

Otanta-aineistojen analyysi, kevät 2014

Syventävät opinnot (pakollinen harjoitustyö)

HARJOITUSTYÖ

Harjoitustyö tehdään henkilökohtaisena työnä.

Harjoitustyössä on viisi tehtävää: kaikille yhteiset tehtävät (1-3) ja valinnainen tehtävä (4 tai 5).

Loppuotteena on 15–20 sivun raportti liitteinen, esimerkiksi rakenteella:

Kansilehti (otsikko, tekijä, kurssi, päiväys)

Tekstiosa (jaa soveltuviin lukuihin ja alilukuihin)

Viitteet (kirjallisuus)

Liitteet (valitut ohjelmakoodit ja tulostusotteet).

Työssä voi käyttää SAS-versiota 9.3 tai soveltuvia SPSS- tai R-työkaluja.

Työn palautus: **31.3.2014 mennessä** liitetiedostona (PDF) sähköpostitse: risto.lehtonen@helsinki.fi tai postitse:

Risto Lehtonen

Helsingin yliopisto, Matematiikan ja tilastotieteen laitos

PL 68 (Gustaf Hällströmin katu 2b), 00014 Helsingin yliopisto

Harjoitustyössä analysoidaan työterveyshuoltotutkimuksen aineistoa (OHC-data). Olkoon kiinnostuksen kohteena psyykinen rasittuneisuus (OHC-aineiston tulosmuuttujat PSYCH ja PSYCH2). Halutaan selvittää, millainen yhteys psyykkisen rasittuneisuuden kokemisella on työoloihin ja sairastavuuteen. Tätä varten tulosmuuttujia mallinnetaan soveltuvan tilastollisen mallin sekä sopivien selittävien muuttujien avulla (esimerkiksi työn fysikaaliset haitat ja pitkäaikaissairastavuus sekä sukupuoli ja ikä). Erityisesti kiinnostuksen kohteena on sisällöllisten selittävien muuttujien ikä-sukupuolivakioitu yhteys psyykkiseen rasittuneisuuteen. Analyysissa tulee ottaa huomioon otanta-asetelman erityispiirteet asetelmaperusteisella metodiikalla (tehtävä 1) ja malliperusteisella metodiikalla (tehtävä 2).

Tehtävä 1. Asetelmaperusteinen analyysi

- Kuvaa analyysiasetelman ja siihen liittyvän aineiston ja otanta-asetelman erityispiirteet. Mitä asioita tässä yhteydessä tulee ottaa huomioon, kun analyyttistä lähestymistapaa ja analyysityökaluja valitaan?
- Valitse **asetelmaperusteista analyysia** varten tilastollinen (kiinteiden tekijöiden) malli ja soveltuva SAS- ohjelmiston, SPSS:n tai R:n analyysiproseduuri molemmille tulosmuuttujille (jatkuva PSYCH ja binäärinen PSYCH2). Perustele, miksi päädyit valitsemaasi malliin ja analyysiproseduriin kummassakin tapauksessa.
- Suorita asetelmaperusteinen analyysi valitsemallasi analyysiproseduurilla. Etsi tulkittavaksi kelvollinen ”lopullinen” malli. Raportoi mallin valintaproseduurin tulokset taulukkomuodossa ja verbaalisesti. Raportoi lopullisen mallin parametriestimaatit, keskivirheet, asetelmakertoimet, testaustulokset ja muut valitsemasi relevantit tunnusluvut. Pohdi diagnostisesti ”lopullisen” mallin hyvyttä.

Tehtävä 2. Malliperusteinen analyysi

- a) Valitse **malliperusteista analyysia** varten tilastollinen sekamalli ja soveltuva SAS- ohjelmiston, SPSS:n tai R:n analyysiproseduuri molemmille tulosmuuttujille (jatkuva PSYCH ja binäärinen PSYCH2). Perustele, miksi päädyit valitsemaasi malliin ja analyysiproseduriin kummassakin tapauksessa.
- b) Suorita malliperusteinen analyysi valitsemallasi analyysiproseduurilla. Etsi tulkittavaksi kelvollinen ”lopullinen” malli. Raportoi mallin valintaproseduurin tulokset taulukkomuodossa ja verbaalisesti. Raportoi lopullisen mallin parametriestimaatit, keskivirheet, testaustulokset ja muut valitsemasi relevantit tunnusluvut. Pohdi diagnostisesti ”lopullisen” mallin hyvyyttä.
- c) Vertaa malliperusteisen analyysin tuloksia asetelmaperusteisiin tuloksiin (tehtävä 1) ja raportoi vertailutulokset sekä johtopäätökset.

Tehtävä 3. Logit-ANOVA-mallin asetelmaperusteinen GWLS-estimointi

- a) Muokkaa luennoilla ja harjoituksissa käsiteltyä SAS/IML-kielistä GWLS-ohjelmaa (mallimatriisit) niin, että voit käyttää ohjelmaa logit-ANOVA-mallille, jossa on binäärinen tulosmuuttuja ja kaksi kaksiluokkaista diskreettiä selittävää muuttujaa.
- b) Sovella ohjelmaa tilanteessa, jossa tulosmuuttujana on PSYCH2 ja diskreetteinä selittäjinä SEX ja jokin muuttujista AGE2, PHYS tai CHRON. Sovita päävaikutusmalli. Raportoi tulokset. Vertaa tuloksia vastaavan mallin sovittamiseen SAS-ohjelmiston SURVEYLOGISTIC-proseduurilla. Raportoi vertailutulokset.

Valinnainen tehtävä Valitse toinen seuraavista:

Tehtävä 4. Jackknife-menetelmä

Varianssiestimointi Jackknife-menetelmällä modifioidussa OHC-asetelmassa (ositemuuttujana NEWSTR ja ryväsmuuttujana NEWCLU).

- a) Laajenna Jackknife-varianssiapproksimoinnin SAS-makroa niin, että makro kattaa perusestimaattorin $v(1.jrr)$ lisäksi luennoilla esitetyistä vaihtoehtoisista JRR-varianssiapproksimaatioista valitsemasi kaksi muuta versiota (ks. Lehtonen ja Pahkinen 2004, s. 156–157).
- b) Sovella kohdassa a) laadittuja kolmea JRR-varianssiapproksimaatioita tulosmuuttujan PHYS osuuden varianssin approksimointiin naisten osajoukossa. Suorita vertailu JRR-estimaattoreiden kesken. Vertaa tuloksia linearisointimenetelmällä (SAS/SURVEYMEANS) saatavaan varianssiestimaattiin. Raportoi vertailutulokset.

Tehtävä 5. Bootstrap-menetelmä

Varianssiestimointi BOOT-menetelmällä (Bootstrap) modifioidussa OHC-asetelmassa (ositemuuttujana NEWSTR ja ryväsmuuttujana NEWCLU).

- a) Täydennä BOOT-makroa niin, että käytät Bootstrap-varianssiestimaattorissa koko otoksesta lasketun parent-estimaattorin sijasta bootstrap-otoksista laskettujen estimaattien keskiarvoa ($K=1000$ bootstrap-otosta) (ks. Lehtonen ja Pahkinen 2004, s. 160–161).
- b) Sovella BOOT-makroa tulosmuuttujan PHYS osuuden varianssin approksimointiin naisten osajoukossa. Vertaa tuloksia jackknife-menetelmien antamiin tuloksiin ja linearisointimenetelmällä (SAS/SURVEYMEANS) saatavaan varianssiestimaattiin. Raportoi vertailutulokset.