Tilastotieteen harjoitustyö 2018

Samuli Virtanen

spevir@utu.fi (mailto:spevir@utu.fi)

op. nro. 511178

Tulostedokumentin tulosteet ovat tehtäviä vastaavassa järjestyksessä lukuunottamatta osaa 1.3. Toistomittausaineisto, jonka tulosteet löytyvät tulostedokumentin lopusta.

1. Numeeristen vastemuuttujien mallinnus

- Aineisto: Elinolo2018 (Tilastokeskuksen elinolotutkimuksen aineisto, N=2199)
- Datasta poimittu 600 kokoinen satunnaisotos käyttämällä omaa opiskelijanumeroa satunnaislukugeneraattorin siemenlukuna, tehtävän ohjeiden mukaisesti.

1.1. Varianssianalyysi

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

1.1.1. Suunnitelma

Kyseessä on kaksi kategorista selittävää muuttujaa, sekä yksi numeerinen selitettävä muuttuja, joten käytetään kaksisuuntaista varianssianalyysiä.

1.1.2. Normaalijakaumaoletus

Ensimmäiseksi tutkitaan jakaumia sukupuolittain asumisahtauden perusteella.

Jaetaan data kahteen osaan sukupuolen perusteella.

```
SORT CASES BY supu.
SPLIT FILE SEPARATE BY supu.
```

Luodaan tarvittavat kuviot ja tiedot SPSS:n Explore-työkalulla.

```
EXAMINE VARIABLES=pala BY ahtas
  /PLOT BOXPLOT NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Tulostedokumentista löytyvän Shapiro-Wilk-testin tulosten perusteella neljä ryhmää ei ole normaalisti jakautunut:

- mies * normaali

- mies * tilava
- nainen * normaali
- nainen * tilava

Näiden ryhmien havaintojen suuresta määrästä johtuen voidaan kuitenkin käyttää parametrista testiä. Myöskin visuaalinen tarkastelu osoittaa ryhmien olevan jokseenkin normaalijakautuneita. Kahden jäljelle jäävän ryhmän havaintojen määrä on pieni, mutta ne ovat testin mukaan normaalijakautuneita.

1.1.3. Kaksisuuntainen varianssianalyysi

Tehdään kaksisuuntainen varianssianalyysi SPSS:n Univariate-työkalulla.

```
SPLIT FILE OFF.

UNIANOVA pala BY supu ahtas
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PLOT=PROFILE(ahtas*supu)
  /PRINT=DESCRIPTIVE HOMOGENEITY
  /CRITERIA=ALPHA(.05)
  /DESIGN=supu ahtas supu*ahtas.
```

Tuloksista nähdään että selittävien muuttujien päävaikutukset ovat merkitseviä merkitsevyystasolla $p < 0,05$, mutta yhdysvaikutus ei ole merkitsevä. ($p \sim 0.638$)

Myöskään hajontojen yhtäsuuruusoletus ei näytä pitävän paikkaansa (Levenen testin p-arvo ~ 0.014)

Voidaan tulkita että miehillä on keskimäärin suuremmat asunnot kuin naisilla, ja asunnon pinta-ala kasvaa koetun asumisahtauden mukaan järjestyksessä ahdas -> normaali -> tilava (pienestä pinta-alasta suurempaan). Koska yhdysvaikutuksella ei ole merkitystä, voidaan tulkita että sukupuolella ei ole väliä asumisahtauden eroissa.

1.1.4. Monivertailut

Tutkitaan vielä monivertailut estimoiduilla keskiarvoilla.

```
UNIANOVA pala BY supu ahtas
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PLOT=PROFILE(ahtas*supu)
  /EMMEANS=TABLES(supu) COMPARE ADJ(SIDAK)
  /EMMEANS=TABLES(ahtas) COMPARE ADJ(SIDAK)
  /EMMEANS=TABLES(supu*ahtas)
  /PRINT=ETASQ DESCRIPTIVE HOMOGENEITY
  /CRITERIA=ALPHA(.05)
  /DESIGN=supu ahtas supu*ahtas.
```

Tulostedokumentista löytyvän *Pairwise comparisons* -taulukon perusteella voidaan todeta kaikkien asumisahtaustasojen välillä olevan tilastollisesti merkittäviä eroja. ($p < 0.001$)

1.2. Regressiomalli

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

Ratkaisu: Sovitetaan regressiomallia, jossa selittäjinä ovat on kuluttajayksiköiden lukumäärä, asumismenot yhteensä sekä alueella asumisaika. Selitettävänä muuttujana on asunnon pinta-ala.

Tutkitaan sirontakuvioita, joka löytyy tulostedokumentista.

