Surveymetodiikka

Otanta-asetelman käyttämisestä SAS.ssa ja SPSS:ssä

Seppo 8.10.2014

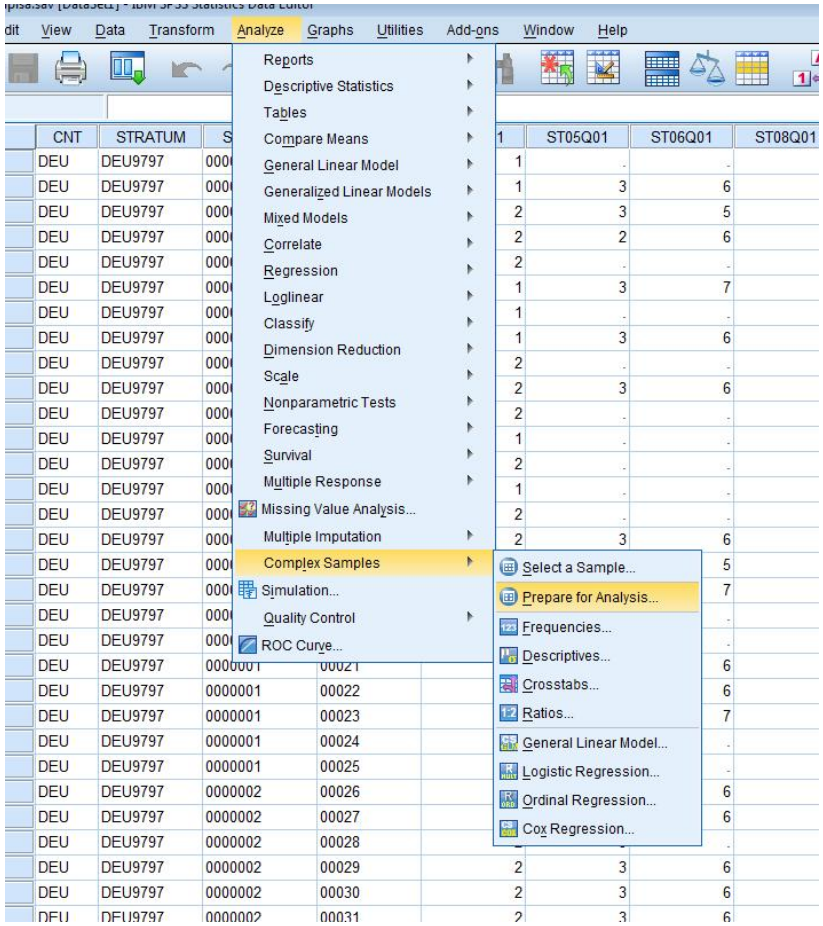
Yleisesti: Kummassakin ohjelmistossa pitää asettaa ajoon mukaan otanta-asetelmaa koskevat tiedot eli muuttujat sellaisina kuin ne ovat datassa. ESS:ssä on valitettavasti vain yksi muuttuja eli jälkiosituspohjainen otospaino joka on skaalattu siten että kussakin maassa painojen keskiarvo = 1. On mahdollista muuttaa tämä painoksi joiden summa kussakin maassa on tavoiteperusjoukon (15+ -vuotiaiden) koko, kuten on toisessa kalvossa esitetty.

Huomaa kuitenkin että joissakin maakohtaisissa aineistoissa on samat otanta-asetelmamuuttujat kuin PISAssa. Esimerkkinä on Espanja, jolle olevat tiedot RYPPÄÄSTÄ (PSU) ja OSITTEESTA (STRATIFY) löydät sivulta:
http://www.europeansocialsurvey.org/data/download.html?r=5
Jos joku haluaa tehdä analyysiä jonkin maan ESS-datalla se sopii. Mutta kaikille maille ei ole samoja tietoja.

