

Tilastollisten ohjelmistojen käyttö Terveys 2000 – tutkimuksen yhteydessä

1 Aineiston rakenne ja erikoisvaatimukset

Aineiston keräämiskustannusten pienentämiseksi havaintoyksilöt poimittiin 80 *rypäästä* (terveyskeskuspiiristä), jotka puolestaan poimittiin 20 *ositteesta* (15 suurimmasta kaupungista ja 5 yliopistosairaala- eli ns. *miljoonapiiristä*). Tällöin havaintoyksilöiden tutkiminen voitiin keskittää muutamiin terveyskeskuksiin ja kotikäynnit tutkittavien luokse oli edullisempaa järjestää lyhyempien välimatkojen ansiosta kuin yksinkertaisen satunnaisotannan tapauksessa, jolloin tutkittavat olisivat jakautuneet tasaisesti lähes kaikkiin terveyskeskuspiireihin.

Tällainen *kaksivaiheinen otanta* aiheuttaa kuitenkin omat haasteensa tilastollisissa analyysissä, koska useimmat tilasto-ohjelmistot eivät osaa ottaa huomioon osituksesta ja ryvästyksestä johtuvaa riippuvuutta samasta rypäästä poimittujen havaintoyksilöiden välillä. Jos riippuvuutta ei oteta huomioon, niin tuloksena saatavat *p*-arvot ovat yleensä liian pieniä ja luottamusvälit liian kapeita, eli tulokset ovat liian liberaaleja.

Toinen tärkeä ominaisuus Terveys 2000 –tutkimuksen otannassa on *painotus*, joilla korjataan havaintoaineiston ikä-, sukupuoli-, alue- ja kielijakaumia vastaamaan populaation todellisia jakaumia, jolloin vähintään 80-vuotiaiden kaksinkertaisen poimintatodennäköisyyden vaikutus ja myös kadon vaikutuksia voidaan kompensoida erityisesti perustunnuslukuja, kuten keskiarvoja ja prevalensseja, laskettaessa. Painokertoimia on kahdentyyppisiä: *analyysi-* ja *korottavia* (tai väestö-) *painokertoimia*. Analyysipainojen summa on havaintoyksilöiden lukumäärä ja korottavien painojen perusjoukon koko. Analyysipainoja käytetään lähinnä regressioanalyysissä ja korottavia painoja erilaisia kuvailevia tunnuslukuja laskettaessa. Useat tilasto-ohjelmistot osaavat huomioida painokertoimet analyysissä, mutta painokertoimet eivät korjaa havaintojen välisistä riippuvuuksista johtuvia virheitä.

Jos analyysijä rajoitetaan pieniin osajoukkoihin, niin tulosten yleistettävyyden koko väestöön voi olla heikko. Analyysijä vaikeuttavat mm. äidinkieleen, alueeseen tai johonkin harvinaiseen ilmiöön liittyvät osajoukot. Suomen- ja ruotsinkielisten vastausaktiivisuus oli selvästi parempi kuin muunkielisten, joita on muutenkin hyvin pieni määrä havaintoaineistossa, joten ainoastaan suomen- ja ruotsinkielisiä voidaan verrata ryhminä. Tulosten yleistettävyyden liittyy myös alueelliseen kattavuuteen, koska miljoonapiirejä hienommalla alueellisella jaolla otokseen valittujen/poimittujen terveyskeskuspiirien ja havaintoyksilöiden lukumäärä voi jäädä pieneksi. Esimerkiksi Kainuun maakunnasta mukana on ainoastaan Kajaanin kaupunki, josta poimittiin 60 havaintoyksilöä, jolloin Terveys 2000 –aineiston edustavuus Kainuun maakunnan suhteen on heikko sekä alueellisesti että määrällisesti. Jos havaintojen lukumäärä osajoukossa jää vähäiseksi (eli muutamaan kymmeneen tai alle), tulosten yleistettävyyden koko väestössä tulisi suhtautua varauksella. Osajoukkojen analysoimisesta on huomionarvoisia mainintoja myös kappaleissa 2 ja 5.3.