```
GRAPH
  /SCATTERPLOT(MATRIX)=rkyks asmenot alaika pala
  /MISSING=LISTWISE.
```

Alueella asumisaika näyttäisi kasvaessaan vaikuttavan pienentäväsi asumismenoihin. Muiden selittävien muuttujien välistä korrelaatiota ei visuaalisesti tarkasteltuna esiinny, jokaisen selittvän muuttujan välinen sirontakuvio on siis suunnilleen ellipsin muotoinen.

Kovin selkeitä korrelaatioita ei esiinny. Näyttäisi kuitenkin hieman siltä että kodin kuluttajayksiköiden määrän kasvaessa myös asunnon pinta-ala kasvaa ja toisaalta samoin käy myös asumismenojen kasvaessa. Alueella asumisaika ei sen sijaan vaikuta vaikuttavan asunnon pinta-alaan.

Lasketaan Pearsonin ja Spearmanin korrelaatiokertoimet.

```
CORRELATIONS
  /VARIABLES=alaika asmenot rkyks pala
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
NONPAR CORR
  /VARIABLES=alaika asmenot rkyks pala
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Tulostedokumentin tuloksista huomataan, että alueella asumisajan ja asumismenojen yhteensä välillä on merkittävä (Pearsonin korrelaatiokerroin ~ 0.29) yhteys. Huomataan myös että alueella asumisajan ja asunnon pinta-alan välillä sekä kotitalouden kuluttajayksiköiden määrän ja pinta-alan välillä on vielä merkittävämpi yhteys.

Vastaavat Spearmanin korrelaatiokertoimet ovat hieman korkeampia kuin Pearsonin, mutta eivät paljon. Suhde ei siis ole täydellisen lineaarinen,mutta ei epälineaarinenkaan.

Luodaan useamman selittäjän lineaarinen malli ja lasketaan sille kertoimet.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT pala
  /METHOD=ENTER alaika asmenot rkyks
  /SAVE ZRESID.
```

Mallin kaavaksi saadaan *23.58+24.83*rkyks+0.364*alaika*. Selitysaste on vaatimattomat 27,3%. Asumismenojen kerroin ja p-arvo on < 0.001 joten se jätetään mallista pois.

1.3. Toistomittausmalli

> *Huom. Tämän osion tulosteet löytyvät tulostedokumentin lopusta. Tulosteet eivät jostain syystä tallentuneet silloin, kun tein tehtävän.*

- Aineisto: Toistomittausaineisto2018

- Poimitaan 700 otoksen suuruinen satunnaisotos datasta Toistomittausaineisto.sav samalla tavalla kuin tehtävässä 1.1.

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.

1.3.1. Suunnitelma

Tehtävässä tulee luoda varianssianalyysimalli, jossa on mukana kahden mittauksen toistotekijä ja luokitteleva tekijä.

Ongelmat:

- *Onko potilaan mielestä saatu ohjaus leikkauksen jälkeen vähäisempää kuin odotettu ennen leikkausta?*
- *Poikkeavatko erot mittausten välillä sukupuolittain?*

Ratkaisu: *Jos varianssianalyysin oletukset ovat voimassa, ongelma voidaan ratkaista toistettujen mittausten varianssianalyysillä.*

Käytetään siis toistettujen mittausten varianssianalyysiä. Voidaan seurata varmasti hyvin Tilastollisten mallien peruskurssin kalvoista löytyvää esimerkkiä.

1.3.2. Normaalijakaumatestaus

Tutkitaan onko normaalijakaumaoletus voimassa.

```
EXAMINE VARIABLES=Functional_M2 Functional_M1 BY D2
  /PLOT BOXPLOT HISTOGRAM NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Shapiro-Wilk testin perusteella ($p < 0.001$) voidaan todeta että normaalijakaumaoletus ei ole voimassa. Tulostedokumentin histogrammeista myös nähdään että otoksen Functional mittausten arvot ovat vasemmalle vinot, ja arvot painottuvat voimakkaasti asteikon yläpäähän. (Otoksen näytteistä suurimman osan arvot ovat 4.0, joka on myös suurin arvo.)