Otanta-asetelman käytöstä SAS.ssa ja SPSS:ssä

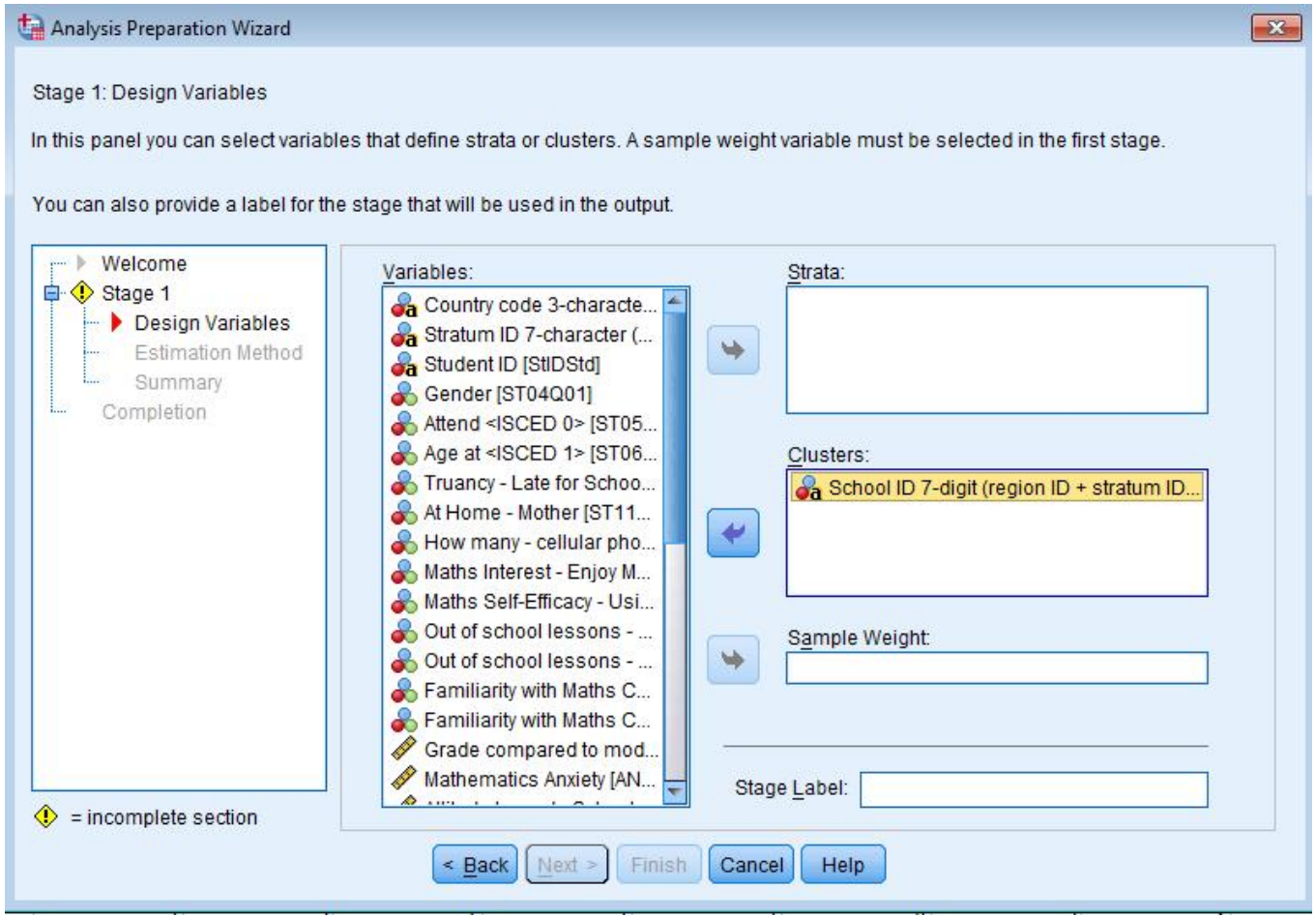
PISAssa on valmiina oppilaspaino jonka summa kussakin maassa on ’PISA-oppilaiden’ määrä. Laske harjoitukseksi Suomen painojen summa, joka vastaa siis PISA-tavoiteperusjoukon kokoa. PISAssa on myös kaksi muuta otanta-asetelman muuttujaa eli otoksessa (sen ensimmäisessä asteessa) käytetty ryväs = koulutunnus = schoolid sekä ositetunnus stratum.