Eri tutkimusosioissa oleva kato vaihtelee merkittävästi, ja tämä on vaikeuttanut painokertoimien muodostamista. Tämän takia Tilastokeskus on muodostanut kolmet painokertoimet: (i) *ravintokysely*, (ii) *unionijoukko*, johon on laskettu ainakin johonkin mittauspisteeseen osallistuneet havaintoyksilöt, ja (iii) *leikkausjoukko*, jossa olevat havaintoyksilöt ovat vastanneet kaikissa mittauspisteissä muutamaa keskeiseen kysymykseen tai mittaukseen. Unionijoukkopainoja käytetään oletusarvoisesti, koska ne mahdollistavat lähes kaikkien havaintojen käytön.

2 Ohjelmistot

Tällä hetkellä ehkä kattavimmat ominaisuudet otanta-asetelman huomioimiseksi analyyseissä tarjoaa SUDAAN, josta on olemassa sekä SAS-versio, että *stand alone* -versio, joka toimii ilman SAS-ohjelmistoa.

Toinen sopiva ohjelmisto on Stata, jonka hankkiminen on edullisempaa kuin SAS- ja SUDAAN-ohjelmistojen. Ominaisuudet ovat kuitenkin mallivakioinnin osalta rajoittuneemmat.

Mikäli tutkija osaa entuudestaan käyttää SAS-ohjelmistoa tai tutkijan työympäristössä on käytössä SAS, on SUDAAN-vaihtoehto yleensä suositeltavampi.

Osajoukkojen analysoinnissa suositellaan käytettäväksi kaikkia havaintoja, ja osajoukon rajaamista SUBPOP- (SUDAAN) ja subpop()- (Stata) optioita käyttäen. Jos havaintotiedosto rajataan ennen analysointia tai käytetään *if*- tai *in*-optioita, niin joissakin tapauksissa varianssiestimaatit saattavat olla virheellisiä.

Muita ohjelmistoja, joita ovat esim. R ja SAS:n SURVEY*-proseduurit, ei tueta Terveys 2000 -organisaation puolelta. SPSS-ohjelmiston versioon 12 on saatavana *SPSS Complex Samples 12.0* -paketti, jolla voi tuottaa otanta-asetelman huomioivat perustunnusluvut sekä 1- ja 2-ulotteisia taulukoita, mutta esim. regressioanalyysien tekeminen ei edelleenkään onnistu. Rakenneyhtälömalleja on mahdollista analysoida Lisrel-ohjelmistolla.

3 Ohjelmistopalvelut

3.1 Ohjelmistojen ylläpito

Kansanterveyslaitoksen Terveiden ja toimintakyvyn osasto (TTO) antaa tukea em. ohjelmistojen käyttöä varten: Tarvittavia perusmakroja tuotetaan SAS/SUDAAN- ja Stata-ympäristöille. Perustaulukoita varten on tuotettu SAS/SUDAAN-makro, ja Stataa varten tuotetaan makro mallivakiointia varten. Tarvittaessa laaditaan myös SAS/SUDAAN-makro poikkileikkaustutkimuksia varten.

3.2 Yhteyshenkilöt

Jokaisessa tutkimusryhmässä pitää olla ainakin yksi menetelmällisesti orientoitunut tukihenkilö, johon ryhmän tutkijat ensisijaisesti tukeutuvat tilastollisen tutkimuksen kysymyksissä. Tukihenkilöille järjestetään koulutustilaisuus, jonka vetäjinä ovat Tommi Härkänen ja Esa Virtala. Tarvittaessa tukihenkilö voi välittää kysymyksen TTO:lle, jolloin kontaktihenkilöinä ovat:

- Tommi Härkänen (otanta-asetelman huomioiminen tilastollisessa analyysissä)
- Esa Virtala (ohjelmistot)
- Harri Rissanen (tulosteiden massatuotanto)
- Paul Knekt (epidemiologisen tutkimuksen tilastolliset menetelmät).

Em. henkilöiden sähköpostiosoitteet ovat muotoa etunimi.sukunimi@ktl.fi.
 Satunnaista käyttöä varten on mahdollista lainata muutamia SAS/SUDAAN- tai Stata-ohjelmistoilla varustettuja kannettavia tietokoneita esim. tutkijoille, joilla ei ole mahdollisuutta saada omalta laitoksestaan käyttöönsä asianmukaisia ohjelmistoja.

