sukupuolen 2011 eduskuntavaaleissa äänestäminen (`k23`). Tee ensin yksiulotteiset frekvenssijakaumat ja kolmen muuttujan ristiintaulu.**

   > *Onko 3-ulotteisessa ristiintaulussa nollasoluja?*

   Tehdään frekvenssitaulukot:

   ```
   table(d2)
   table(d32)
   table(k23)
   ```

   Huomataan, että puuttuvia arvoja ovat:

   - sukupuoli (`d2`): 15 arvoa
   - työttömyys viimeisen 12 kuukauden aikana (`d32`): 5 arvoa arvoa
   - oman sukupuolen 2011 eduskuntavaaleissa äänestäminen (`k23`): 141 arvoa

   Kolmen muuttujan ristiintaulu:

   ```
   ftable(table(d2, d32, k23))
   ```

Ristiintaulussa ei ole nollasoluja. Pienin solufrekvenssi on 17.

2. **Tarkastele kolmen muuttujan välisiä riippuvuuksia loglineaaristen mallien avulla. Ota mukaan muuttujista vain ne luokat, joissa havaintoja on yli 10.**

Loglineaarinen mallitus:

```
library(MASS)
mytable <- xtabs(~ d2 + d32 + k23, data=oma.otos3)
malli <- loglm(~ d2 + k23 + d32 + d2*k23+d32, mytable)
malli
```

> *Millaiset riippuvuudet muuttujien välillä askeltavan menetelmän avulla valittuun malliin jäivät?*

Sukupuoli ja oman sukupuolen äänestäminen (`d2` ja `k23`) jäivät riippuvaisiksi.

> *Mikä on mallin generoiva luokka?*

Mallin generoiva luokka on { `d2` * `k23` + `d32` }.

> *Mikä on mallin yhteensopivuustestin p-arvo?*

Mallin yhteensopivuustestin p-arvo on n. 0,067.

> *Mikä on standardoitujen jäännösten vaihteluväli?*

```
stdres = residuals(malli, "pearson")
summary(stdres)
```

Standardoitujen jäännösten vaihteluvälin alaraja on -1,871 ja yläraja 1,493.

3. **Tee mallin mukainen yhteyksien jatkotarkastelu ristiintauluin ja tulkitse malli riviprosenttien avulla.**

Tehdään jatkotarkastelu ristiintauluin:

```
taulu1 <- table(d2, k23)
prop.table(taulu1, 1)
taulu2 <- table(d32, k23)
prop.table(taulu2, 1)
```

Huomataan, että miehet äänestävät omaa sukupuoltaan useammin kuin naiset. Huomataan myöskin, että viimeisen 12 kk:n aikana työttömänä olleet äänestävät

omaa sukupuoltaan harvemmin kuin ne jotka eivät ole olleet työttömänä viimeisen 12 kk:n aikana.

2.2. Kaksiluokkainen selitettävä muuttuja

1. **Tutki muuttujien sukupuoli (`d2`) ja ikä yhteyttä työttömyyteen viimeisen 12 kuukauden aikana (`k32`) käyttämällä logistista regressiomallia.**

   Logistinen binäärinen regressio:

```
tyottomyys <- glm(d32 ~ d1 + d2, data=oma.otos3, family=binomial)
summary(tyottomyys)
```

2. > *Mitkä muuttujat selittävät työttömyyttä ?*

   Huomataan, että ikä on tilastollisesti merkitsevä tekijä ($p < 0{,}001$). Sen sijaan sukupuoli ei ole tilastollisesti merkitsevä tekijä ($p = 0{,}276$).

   > *Tulkitse yhteydet OR:ien avulla. Raportoi myös luottamusvälit OR:ille*

   Selitetään "*työtön viimeisen 12 kk:n aikana*" kyllä/ei -suhdetta:

```
exp(cbind(OR=coef(tyottomyys), confint(tyottomyys)))
```

   Muuttujan `d1` (ikä) OR on n. 1,04. Muuttuja `d2` (sukupuoli) ei ollut tilastollisesti merkitsevä tekijä, minkä vuoksi sitä ei tarkastella.

   OR:n 95% luottamusvälin alaraja on n. 1,03 ja yläraja puolestaan n. 3,41.

   > *Mikä on mallin Nagelkerke selitysaste ?*

```
install.packages("fmsb")
library(fmsb)
data.nagel <- NagelkerkeR2(tyottomyys)
data.nagel
```

   Mallin *Nagelkerke* selitysaste on n. 0,092.