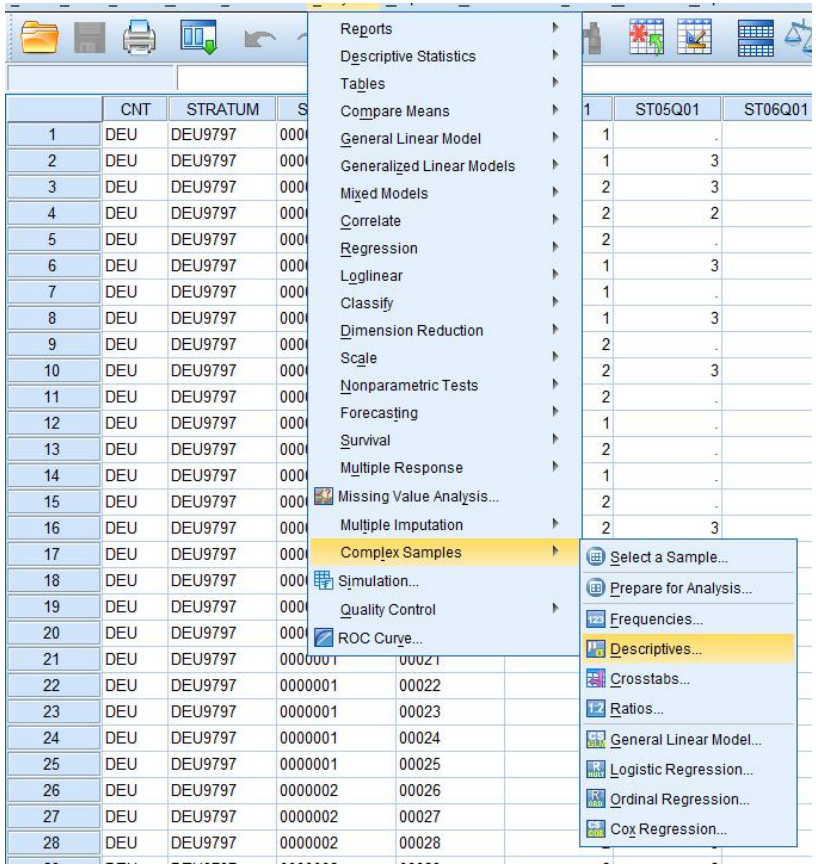
Analyysi on nyt helppoa. SAS:ssa on valittava sopiva ohjelmisto, mikä aina alkaa sanalla ’survey’. Niinpä jos lasketaan keskiarvoja, käytetään PROC SURVEYMEANSia ja vastaavasti SURVEYFREQiä sekä malleissa SURVEYREGiä tai SURVEYLOGISTICia. Kaikissa näissä toimitaan kuten muutenkin mutta lisätään seuraavat ’optiot’ tai asetukset:
WEIGHT <otospaino>; CLUSTER <ryväs>; STRATUM <osite>;
Jos SURVEYMEANSllä haluaa keskiarvoja taustatekijöiden mukaan aseta DOMAIN <taustamuuttujien nimet peräkkäin>; Jos samaa haluaa SURVEYFREQllä aseta BY <taustamuuttuja>;
Analyyseihin on toki erilaisia lisämahdollisuuksia mutta automaattisesti saat tietyt tulokset jotka näet kun kokeilet.
SPSSssä sama tehdään hieman toisella tavalla kuten näet seuraavien sivujen kuvasarjasta.