4 Tutkimusasetelmat

Poikkileikkaustutkimuksessa käytetään ohjelmistoa, joka kykenee huomioimaan sekä painokertoimet että otanta-asetelman ositus- ja ryvästysrakenteen. Tällöin tulokset voidaan yleistää perusjoukkoon. Sopivia ohjelmistoja ovat siis esim. SUDAAN ja Stata.

Kahden otoksen, eli Terveys 2000 ja Mini-Suomi -otosten vertailu tapahtuu vastaavalla tavalla kuin poikkileikkaustutkimuksessakin. Kaikkien perustunnuslukujen ja mallivakiointien yhteydessä tulee kuitenkin käyttää *korottavia painoja* (eli väestöpainoja), mikä korjaa erilaisista sisällyttämistodennäköisyyksistä johtuvaa harhaa estimaateissa. Molemmille otoksille on muodostettu omat painokertoimet. Regressiomalleissa tulee käyttää otoksen identifioivaa muuttujaa selittäjänä.

Mini-Suomi –terveystutkimusaineiston toistomittaukset Terveys 2000 – tutkimuksessa perustuvat vuonna 2000 yliopistosairaalaapaikkakunnilla tai niiden läheisyydessä asuneisiin Mini-Suomi –tutkimukseen osallistuneiden tutkimuksiin. Tämä aiheuttaa yhdessä esim. muuttoliikkeen vaikutuksen kanssa monimutkaisen valikoitumisen, jolloin tulosten yleistäminen koko perusjoukkoon (esim. painokertoimia käyttämällä) käy lähes mahdottomaksi. Tämän takia tulokset pitää esittää vain ko. havaintoyksilöihin liittyen. Mahdollisia analysointimenetelmiä ovat erilaiset toistomittausmallit, sekamallit ja ehdollinen logistinen regressiomalli ilman painokertoimia.

Kohortti- ja upotettua tapaus-verrokkitutkimusasetelmaa ei käsitellä tässä muistiossa.

4.1 Mallivakioinnista

Erilaisten determinanttien vaikutuksen havainnollistamiseen voidaan käyttää esim. *predictive margin* -mallivakiointia (alkuperäinen viite Lee 1981, varianssiestimointi otanta-asetelmatilanteessa Graubard ja Korn 1999). Tässä menetelmässä jokaiselle havaintoyksilölle lasketaan ennustearvo lineaarisen tai logistisen regressiomallin avulla asettamalla yhden tai useamman selittäjän (determinantin) arvo tiettyyn vakioarvoon. Lineaarisen mallin tapauksessa ennustearvo on vastemuuttujan odotettavissa oleva arvo ja logistisen mallin tapauksessa todennäköisyys sille, että vastemuuttujan arvo on 1. Keskiarvo näistä yksilöllisistä ennustearvoista kuvaa väestökeskiarvoa tai -prevalenssia siinä hypoteettisessa tilanteessa, että determinanttimuuttujan arvo olisi sama kaikilla yksilöillä. Tällä tavalla voidaan poistaa sekoittavien tekijöiden vaikutus tarkasteltaessa determinanttimuuttujan vaikutusta vastemuuttujaan.

Esimerkiksi ikäjakaumat ovat sukupuolittain erilaisia, koska naiset elävät keskimäärin pidempään kuin miehet. Jos tarkastellaan korkean systolisen verenpaineen havaittua prevalenssia sukupuolittain ilman iän sekoittavan vaikutuksen huomioimista, niin vanhojen naisten suuri osuus yliarvioi prevalenssia verrattuna miehiin. Muodostetaan logistinen regressiomalli, jossa selittäjinä ovat ikä ja sukupuoli (sekä mahdollisesti muita selittäjiä). Mallivakioinnissa tarkastellaan tässä tapauksessa kahta hypoteettista tilannetta, eli "kaikki ovat naisia" ja "kaikki ovat miehiä". Tuloksena saadaan mallivakioidut prevalenssit, jotka ovat keskenään vertailukelpoisia, koska ikä- ja muiden selittäjien jakaumat ovat samoja kummassakin hypoteettisessa tilanteessa.

5 Käyttöesimerkkejä

5.1 Esimerkkiaineisto

Tutkijoille toimitettavassa tiedostossa ovat muuttujat `osite` ja `ryvas`, joilla kuvataan otanta-asetelman rakenne. Lisäksi tilastollisessa analyysissä tarvitaan painokertoimet, jotka ovat muuttujassa `paino`.