Näin suurilla otoksilla normaalijakaumaoletuksen voisi kenties hylätä, mutta noin voimakas vinous voinee altistaa virheellisille ennusteille, sillä ANOVA olettaa datan olevan kuitenkin normaalijakautunutta. Yritetään siis ensin Friedmannin ja Wilcoxonin testejä.

1.3.3. Analyysi alkuperäisille muuttujille

Koska normaalijakaumaoletus ei ole voimassa, käytetään Friedmanin epäparametrista testiä muuttujien sijaintien erojen testaamiseen.

```
NPAR TESTS
  /FRIEDMAN=Functional_M1 Functional_M2
  /MISSING LISTWISE.
```

Tulokseksi saadaan, että muuttujien väliset arvot ovat tilastollisesti merkitseviä. (p<0.001) Muuttujien väliset vertailut pareittain voidaan siten suorittaa bonferroni-korjatuin Wilcoxonin testein.

```
NPAR TESTS
  /WILCOXON=Functional_M1 WITH Functional_M2 (PAIRED)
  /MISSING ANALYSIS.
```

Bonferroni-korjausta ei tarvita koska kyseessä on vain yksi vertailu. Mittausten välinen ero on tilastollisesti merkitsevä (p<0.001). Tulos on sama myös molemmille sukupuolille erikseen.

Suunnitelman muutos

Huomasin että eroa mittausten välillä sukupuolittain ei pysty tällä tavalla tekemään, ainakaan omilla taidoillani. Käytetään normaalijakaumaoletuksen toteutumattomuudesta huolimatta toistettujen mittausten varianssianalyysiä, jotta saamme selville mittausten välisen eron sukupuolittain.

```
GLM Functional_M1 Functional_M2 BY D2
  /WSFACTOR=mittaus 2 Polynomial
  /METHOD=SSTYPE(3)
  /PLOT=PROFILE(mittaus*D2)
  /EMMEANS=TABLES(mittaus)
  /PRINT=DESCRIPTIVE
  /CRITERIA=ALPHA(.05)
  /WSDESIGN=mittaus
  /DESIGN=D2.
```

Tuloksista selviää, että

- Sukupuolten välinen ero ei ole tilastollisesti merkitsevä (p=0.482).
- Mittausten väliset erot ovat tilastollisesti merkitseviä (p<0.001).
- Yhdysvaikutus on tilastollisesti merkitsevä (p=0.008).

Tulosteesta "Profile plots" nähdään, että jälkimmäisen mittauksen tulokset ovat olleet keskimäärin matalampia kuin ensimmäisen. Muutos on ollut laskeva molemmilla sukupuolilla, mutta naisilla pudotus tyytyväisyydessä on ollut voimakkaampi.

1.3.4. Lopputulos

Tulosten merkitsevyys ilmoitettu p-arvolla 0.05.

- *Onko potilaan mielestä saatu ohjaus leikkauksen jälkeen vähäisempää kuin odotettu ennen leikkausta?*

Kyllä.

- *Poikkeavatko erot mittausten välillä sukupuolittain?*

Kyllä - naisten tyytyväisyys putosi enemmän kuin miesten, ja sukupuolen ja mittausten yhdysvaikutus oli merkitsevää.

2. Kategoristen vastemuuttujien mallitus

- Käytetty aineisto: EK2011 (eduskuntavaaliaineisto vuodelta 2011, N=1318)
- Datasta poimittu 800 kokoinen satunnaisotos käyttämällä omaa opiskelijanumeroa satunnaislukugeneraattorin siemenlukuna, tehtävän ohjeiden mukaisesti.