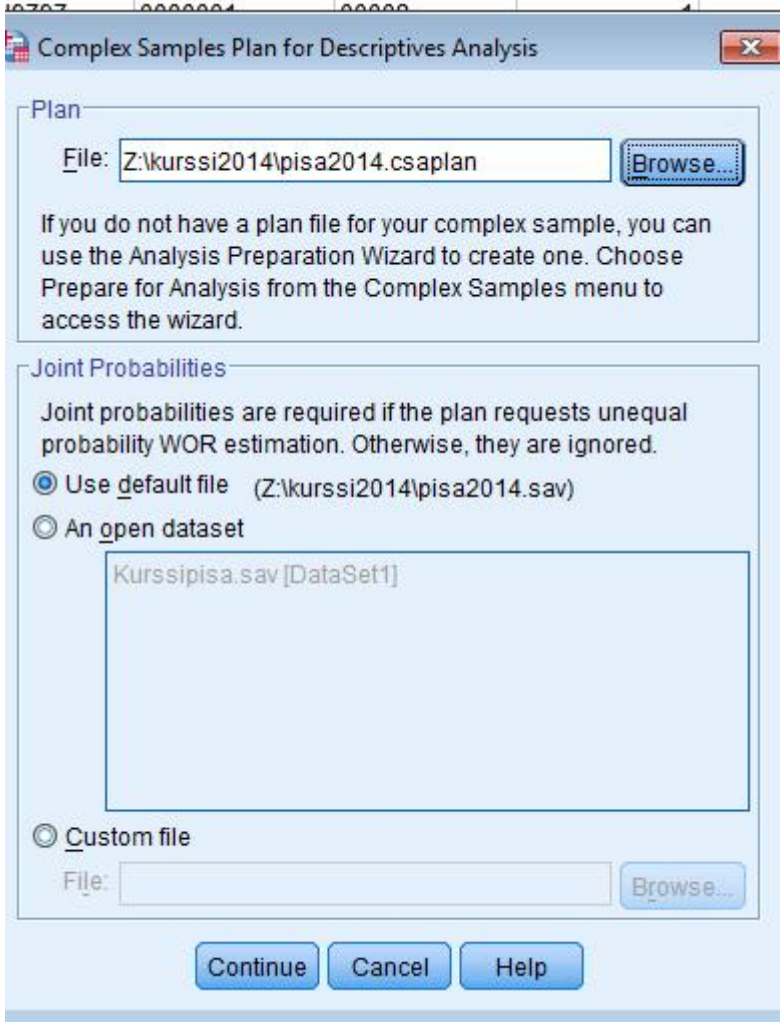
Aloita siis tuosta Prepare for Analysis klikkaamalla. Avautuu uusi ikkuna, alla. Aseta ensin otanta-asetelmallesi sopiva nimi ja säästä se paikkaan josta löydät sen. Sitten klikkaa 'next'.
SPSS:n filosofia on siis hieman erilainen. Ensin päätetään otanta-asetelman muuttujat ja sitten mennään analyyseihin joiden vaihtoehdot näkyvät myös oheisesta ikkunasta. Harjoituksissa kokeilemme FREQUENCIES, DESCRIPTIVES, GENERAL LINEAR MODEL ja jos aikaa on LOGISTIC REGRESSION

Nyt aseta kuhunkin lokeroon oikea muuttuja. Olen laittanut vain ryväsmuuttujan, mutta jatka siitä. Kun siellä on niitä riittävästi, boxi ’Finish’ aktivoituu alempana. Silloin paina sitä ja sulla on tarvittava tieto tallessa. Jos on tarpeen, voit luoda niitä lisää ja erilaisia (esim. molemmilla painoilla). NYT olet valmis varsinaiseen analyysiin. Sitä varten palaa alkuun ja valitse listasta sopiva ohjelma, seuraava sivu. Huomaa etten ole täyttänyt esimerkkiä loppuun. Toki jollet halua kaikkia käyttää, niin tulos muuttuu.