Vasteena on systolinen verenpaine `SystBP2`, jota käsitellään jatkuvana, binäärisenä (kynnysarvon 140 ylittävälle vasteen arvo on 1 ja muille 0) ja moniluokkaisena (kynnysarvot 120 ja 160 tuottavat vasteen arvot 1, 2 ja 3). Selittäjinä ovat

- ikä (jatkuvana tai 6-luokkaisena) `ika` ja `ika6`
- sukupuoli `sp2`
- siviilisääty `aa01`
- painoindeksi (*body mass index*) `BMI`
- kokonaiskolesteroli `T114`
- HDL-kolesteroli `T115` ja
- toimintakykyä kuvaavana muuttujana portaiden nouseminen `PortaanNousu`.

On huolehdittava, että binäärisen selittävän muuttujan arvot ovat joko 1 tai 2, ja että k -luokkaisen luokitellun muuttujan arvoalue on 1, 2, ..., k .

5.2 SAS/SUDAAN

Otanta-asetelma huomioidaan kaikissa SUDAANin proseduureissa `NEST-`lauseella, jonka ensimmäinen argumentti on ositemuuttuja ja toinen ryvästysmuuttuja. Painomuuttuja esitellään `WEIGHT-`lauseella.

Kuvailevia tunnuslukuja voidaan tulostaa `DESCRIPT-`, `CROSSTAB-` ja `RATIO-`proseduureilla. Jatkuvien ja luokiteltujen muuttujien tunnuslukuja voidaan laskea `DESCRIPT-`proseduurilla, luokiteltujen muuttujien erilaisia taulukoita `CROSSTAB-`proseduurilla ja suhde-estimaatteja `RATIO-`proseduurilla. Systolisen verenpaineen ja kokonaiskolesterolin solukeskiarvot sukupuolen, 6-luokkaisen ikämuuttujan ja tutkimuskohortin mukaan saadaan

```
proc descript data=work.t2k data;
  setenv decwidth=4 colwidth=13;
  nest osite ryvas;
  weight wan unioni;
  subgroup t2k sp2   ika6;
  levels   2   2     6;
```

```
var systbp2 T114;
tables t2k*sp2*ika6;
run;
```

Lukumäärien taulukointiin voidaan käyttää CROSSTAB-proseduuria. Esim. lasketaan solukohtaisesti havaintojen lukumäärät, painotetut lukumäärät keskivirheineen ja suhteelliset osuudet sarakkeittain ja riveittäin keskivirheineen:

```
proc crosstab data=work.t2k data;
  setenv decwidth=4 colwidth=13;
  nest osite ryvas;
  weight wan unioni;
  subgroup t2k sp2 ika6 aa01 Portaannousu;
  levels 2 2 6 4 2;
  tables t2k*sp2*ika6;
  print nsum wsum sewgt colper rowper secol serow;
run;
```

Luokittelevia selittäjiä voidaan käyttää komentojen SUBGROUP ja LEVELS avulla. SUBGROUP-komennolla luetellaan muuttujat, ja LEVELS-komennolla eri muuttujien eri arvojen lukumäärät vastaavasti. Komennolla REFLEVEL muuttujan nimi=luokka; voidaan valita luokittelevan selittäjän referenssiluokka. Huomaa option defft4 käyttäminen osituksen vaikutuksen arvioimisessa.

Logistisia ja lineaarisia regressiomalleja estimoidaan RLOGIST- ja REGRESS-proseduureilla. Lukumäärävasteiden mallittamiseen on käytettävissä LOGLINK-proseduuri, ja suhteellisten riskien mallittamiseen SURVIVAL-proseduuri.

Sudaan kykenee laskemaan sekä *predictive margin* että ehdollisia mallivakiointeja luokiteltujen selittäjien eri arvoille erilaisten regressioanalyysien yhteydessä. Esim. PREDMARG ika6; laskee *predictive margin* -tunnusluvut ikäluokkamuuttujan ika6 eri arvoilla. Jatkuvan vasteen tapauksessa mallivakioidut tunnusluvut ovat ennustekeskisarvoja ja binäärisen vasteen tapauksessa ennusteprevalensseja olettaen, että kaikki yksilöt olisivat kuuluneet samaan ika6-muuttujan luokkaan. Katso esimerkkejä luvusta 7.