2.1. Muuttujien riippuvuusrakenne

2.1.1. Tarkastellaan muuttujia sukupuoli (d2), työttömyys viimeisen 12 kuukauden aikana (k32) ja oman sukupuolen 2011 eduskuntavaaleissa äänestäminen (k23). Tee ensin yksiulotteiset frekvenssijakaumat ja kolmen muuttujan ristiintaulu

Onko 3-ulotteisessa ristiintaulussa nollasoluja?

```
FREQUENCIES VARIABLES=d2 k23 d32
  /ORDER=ANALYSIS.

CROSSTABS
  /TABLES=d32 BY k23 BY d2
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

Huomataan, että puuttuvia arvoja on sekä sukupuolessa että viimeisen 12kk työttömyydessä 13 kappaletta, ja omaa sukupuolta äänestäneissä 125 kappaletta. Omaa sukupuolta äänestäneiden puuttuvien arvojen määrä on huomattava, joka on huomioitava analyysissä.

Ristiintaulussa ei ole nollasoluja.

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

Kaikissa luokissa on havaintoja yli 10 kappaletta, joten otetaan mukaan kaikki luokat.

Suoritetaan analyysi seuraavalla komennolla

```
HILOGLINEAR d2(1 2) d32(1 2) k23(0 1)
  /METHOD=BACKWARD
  /CRITERIA MAXSTEPS(10) P(.05) ITERATION(20) DELTA(.5)
  /PRINT=FREQ RESID
  /DESIGN.
```

Millaiset riippuvuudet muuttujien välillä askeltavan menetelmän avulla valittuun malliin jäivät?

Riippuvuuksiksi jäivät sukupuolen (d2) ja omaa sukupuolta äänestämisen (k23) yhteisvaikutus.

Mikä on mallin generoiva luokka?

Askeltavan mallin valitseman mallin generoiva luokka on d2*k23 + d32. (sukupuoli*omaa sukupuolta äänestäminen + työttömyys viim. 12kk aikana).

Mikä on mallin yhteensopivuustestin p-arvo? Mikä on standardoitujen jäännösten vaihteluväli?

Yhteensopivuustestin p-arvo on 0.561.

Mallin standardoidut jäännökset ovat pieniä [-0.657, 1.046], joten malli sopii kuvaamaan kolmen muuttujan keskinäisiä yhteyksiä.

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

Mallin ainoa riippuvuus on sukupuolen ja omaa sukupuolta äänestämisen välillä, joten sen jatkotarkastelu riittää. Jatkotarkasteluksi riittää sukupuolen ja omaa sukupuolta äänestämisen välisen yhteyden tarkastelu marginaalitaulusta.

```
CROSSTABS
  /TABLES=d2 BY k23
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT ROW SRESID
  /COUNT ROUND CELL.
```

Tuloksista huomataan, että miehet äänestävät naisia useammin omaa sukupuolta olevaa ehdokasta. Tässä aineistossa miehistä 66,6% äänestää omaa sukupuolta, naisista vain 55,8%.

2.2. Kaksiluokkainen selitettävä muuttuja

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

Tutkitaan dataa SPSS:n Binary logistics -työkalulla. Otetaan mukaan myös luottamusvälit.

```
LOGISTIC REGRESSION VARIABLES d32
  /METHOD=ENTER d2 ika
  /CONTRAST (d2)=Simple
  /PRINT=GOODFIT CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

2.2.2. Mitkä muuttujat selittävät työttömyyttä ?

Tuloksista huomataan, että sukupuolella ei ole tilastollisesti merkitsevää yhteyttä työttömyyteen viimeisen 12kk aikana ($p = 0.439$).

Sen sijaan mitä enemmän on ikää, sitä todennäköisempää on, että ei ole ollut työttömänä viimeisen 12 kk aikana. Iän yhteydelle työttömyyteen $p < 0.001$.

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

Mitä enemmän on ikää, sitä todennäköisempää on, että ei ole ollut työttömänä viimeisen 12kk aikana. (OR = 0.963).

Koska sukupuoli ei ole merkittävä tekijä mallissa, sitä ei huomioida tarkastelussa.

Iän kasvaessa todennäköisyys luokittua työttömäksi vähenee 0.038 yksikköä per ikävuosi.

OR:n luottamusväli on [0.951, 0.975].

Mikä on mallin Nagelkerke-selitysaste?

Mallin Nagelkerke-selitysaste on 0.086.

3. Monimuuttujamenetelmät

- Aineisto: pankkiotos2018 (todellinen asiakasaineisto, N=2453)

- Datasta poimittu 1200 kokoinen satunnaisotos käyttämällä omaa opiskelijanumeroa satunnaislukugeneraattorin siemenlukuna, tehtävän ohjeiden mukaisesti.

