

## Otanta-aineistojen analyysi, kevät 2010

### Syventävät opinnot (pakollinen harjoitustyö)

## HARJOITUSTYÖ

Harjoitustyö tehdään henkilökohtaisena työnä.

Harjoitustyössä on kaksi tehtävää, kaikille yhteinen (tehtävä 1) ja valinnainen (tehtävä 2, 3 tai 4).

Lopputuotteena on noin 10 sivun raportti liitteineen, 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 9.1.3 tai 9.2.

Työn palautus: **30.3.2010 mennessä** liitetiedostona (PDF) sähköpostitse: [risto.lehtonen@helsinki.fi](mailto:risto.lehtonen@helsinki.fi) tai postitse:

Risto Lehtonen

Helsingin yliopisto, Matematiikan ja tilastotieteen laitos

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

### Tehtävä 1.

Data-analyysia. Olkoon kiinnostuksen kohteena psyykinen rasittuneisuus (OHC-aineiston tulosmuuttuja PSYCH2). Halutaan selvittää, millainen yhteys ilmiöllä on työoloihin ja sairastavuuteen. Tätä varten tulosmuuttujaa mallinnetaan soveltuvan tilastollisen mallin sekä selittävien muuttujien sukupuoli (SEX), ikä (AGE), työn fyysiset haitat (PHYS) ja pitkäaikaissairastavuus (CHRON) avulla. Erityisesti kiinnostuksen kohteena olevat selittäjät ovat PHYS ja CHRON, joiden ikä-sukupuolivakioitu yhteys psyykkiseen rasittuneisuuteen halutaan selvittää.

Suorita seuraavat työvaiheet:

- Kuvaa lyhyesti analyysiasetelman ja siihen liittyvän aineiston ja otanta-asetelman erityispiirteet. Mitä asioita tässä yhteydessä tulee ottaa huomioon, kun analyyttistä lähestymistapaa ja analyysityökaluja valitaan?
- Valitse analyysitilanteeseen soveltuva tilastollinen malli ja soveltuva SAS- ohjelmiston analyysiproseduuri (tulosmuuttujana binäärinen PSYCH2). Perustele, miksi päädyit valitsemaasi malliin ja analyysiproseduuriin.
- Sovita tulosmuuttujalle malli asetelmaperusteisesti valitsemallasi analyysiproseduurilla. Etsi tulkittavaksi kelvoinen ”lopullinen” malli. Raportoi mallin valintaproseduurin tulokset taulukkomuodossa ja verbaalisesti. Raportoi lopullisen mallin parametristimaatit, keskivirheet, asetelmakertoimet, testaustulokset ja muut valitsemasi relevantit tunnusluvut. Pohdi diagnostisesti ”lopullisen” mallin hyvyttä.
- Suorita kohdan c) operaatiot malliperusteisesti (SRS-oletus) soveltuvalla SAS-ohjelmiston analyysiproseduurilla. Vertaa tuloksia asetelmaperusteisiin tuloksiin ja raportoi vertailutulokset sekä johtopäätökset.

## **Valinnainen tehtävä Valitse yksi seuraavista:**

### **Tehtävä 2.**

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 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.

### **Tehtävä 3.**

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

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

### **Tehtävä 4.**

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

- a) Täydennä BOOT-makroa niin, että käytät Bootstrap-variانسsiestimaattorissa 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 BOOT-variانسsiestimaattoreiden ( $K=1000$  bootstrap-otosta) antamiin tuloksiin ja linearisointimenetelmällä (SAS/SURVEYMEANS) saatavaan variانسsiestimaattiin. Raportoi vertailutulokset.