Jos halutaan estimoida lukumääriä perusjoukossa, niin tarvitaan *korottava* (eli väestö-) *painomuuttuja* w_unioni (ja w_ravinto), joka toimitetaan tutkimustiedoston mukana. Painojen w_unioni (ja w_ravinto) summa on tällöin Terveys 2000 -aineistossa 3 254 681, joka on yli 30-vuotiaiden kohdeperusjoukon koko, ja painomuuttujan arvo ilmaisee, kuinka montaa kohdeperusjoukon henkilöä otokseen poimittu kohdehenkilö vastaa. Tämän jälkeen esim. korkeasta (>140) systolisesta verenväenpaineesta kärsivien lukumääräestimaatit sukupuolittain ja ikäryhmittäin saadaan komennolla

```
proc descript data=work.t2k;
  setenv decwidth=4 colwidth=16;
  nest osite ryvas;
  weight w unioni;
  subgroup sp2 ika6;
  levels 2 6;
  var systbp2 01;
  tables sp2*ika6;
  print total settotal;
run;
```

Huomaa, että muuttuja systbp2_01 saa arvoja 0 tai 1.

5.2.1 Poikkileikkaustutkimus

Esimerkkejä jatkuvan ja binäärisen vasteen analysoimiseksi SUDAANilla. Analyysi rajataan Terveys 2000 -aineistoon käyttämällä komentoa subpopn t2k=1;. Malleissa oletetaan, että kokonaiskolesterolin vaikutus riippuu

sukupuolesta (sp2*T114), mutta muiden selittäjien välillä ei ole yhdysvaikutuksia.

Mallivakioidut keskiarvot lasketaan komentoa `predmarg ika6`; käyttäen.

5.2.1.1 Jatkuva vaste

```
proc regress data=work.t2k data;
  setenv decwidth=4 colwidth=13;
  nest osite ryvas;
  weight wan unioni;
  test satadjf;
  subpopn t2k=1 / name="Terveys 2000";
  subgroup t2k sp2 ika6 aa01 Portaannousu;
  levels 2 2 6 4 2;
  model systbp2 = bmi sp2*T114 ika6 aa01 Portaannousu;
  refllevel ika6=1;
  predmarg ika6;
run;
```

5.2.1.2 Binäärinen vaste

```
proc rlogist data=work.t2k data;
  setenv decwidth=4 colwidth=13;
  nest osite ryvas;
  weight wan unioni;
  test satadjf;
  subpopn t2k=1 / name="Terveys 2000";
  subgroup t2k sp2 ika6 aa01 Portaannousu;
  levels 2 2 6 4 2;
  model systbp2 01 = bmi sp2*T114 ika6 aa01 Portaannousu;
  refllevel ika6=1;
  predmarg ika6;
run;
```

5.2.1.3 Moniluokkainen vaste

Moniluokkaisen vasteen tapauksessa on huomattava, että vastemuuttuja on kuvattava subgroup- ja levels-lauseissa:

```
proc multilog data=work.t2k data;
  setenv decwidth=4 colwidth=13;
  nest osite ryvas;
  weight wan unioni;
  test satadjf;
  subpopn t2k=1 / name="Terveys 2000";
  subgroup systbp2 123 t2k sp2 ika6 aa01 Portaannousu;
  levels 3 2 2 6 4 2;
  model systbp2 123 = t2k*bmi sp2*T114 ika6 aa01 Portaannousu;
  refllevel ika6=1;
  predmarg ika6;
run;
```

5.2.2 Kahden otoksen vertailu

Aineistojen vertailussa malli on sama kuin poikkileikkaustilanteessa lukuun ottamatta oletusta, jonka mukaan BMI vaikuttaa systoliseen verenpaineeseen eri tavalla Mini-Suomi- ja Terveys 2000 –aineistoissa.