3. Monimuuttujamenetelmät

Satunnaisotos (1500 otosta) valittu käyttämällä opiskelijanumeroa siemenlukuna

`pankkiotos2020.sav` (todellinen asiakasaineisto, N=2453)

```
library(foreign)
ht4.dat<-read.spss("pankkiotos2020.sav", to.data.frame=TRUE)
attach(ht4.dat)

set.seed(518467)
oma.otos4<-ht4.dat[sample(nrow(ht4.dat), 1500), ]
attach(oma.otos4)
```

3.1. Muuttujien ryhmittely

1. **Muodosta pääkomponenttianalyysilla luokitelluista muuttujista (41 kpl: autom_lainan_perinta_luok - kulutusluotot1_luok) pääkomponentteja ominaisarvokriteerin mukaan. (Promax-rotaatio)**

   Poistetaan muuttuja `toimeksianto_a_kpl_luok`, sillä sen kaikki arvot ovat 0, mikä johtaa koodia ajaessa virheeseen:

   ```
   drops <- c("toimeksianto_a_kpl_luok")
   oma.otos5 <- oma.otos4[ , !(names(oma.otos4) %in% drops)]
   ```

   Tehdään pääkomponenttianalyysi:

   ```
   # korrelaatiokertoimet
   data.kor <- cor(oma.otos5, method="pearson", use="complete.obs")

   # pääkomponenttianalyysi
   pca <- prcomp(data.kor, center=TRUE, scale=TRUE)
   summary(pca)
   ```

   Nähdään että pääkomponentteja muodostui 10 kappaletta kun ominaisarvokriteeri on se että ominaisarvo on suurempi kuin 1.

   Promax-rotaatio:

   ```
   pca.chosen <- pca$rotation [, 1:10]
   pca.promax <- promax(pca.chosen)
   pca.promax
   ```

2. **Talleta havaintomatriisiin uusiksi muuttujiksi pääkomponenttipistemäärät.**

   Pääkompontit tallennettiin yllä olevalla koodikatkelmalla.

3. **Nimeä uudet muuttujat (pääkomponentteihin latautuneiden muuttujien mukaisesti).**

   - PC1: **Tilin aktiivisuus** - panot ja otot jokseenkin latautuneita
   - PC2: **Maksupalvelu-/tiski-maksuaktiivisuus** - maksupalvelu- ja tiskimaksut latautuneita

- PC3: **Rahastoaktiivisuus** - rahastoihin liittyvät muuttujat latautuneita
- PC4: **Laina-aktiivisuus** - lainoihin liittyvät muuttujat latautuneita
- PC5: **Luottokorttien lkm. / vakuutusaktiivisuus** - sekä vakuutus-, että luottokorttien lkm. -muuttujat latautuneita
- PC6: **Osakeaktiivisuus** - osakkeisiin liittyvät muuttujat latautuneita
- PC7: **Kouluttautuneisuus** - korkeakoulutus-muuttuja latautunut
- PC8: **Käyttötilin velkaisuus** - käyttötilin velka -muuttuja latautunut
- PC9: **Asuntolaina-/rahastoaktiivisuus** - sekä asuntolaina- (b), että rahasto (a1) -muuttujat latautuneita
- PC10: **Toimeksianto/kansainväliset maksukortit** - sekä toimeksianto- (b), että kv. maksukortit -muuttujat latautuneita

3.2. Havaintojen ryhmittely

4. **Käytä näitä uusia muuttujia klusterianalyysissä, jossa muodostat asiakasryhmiä K-means menetelmällä lähtien kahdesta klusterista viiteen tai kuuteen klusteriin saakka. Kuvaile muodostamiasi ryhmiä.**

   Tehdään klusterianalyysi:

```
km = pca.chosen[,c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)]
```

   Kuvaillaan muodostuneita ryhmiä:

k = 2

```
set.seed(518467)
km2 = kmeans(km, 2, nstart=100)
km2
```

Klusteri	Havaintoja	Profiloivat pääkomponentit
1	14	◦ PC2 (Maksupalvelu-/tiski-maksuaktiivisuus) ◦ PC3 (Rahastoaktiivisuus) ◦ PC4 (Laina-aktiivisuus) ◦ PC5 (Luottokorttien lkm. / vakuutusaktiivisuus) ◦ PC7 (Kouluttautuneisuus) ◦ PC8 (Käyttötilin velkaisuus) ◦ PC9 (Asuntolaina-/rahastoaktiivisuus) ◦ PC10 (Toimeksianto/kansainväliset maksukortit)

Klusteri	Havaintoja	Profiloivat pääkomponentit
2	26	◦ PC1 (Tilin aktiivisuus) ◦ PC6 (Osakeaktiivisuus)

k = 3

```
set.seed(518467)
km3 = kmeans(km, 3, nstart=100)
km3
```

Klusteri	Havaintoja	Profiloivat pääkomponentit
1	5	*Matalat arvot*
2	8	*Matalat arvot*
3	27	◦ PC1 (Tilin aktiivisuus) ◦ PC5 (Luottokorttien lkm. / vakuutusaktiivisuus)

k = 4

```
set.seed(518467)
km4 = kmeans(km, 4, nstart=100)
km4
```

Klusteri	Havaintoja	Profiloivat pääkomponentit
1	22	◦ PC2 (Rahastoaktiivisuus)
2	10	*Matalat arvot*
3	3	*Matalat arvot*
4	5	*Matalat arvot*

k = 5

```
set.seed(518467)
km5 = kmeans(km, 5, nstart=100)
km5
```

Klusteri	Havaintoja	Profiloivat pääkomponentit
1	8	*Matalat arvot*
2	19	◦ PC1 (Tilin aktiivisuus)
3	5	*Matalat arvot*
4	5	*Matalat arvot*
5	3	*Matalat arvot*