3.1.Muuttujien ryhmittely

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

Muodostetaan pääkomponentit SPSS:n Dimension Reduction - Factor -työkalulla. Ensimmäisellä kerralla mukaan valittiin kaikki muuttujat, mutta komennon suoritus päättyi virheeseen. Silmämääräisellä tarkistuksella huomattiin, että muuttujalla *toimeksianto_b_kpl_luok* kaikki arvot olivat 0. Tämä muuttuja ei siten tuo mitään tietoa analyysiin, ja sen voi jättää analyysista pois. Poisjättämisen jälkeen syntaksin suoritus onnistui.

```
FACTOR
  /VARIABLES autom_lainan_perinta_luok lainojen_lukumaara_luok asuntoluotot1_luok
    automaattinostoja_luok vakuutus_a_luok asuntolaina_a_kpl_luok
asuntolaina_b_kpl_luok
    vakuutus_b_luok vakuutus_c_luok korkeakork_kpl_luok rahasto_a1_luok
pankkikorttilkm_luok
    luottokortteja_yhteensa_luok maaraaikaistileja_luok
maksuautomaattitapahtumia_luok
    kayttotili_tal_luok kayttotili_vel_luok asuntolaina_c_kpl_luok
osakkeet_euroa_1_luok
    eri_osakesarjoja_luok rahasto_b1_luok ottoja_luok pkorttimaksuja_luok
panoja_luok
    asuntolaina_d_kpl_luok palveluja_kpl_luok rahastolajeja_luok lainarastit_luok
saastotililla_luok
    asuntolaina_e_kpl_luok suoraveloituksia_luok netissa_maksut_luok
maksupalvelussa_maksut_luok
    tiskilla_maksut_luok tilinylityspaivat_luok toimeksianto_a_kpl_luok
kv_maksukortit_luok
    rahasto_c1_luok korttiluotot1_luok kulutusluotot1_luok
  /MISSING LISTWISE
  /ANALYSIS autom_lainan_perinta_luok lainojen_lukumaara_luok asuntoluotot1_luok
    automaattinostoja_luok vakuutus_a_luok asuntolaina_a_kpl_luok
asuntolaina_b_kpl_luok
    vakuutus_b_luok vakuutus_c_luok korkeakork_kpl_luok rahasto_a1_luok
pankkikorttilkm_luok
    luottokortteja_yhteensa_luok maaraaikaistileja_luok
maksuautomaattitapahtumia_luok
    kayttotili_tal_luok kayttotili_vel_luok asuntolaina_c_kpl_luok
osakkeet_euroa_1_luok
    eri_osakesarjoja_luok rahasto_b1_luok ottoja_luok pkorttimaksuja_luok
panoja_luok
    asuntolaina_d_kpl_luok palveluja_kpl_luok rahastolajeja_luok lainarastit_luok
saastotililla_luok
    asuntolaina_e_kpl_luok suoraveloituksia_luok netissa_maksut_luok
maksupalvelussa_maksut_luok
    tiskilla_maksut_luok tilinylityspaivat_luok toimeksianto_a_kpl_luok
kv_maksukortit_luok
    rahasto_c1_luok korttiluotot1_luok kulutusluotot1_luok
  /PRINT INITIAL SIG KMO AIC EXTRACTION ROTATION
  /FORMAT SORT
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION PROMAX(4)
  /SAVE REG(ALL)
  /METHOD=CORRELATION.
```

Pääkomponentteja muodostui 14 kappaletta. Ominaisarvokriteerinä pidettiin ominaisarvon olemista suurempi kuin 1.

3.1.2. Talleta havaintomatriisiin uusiksi muuttujiksi pääkomponenttipistemäärät

Pääkomponentit tallettuivat havaintomatriisiin syntaksin

```
/SAVE REG(ALL)
```

komennon ansiosta.

3.1.3. Nimeä uudet muuttujat (pääkomponentteihin latautuneiden muuttujien mukaisesti)

Noin ensimmäisen neljän pääkomponentin osalta on jokseenkin tulkittavissa, että komponentteihin latautuu toisiinsa liittyviä muuttujia. Tästä eteenpäin komponenttien lataukset ovat niin hajanaiset, että muuttujien nimeäminen olisi jo harhaanjohtavaa.