5.2.2.1 Jatkuva vaste

```
proc regress data=work.t2k data;
  setenv decwidth=4 colwidth=13;
  nest osite ryvas;
  weight wan unioni;
  test satadjf;
  subgroup t2k sp2 ika6 aa01 Portaannousu;
  levels 2 2 6 4 2;
```

```

model systbp2 = t2k*bmi sp2*T114 ika6 aa01 Portaannousu;
reflevel ika6=1;
run;

```

5.2.2.2 Binäärinen vaste

```

proc rlogist data=work.t2k data;
setenv decwidth=4 colwidth=13;
nest osite ryvas;
weight wan unioni;
test satadjf;
subgroup t2k sp2 ika6 aa01 Portaannousu;
levels 2 2 6 4 2;
model systbp2 01 = t2k*bmi sp2*T114 ika6 aa01 Portaannousu;
reflevel ika6=1;
run;

```

5.2.2.3 Moniluokkainen vaste

```

proc multilog data=work.t2k data;
setenv decwidth=4 colwidth=13;
nest osite ryvas;
weight wan unioni;
test satadjf;
subgroup systbp2 123 t2k sp2 ika6 aa01 Portaannousu;
levels 3 2 2 6 4 2;
model systbp2 123 = t2k*bmi sp2*T114 ika6 aa01 Portaannousu;
reflevel ika6=1;
run;

```

5.2.3 Toistomittaus

Toistomittauksen mallituksen voi suorittaa käyttämällä jotakin pitkittäistutkimusmenetelmää, esim. SAS-ohjelmiston `mixed` tai `genmod`-proseduureja `repeated`-optiolla.

5.3 Stata, versio 8

Statan komentoja voi suorittaa valitsemalla suoritettavat komennot *valikoista*, kirjoittamalla komentoja *komentoikkunaan* ("Stata Command") painaen `enter`- (eli rivinvaihto-) näppäintä sen jälkeen tai ajamalla **.do-komentotiedostoihin* talletettuja ohjelmia. Käytettäessä komentoikkunaa tai valikoita käytetyt komennot kirjautuvat "Review"-ikkunaan, josta ne voidaan tarvittaessa tallentaa komentotiedostoksi painamalla hiiren oikeanpuoleista painiketta ko. ikkunan päällä. Tällä keinolla voi säästää vaivaa, koska rutiininomaisia aineiston luku- ja otanta-asetelman kuvaamiskomentoja ei tarvitse kirjoittaa aina uudelleen: komentotiedoston voi ajaa valitsemalla tiedoston `File`-valikon kohdasta `Do...` Tiedostoja voi muokata editorilla, joka käynnistyy `Window`-valikon kohdasta `Do-file editor`. Editorin ylälaudassa on rivi painonappeja, joista toiseksi oikeanpuoleinen suorittaa tiedostossa olevat komennot. Mikäli komentoista on valittu ("maalattu") vain osa, niin vain ko. komennot suoritetaan.

Muistissa olevat muuttujat näkyvät ikkunassa "Variables", josta niitä voi valita hiiren vasemmanpuolisella painikkeella. Muuttujien arvoja voi katsella ja muokata `data-editorilla` (`Window/ Data-editor`).

Statan käyttämisestä ja komentoista löytyy apua `Help`-valikosta sekä painamalla nappia, jossa on kysymysmerkin kuva. On syytä huomata, että Stata ymmärtää isot ja pienet kirjaimet eri merkkeinä toisin kuin esim. SAS tai SUDAAN, eli Stata ei esim. ymmärrä muuttujanimien `Ika` tai `IKA`

tarkoittavan muuttujaa `ika`. Etenkin komentotiedostoja käytettäessä on hyvä huomata, että oletusarvoisesti yksittäistä komentoa ei voi jakaa useammalle riville.

Stata ei lue suoraan SAS-datatiedostoja, joten aineisto toimitetaan tarvittaessa Stata-muotoisena tiedostona (`*.dta`). Stata-datatiedostojen lukeminen tapahtuu komennolla `use` ja tallettaminen komennolla `save`, esim. `save c:\Data\T2000\oma` tallentaa aineiston tiedostoon `oma.dta`, jonka voi lukea komennolla `use c:\Data\T2000\oma`. Valikoita käyttämällä valitaan File/Open tai File/Save. Pienimpiä aineistoja lukuun ottamatta oletusarvona varattu muisti (1 megatavu) ei riitä, vaan aineiston lukeminen päättyy virheilmoitukseen. Lisää muistia voi varata komennolla esim. `set memory 20m, permanently`, joka varaa muistia 20 megatavua.