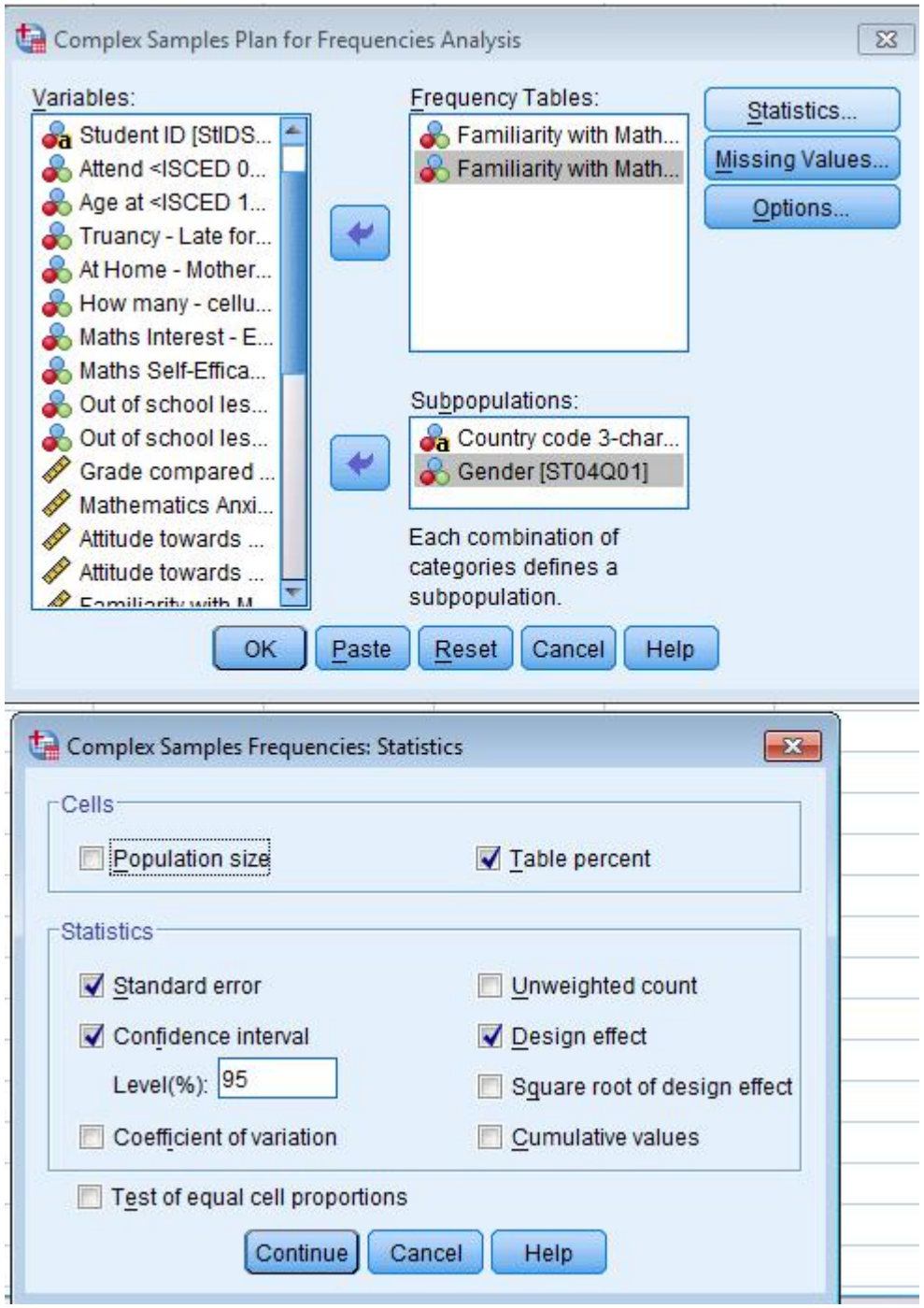




Tässä olen asettanut ’liipasimelle’ ohjelman jolla saa mm. keskiarvoja estimoitua ja niille keskivirheet, luottamusvälit ja DEFF-luvut esimerkiksi. Ohjelma ehdottaa analyysiin edellä luotua otanta-asetelmatiedostoa. Jos haluat vaihtaa sen, vaihda. Kun olet tyytyväinen paina ’Continue’ ja sulle avautuu ikkuna varsinaiseen analyysiisi.