Nämä ovat toki tulkintakysymyksiä mutta kun kyseessä on todellinen aineisto helposti ymmärrettävästä aiheesta, tulkinnan tekeminen selkeästi tulkittavissa olevista pääkomponenttien latauksista ei johda tarpeettomasti harhaan.

Ehdotukseni joidenkin pääkomponenttien nimeksi:

1. **Tilinkäyttöaktiivisuus** (Pääkomponentissa painottuvat erityisesti ottojen, nostojen, panojen ja maksujen määrä.)
2. **Säästämisaktiivisuus** (Painottuvina rahastot, osakkeet yms.)
3. *Vaikeatulkintainen.* (Lataukset matalia) Jätetään nimeämättä.
4. **Asuntolainallisuus** (painottuvina asuntolainaan ja lainanlyhennykseen liittyvät muuttujat)
5. **Osakesijoittaminen**
6. *Vaikeatulkintainen*
7. *Vaikeatulkintainen*
8. *Vaikeatulkintainen*
9. *Vaikeatulkintainen*
10. *Vaikeatulkintainen*
11. **Käyttötilivelallinen**
12. **Rahasto a**
13. **Vakuutukselliset**
14. **Asuntolaina b**

Klusterianalyysiä varten jätetään komponentit kuitenkin toistaiseksi nimeämättä, ja käytetään pääkomponenttien numerointia 1-14.

3.2. Havaintojen ryhmittely

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

Suoritetaan klusterianalyysi aina viiteen klusteriin asti:

```
QUICK CLUSTER FAC1_1 FAC2_1 FAC3_1 FAC4_1 FAC5_1 FAC6_1 FAC7_1 FAC8_1 FAC9_1
FAC10_1 FAC11_1
    FAC12_1 FAC13_1 FAC14_1
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(2) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER
  /PRINT INITIAL.

QUICK CLUSTER FAC1_1 FAC2_1 FAC3_1 FAC4_1 FAC5_1 FAC6_1 FAC7_1 FAC8_1 FAC9_1
FAC10_1 FAC11_1
    FAC12_1 FAC13_1 FAC14_1
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(3) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER
  /PRINT INITIAL.

QUICK CLUSTER FAC1_1 FAC2_1 FAC3_1 FAC4_1 FAC5_1 FAC6_1 FAC7_1 FAC8_1 FAC9_1
FAC10_1 FAC11_1
    FAC12_1 FAC13_1 FAC14_1
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(4) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER
  /PRINT INITIAL.

QUICK CLUSTER FAC1_1 FAC2_1 FAC3_1 FAC4_1 FAC5_1 FAC6_1 FAC7_1 FAC8_1 FAC9_1
FAC10_1 FAC11_1
    FAC12_1 FAC13_1 FAC14_1
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(5) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER
  /PRINT INITIAL.
```

Huomataan että viiden klusterin tapauksessa jää klustereihin 2, 4, ja 5 vain 2, 4, ja 4 kappaletta havaintoja. Tällöin on tuskin mielekästä ottaa malliin näin montaa klusteria.

Muiden klusterimäärien havaintojen määrät klustereittain:

k=2:

1. 1134
2. 66

k=3:

1. 70
2. 250
3. 880

k=4:

1. 44
2. 25
3. 251
4. 880

Valitaan analyysin klusterien määräksi 4, ja kuvaillaan muodostuneita ryhmiä.

- Klusteria 1 profiloivat pääkomponentit:
  - 3, 6, 11, 12
- Klusteria 2 profiloivat pääkomponentit:
  - 4, 5, 10, 13
- Klusteria 3 profiloivat pääkomponentit:
  - 1, 2, 7, 8, 9
- Klusteria 4 profiloivat pääkomponentit:
  - 14

Johtopäätökset

Pääkomponentit hajaantuvat eri klustereihin niin, että johtopäätöksiä ryhmien luonteesta on vaikea tehdä, etenkin johtuen pääkomponenttien suuresta määrästä ja pääkomponentteihin latautuneiden muuttujien lataantumisien hajaantumisesta.

Riskinä on virheellisten johtopäätösten teko ryhmien luonteesta. Jätetään ryhmien tulkinta siis tekemättä ja tyydytään kuvailussa analyysin tuottamien tulosten varaan.