Otanta-asetelma ja painot määritellään seuraavasti:

```
svyset [pweight= wan unioni], strata(osite) psu(ryvas)
```

Valikoita käyttämällä valitaan Statistics/ Survey data analysis/ Setup & utilities. Statan help-toiminto kuvaa komennot varsin kattavasti, ja hakusanalla `svy` löytyy lista Statan tarjoamista otanta-aineiston analysointikomennosta.

Jos halutaan estimoida lukumääriä perusjoukossa, niin tarvitaan uudet painomuuttuja `w_unioni` (ja `w_ravinto`), joka toimitetaan tutkimustiedoston mukana. Painojen `w_unioni` (ja `w_ravinto`) summa on tällöin 3 254 681, joka on yli 30-vuotiaiden perusjoukon koko, ja painomuuttujan arvo ilmaisee, kuinka montaa perusjoukon henkilöä otokseen poimittu kohdehenkilö vastaa. Tämän jälkeen asetetaan otanta-asetelmamuuttujat komennolla `svyset [pweight=w_unioni], strata(osite) psu(ryvas)`, ja korkeasta (>140) systolisesta verenpaineesta kärsivien lukumääräestimaatit sukupuolittain ja ikäryhmittäin saadaan komennolla `svytotal systbp2_01, by(sp2 ika6)`. Huomaa, että muuttuja `systbp2_01` saa arvoja 0 tai 1. Komennot `svymean`, `svyratio` ja `svyprop` tuottavat muita perustunnuslukuja.

Komento `svytab` tuottaa kaksiulotteisia frekvenssitauluja. Komento löytyy valikosta Statistics/ Survey data analysis/ Setup & utilities.

Varoitus! `if` ja `in` -optioita analyysien rajaamiseksi johonkin osajoukkoon EI saa käyttää `svy`-proseduurien kanssa. Näiden sijaan pitää muodostaa 0/1-arvoinen muuttuja, ja käyttää `subpop()`-optiota `svy`-proseduurin kanssa. Perustunnuslukujen estimoinnissa voi käyttää myös `by()`-optiota moniluokkaisen luokittelijan yhteydessä. Esim. `svytotal systbp2_01, by(sp2 ika6)` tuottaa sukupuolen ja 6-luokkaisen ikämuuttujan mukaan korkeasta systolisesta verenpaineesta kärsivien lukumääräestimaatit vain Terveys 2000 -aineistossa sukupuolittain ja ikäryhmittäin. Valikoita käyttämällä valitaan Statistics/ Survey data analysis/ Univariate estimators. Taulukoita voi tuottaa `svytab`-komennolla.

Tilastollisia analyysejä varten tarvittavia komentoja ovat mm. `svylogit`, `svymlogit`, `svyreg` ja `predmarg` (joka on erillinen makro). Regressiomalleissa luokitellut selittäjät ja interaktiot määritellään käyttämällä `xi`-makroa. Esim. luokiteltu ikämuuttuja `ika6` ilmoitetaan luokitelluksi selittäjäksi terminä "i.ika6" ja yhdysvaikutus sukupuolen kanssa terminä "i.ika6*i.sp2" lineaarisessa regressiomallissa, jossa vasteena on systolinen verenpaine `systbp2` ja selittäjänä myös jatkuva-arvoinen painoindeksi `bmi`, komennolla `xi: svyregress systbp2 i.aa01 i.ika6*i.sp2 bmi`. Käytettäessä valikoita valitaan Statistics/ Survey data analysis/ Distribution-specific

models/ Linear regression for survey data. Vastemuuttuja kirjoitetaan kenttään `Dependent variable`. "xi:"-tekstiä ei kirjoiteta, vaan kenttään `Independent variables` kirjoitetaan tarvittavat "i."-lausekkeet suoraan, esim. `i.ika6 i.ika6*i.sp2 bmi`. Katso myös esimerkkejä jatkossa. Referenssiluokka luokitellulle selittäjälle asetetaan komennolla `char`, esim. komento `char ika6[omit] 1` asettaa muuttujan `ika6` referenssiluokaksi arvon 1.