Tätä keskiarvo-ohjelmaa en vie loppuun vaan annan Jatkossa frekvenssi-ohjelman. Molemmat ovat saman-tapaisia.



Tässä FREQUENCIES –ohjelmassa olen valinnut analyysiini kaksi kategorista muuttujaa ja pyydän saada ne sukupuolen ja syntymämaan. Kun klikkaan ’Statistics’-boksia, voin valita lisää tulostevaihtoehtoja. Yleensä luottamusväli kannattaa ottaa ja myös DEFF jotta siihen saisi tuntumaa (SPSS:n DEFFiä on tosin kritisoitu mutta ei sun tarvi sitä julkaista).

Nyt on vain jäljellä klikkauksia lisää ’Continue’ > ’OK’ ja voit ryhtyä odottamaan upeita tuloksia jotka on hyvä ymmärtää. Huomaa että ohjelmaa tuntuu tarjoavan boksia ’Population size’ mikä ei ole järkevä yleensä vaan optio ’Table percent’ josta yleensä ollaan kiinnostuneita, ei määristä eli ’Population size’ mikä on sen sijaan keskiarvoissa hyvin järkevää.

Laitan tähän myös miten keskiarvot ja frekvenssit saadaan SAS:lla. Huomaa että frekvenssseja kannattaa laskea vain kategorisille muuttujille. Jos haluat kategorisoida osaamismuuttujia niin valitse jokin taitekohta, kuten esimerkiksi
Jos osaaminen >=600 niin Osaaminen='Mainio', muuten ='Vähemmän kuin mainio'. Syy: koska 600 on keskihajonnan päästä keskiarvosta eli varsin hyvä osaaja on kyseessä.
Tätä voisi käyttää jatkossa mm. logistisessa regressiomallissa.

```
proc surveymeans data=z.pisa2012;
domain cnt *gender;
var math_stat reading science problem;
cluster schoolid; stratum stratum; weight W_FSTUWT;
run;

proc sort; by cnt;
proc surveyfreq data=z.pisa2012;
tables ST08Q01 ST24Q05 ST34Q02;
cluster schoolid; stratum stratum; weight W_FSTUWT;
by cnt; run;
```

Johdatus tilastollisiin malleihin

Tilastollisia malleja voi olla erilaisia:
- Ristiintaulukko on oikein tehtynä luonteeltaan tilastollinen malli. Esimerkiksi teksti ’iän mukaan’ tarkoittaa että selittäjänä on ikä. Keksi itse lisää esimerkkejä.

- Kuvailumalli on aina tehtävissä eli voidaan ’kuvailla’ tai havainnollistaa eri muuttujien välisiä yhteyksiä. Tämä siis vaadi välttämättä suurta syy-seuraus-suhteiden analyysiä eli kyse ei ole kausaalimallista, joka on mietittävä huolellisesti. Tätä mallityyppiä voisi kutsua myös ’uteliaisuuden tyydyttämismalliksi.’
- Kausaali- ja selitysmalli edellyttää että on olemassa perustelut sille miten selittävä muuttuja (selittäjä) liittyy selitettävään muuttujaan. Perustelu löydetään usein tieteen teoriasta eli siis tutkimalla aihealueen lähteitä. Ei silti tarvitse olla niin, että teorian mukainen selittäjä on tilastollisesti merkitsevä tai että sen merkki on teorian mukainen. Syy voi olla huonossa aineistossa, huonossa mallittajassa tai jossain määrin myös itse teoriassa. Voisit saada nimesi ansiokkaasti esille jos kykenisit esittämään vaihtoehtoisen teorian tunnetulle ja arvostetulle olemassa olevalle teorialle.