Lineaarisen tai logistisen regressiomallin, ts. komentojen `svyregress` tai `svylogit` estimoinnin jälkeen voidaan erillisellä makrolla `predmarg` laskea mallivakioituja keskiarvoja ja prevalensseja. Makro on ladattavissa Terveys 2000 sivuilta osoitteesta

<http://www.ktl.fi/terveys2000/indexx.html>

kohdasta Ohjeita tutkijoille. Makro¹ ei tuota keskivirheitä lineaarisen mallin tapauksessa, mikäli vakioitavan muuttujan (tai muuttujien) ja muiden selittäjien välillä on yhdysvaikutus. Esim. komento `predmarg , ref(ika6=2)` tuottaa mallivakioidun tunnusluvun asettamalla kaikki yksilöt ikäluokkaan 2. Tunnusluku on keskiarvo tai prevalenssi riippuen aiemmin suoritetusta estimointikomennosta. `subpop()`-optiolla voidaan määritellä osajoukko, jonka pitää olla sama kuin edeltävässä `svyregress` tai `svylogit` -komennossa käytetty. Esim.

```
xi: svyregress systbp2 i.sp2 i.minis ika, subpop(nuoret50)
predmarg , ref(sp2=1) subpop(nuoret50)
predmarg , ref(sp2=2) subpop(nuoret50)
```

5.3.1 Poikkileikkaustutkimus

Poistetaan aluksi Mini-Suomi -aineisto komennolla `drop if t2k==0`.

5.3.1.1 Jatkuva vaste

`svyregress` suorittaa tavallisen lineaarisen regressioanalyysin. Esim.

```
xi: svyregress systbp2 bmi i.sp2*t114 i.ika6 i.aa01 Portaannousu
```

analysoi regressiomallin pelkästään Terveys 2000-aineistolla (`t2k` on 0/1-arvoinen muuttuja).

5.3.1.2 Binäärinen vaste

Esim.

```
xi: svylogit systbp2_01 bmi i.sp2*t114 i.ika6 i.aa01 Portaannousu
```

5.3.1.3 Moniluokkainen vaste

Luokiteltu ja järjestysasteikollinen logistinen regressioanalyysi suoritetaan komennoilla `svymlogit` ja `svyologit`, esim.

```
xi: svymlogit systbp2_123 bmi i.sp2*t114 i.ika6 i.aa01 Portaannousu
xi: svyologit systbp2_123 bmi i.sp2*t114 i.ika6 i.aa01 Portaannousu
```

¹ Makro toimitetaan tiedostossa `predmarg.ado`, joka kopioidaan johonkin Statan systeemihakemistoista (-kansioista), joiden listauksen saa Statan komennolla `sysdir`. Sopiva paikka on `PERSONAL`-hakemisto, joka on usein `c:\ado\personal`. Mikäli hakemistoa ei ole olemassa, se pitää luoda. Kun Stata käynnistetään seuraavan kerran, makro on käytettävissä.

5.3.2 Kahden otoksen vertailu

5.3.2.1 Jatkuva vaste

Svyregress suorittaa tavallisen lineaarisen regressioanalyysin. Esim. mallissa
xi: svyregress systbp2 i.t2k*bmi i.sp2*t114 i.ika6 i.aa01
i.Portaannousu

painoindeksin oletetaan vaikuttavan systoliseen verenpaineeseen eri tavalla Mini-Suomi- ja Terveys 2000 –aineistoissa.

5.3.2.2 Binäärinen vaste

Esim.

```
xi: svylogit systbp2_01 i.t2k*bmi i.sp2*t114 i.ika6 i.aa01  
i.Portaannousu
```

5.3.2.3 Moniluokkainen vaste

Esim. luokiteltu (svymlogit) ja järjestysasteikollinen (svyologit) logistinen regressioanalyysi suoritetaan komennoilla

```
xi: svymlogit systbp2_123 i.t2k*bmi i.sp2*t114 i.ika6 i.aa01  
i.Portaannousu  
xi: svyologit systbp2_123 i.t2k*bmi i.sp2*t114 i.ika6 i.aa01  
i.Portaannousu
```

5.3.2.4 Toistomittaus

Paneeliaineistojen analysoimiseen tarkoitettuja xt-komentoja voi käyttää, kuten esim. xtglm, xtreg ja xtlogit. Myös ehdollisia logistisia malleja estimoivaa clogit-komentoa voi käyttää.