Johdatus tilastollisiin malleihin

Tilastollisia malleja voi olla erilaisia:

- Ennustemalli tähtää nimensä mukaisesti siis ennustamaan. Aina ei havaita, että ennustamista on ainakin kahta laatua:
    - Tulevaisuuden ennustaminen
    - Menneisyyden ja nykyisyyden ennustaminen.

Tulevaisuuden ennustaminen vaatii arviota siitä mitä tulevaisuudessa voi tapahtua ennustemallin selittäjien osalta. Yksinkertaisin vaihtoehto on olettaa että kaikki menee aikaisemman trendin mukaan eli jatkuu samankaltaisena. Toinen, vaativampi vaihtoehto on miettiä sopivassa tiimissä tulevaisuuden skenaarioita näille selittäjille. Tällöin saadaan vaihtoehtoisia skenaarioita eli tulevaisuuden kuvia.

Surveyn mikrodatalla harvemmin pyritään tulevaisuuden ennustamiseen suoranaisesti vaan sen sijaan nykyisyyden ja lähimenneisyyden arvioimiseen jostakin tavoitteesta lähtien:

Johdatus tilastollisiin malleihin

Surveyn mikrodatalla harvemmin pyritään tulevaisuuden ennustamiseen suoranaisesti vaan sen sijaan nykyisyyden ja lähimenneisyyden arvioimiseen jostakin tavoitteesta lähtien:

- Kuten muistat otanta-asetelman suunnittelussa piti ennustaa miten käy vastausasteen ja ylipeiton ja myös miten korkea sisäkorrelaatio? DEFF-lukuun voidaan myös vaikuttaa ryväsotoskoolla.
- Lopussa meillä on esillä Uudelleenpainotus, jossa yksi menetelmä perustuu ennustamaan vastaustodennäköisyyttä. Tätä varten rakennetaan logistinen tai probit-malli jolla ennustetaan vastaustaipumusta käsillä olevan datan eli nykydatan avulla. Tässä rakennetaan siis hyvä ennustemalli, ei tarvitse välittää kausaliteetista ainakaan kovin paljoa jos malli vaan toimii.
- Lopussa on myös esillä Imputointi, jonka yhtenä osana on rakentaa Imputointimalli. Tämä on myös puhtaasti nykydataan perustuva ennustemalli. Jos hyvin ennustava malli saadaan, voidaan olla onnellisia.
- On paljon muita nykyennusteita, joita saadaan tilastollisilla malleilla. Pääosa tilastollisista malleista onkin juuri näitä. Toki usein uskotaan että sama ’totuus’ usein tapahtuu myös tulevaisuudessa.

Johdatus tilastollisiin malleihin

Esimerkkejä löytyy vaikka kuinka. Voit itse keksiä lisää:

- Tupakointi on aiheuttanut vaikka mitä sairauksia ja niin tapahtunee tulevaisuudessakin ellei keksitä jotain ihmeellistä vaaratonta tupakkaa.
- Huumeita koskevat samat näkökohdat.
- Onnellisuuteen vaikuttavat sekä ulkoiset että sisäiset tekijät eikä lopullista totuutta tiedetä koskaan koska sisäisiä tekijöitä on hankala ennustaa. Monet ulkoiset tekijät sen sijaan kertautuvat kerta toisensa jälkeen surveydatalla.
- Palkkakäyrä iän mukaan on jollain tavalla parabolinen eli työuran alkupäässä matalahko ja jossain 45-50 vuoden tienoilla korkeimmillaan.
- Koulutuksen laatu ja määrä nostaa tulotasoa mutta lineaarinen se ei ole.
- Hyväveli- ja hyväsisarverkostot edistävät monia asioita näihin verkkoihin kuuluvien osalta.
- Surveymetodiikan perusteiden tuntemisesta on iloa ja hyötyäkin.

Johdatus tilastollisiin malleihin

Harjoituksia silmällä pitäen muutama näkökohta.
Tarkoitus on tehdä survey-mielessä oikeaoppisia malleja. Voit pyrkiä siinä ainakin kuvailumalliin mutta yritä miettiä selittäjiä myös kausaalisesti. Ennustamismielessäkin voit mallia tehdä.

Mallilla on aina jokin sen ’voimaa’ kuvaava mittari, kuten selitysaste (R-SQUARE tai R Squared). Sitä on hyvä aina katsoa ja pyrkiä nostamaan sen tasoa

- Ottamalla malliin mukaan uusia hyviä muuttujia
- Pyrkimällä muuntamaan olemassa olevia paremmiksi jos mallista
- Kokeilemalla yhdysvaikutuksia tyypillisesti kahden muuttujan välillä (vaikkapa ikäryhmän ja sukupuolen).

Selittäjien kertoimien tulkinta on minusta kiinnostavinta koska ne kertovat miten selittäjä liittyy selitettävään muuttujaan.

Johdatus tilastollisiin malleihin

Selittäjien kertoimien tulkinta on minusta kiinnostavinta koska ne kertovat miten selittäjä liittyy selitettävään muuttujaan.

Jatkuvan muuttujan osalta tulkinta on helpohko. Katso kertoimen merkki ja onko se merkitsevä ja tulkitse ottaen huomioon selitettävän muuttujan luonteen.

Kategorinen muuttuja voidaan malliin rakentaa eri tavoilla. Kaikissa on ideana verrata kun kategorian kerroinestimaattia johonkin vertailuestimaattiin. Tuo vertailu voidaan asettaa esimerkiksi kaikkien kategorioiden keskiarvoksi mutta perusohjelmisto ei tue tätä automatiikalla. Automatiikassa sen sijaan vertailukategoriaksi otetaan viimeinen kategoria, johon kaikkia muita verrataan. Siis tältä pohjalta lasketaan estimaatin keskivirhe (standard error), t-arvo, p-arvo (SPSS:ssä 'Sig') ja muut tilastolliset tunnusluvut. Jos siis vertailukategoria vaihtuu, kaikki nämä muuttuvat. Yleensä kuitenkin kuvailumallissa ja selitysmallissa tämä riittää. Ennustemallissa taas näistä ei olla kiinnostuneita, vaan itse ennusteista.

Esimerkki PISA-datan regressiomallista jossa selitettävä näkyy alempana. Selittäjiä melko vähän, ja mm. maa puuttuu.

Parameter Estimates[a]

Parameter	Estimate	Std. Error	Design Effect
(Intercept)	410,710	8,148	5,367
[FAMSTRUC=1]	37,126	7,914	4,155
[FAMSTRUC=2]	70,824	8,481	4,500
[FAMSTRUC=3]	,000[b]	.	.
[IMMIG=1]	33,049	4,869	7,026
[IMMIG=2]	2,497	5,136	5,596
[IMMIG=3]	,000[b]	.	.
[gender=1]	33,825	1,711	5,465
[gender=2]	,000[b]	.	.
[ST08Q01=1]	-13,689	4,786	3,560
[ST08Q01=2]	,000[b]	.	.
ATSCHL	6,789	,948	7,040

a. Model: Plausible value in reading - reflect and evaluate = (Intercept) + FAMSTRUC + IMMIG + gender + ST08Q01 + ATSCHL

b. Set to zero because this parameter is redundant.

Perherakenne
FAMSTRUCT kolme kategoriaa
1 (Single Parent) 37,126
2 (Nuclear) 70,824
3 (Mixed) ,000 on vertailuryhmä muille

Vastaavasti kolmelle muulle kategoriselle
Tulkitse ne itse

Ja lopussa on jatkuva muuttuja Asenne koulua kohtaan jonka arvo on positiivinen eli positiivinen asenne auttaa osaamisessa.

Huomaa että DEFF –luvut aika korkeita eli otanta-asetelmamuuttujien mukana ololla on väliä. Siis keskivirhe kasvaa oikeutetusti niiden kautta.