

## 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 PISA:ssa. Esimerkkinä on Espanja, jolle ovat 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 SURVEYMEANSilla haluaa keskiarvoja taustatekijöiden mukaan aseta DOMAIN <taustamuuttujien nimet peräkkäin>; Jos samaa haluaa SURVEYFREQillä 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 analyysiin joiden vaihtoehdot näkyvät myös oheisesta ikkunasta. Harjoituksissa kokeilemme FREQUENCIES, DESCRIPTIVES, GENERAL LINEAR MODEL ja jos aikaa on LOGISTIC REGRESSION

The screenshot shows the SPSS software interface. The 'Analyze' menu is open, displaying various statistical analysis options. The 'Complex Samples' option is highlighted, and its sub-menu is visible, showing options like 'Select a Sample...', 'Prepare for Analysis...', 'Frequencies...', 'Descriptives...', 'Crosstabs...', 'Ratios...', 'General Linear Model...', 'Logistic Regression...', 'Ordinal Regression...', and 'Cox Regression...'. The 'Prepare for Analysis...' option is selected. In the background, a data table is visible with columns labeled 'CNT', 'STRATUM', and 'S'. The table contains data for various cases, including 'DEU' and 'DFU' with their respective 'CNT' and 'STRATUM' values.

The screenshot shows the 'Analysis Preparation Wizard' dialog box. The window title is 'Analysis Preparation Wizard'. The main text reads: 'Welcome to the Analysis Preparation Wizard. The Analysis Preparation Wizard helps you describe your complex sample and choose an estimation method. You will be asked to provide sample weights and other information needed for accurate estimation of standard errors. Your selections will be saved to a plan file that you can use in any of the analysis procedures in the Complex Samples Option.' Below this text, there is a section titled 'What would you like to do?' with three radio button options: 'Create a plan file', 'Edit a plan file', and 'If you already have a plan file you can skip the Analysis Preparation Wizard and go directly to any of the analysis procedures in the Complex Samples Option to analyze your sample.' The 'Create a plan file' option is selected. To the right of the 'Create a plan file' option, there is a text field labeled 'File:' containing the path 'Z:\kursssi2014\pisa2014.csaplan' and a 'Browse...' button. To the right of the 'Edit a plan file' option, there is a text field labeled 'File:' and a 'Browse...' button. At the bottom of the dialog box, there are buttons for '< Back', 'Next >', 'Finish', 'Cancel', and 'Help'.

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.

The screenshot shows the 'Analysis Preparation Wizard' window, specifically 'Stage 1: Design Variables'. The window title is 'Analysis Preparation Wizard' and it has a close button in the top right corner. The main text reads: 'In this panel you can select variables that define strata or clusters. A sample weight variable must be selected in the first stage.' Below this, it says: 'You can also provide a label for the stage that will be used in the output.'

On the left, there is a navigation pane with a tree view showing the following steps: 'Welcome', 'Stage 1' (with a yellow warning icon), 'Design Variables' (highlighted), 'Estimation Method', 'Summary', and 'Completion'. A legend at the bottom left indicates that a yellow warning icon means '= incomplete section'.

The main area is divided into several sections:

- Variables:** A list of variables with a scroll bar. The visible variables are: Country code 3-character..., Stratum ID 7-character (...), Student ID [StIDStd], Gender [ST04Q01], Attend <ISCED 0> [ST05..., Age at <ISCED 1> [ST06..., Truancy - Late for Schoo..., At Home - Mother [ST11..., How many - cellular pho..., Maths Interest - Enjoy M..., Maths Self-Efficacy - Usi..., Out of school lessons - ..., Out of school lessons - ..., Familiarity with Maths C..., Familiarity with Maths C..., Grade compared to mod..., and Mathematics Anxiety [AN....
- Strata:** An empty text box for selecting strata variables.
- Clusters:** A text box containing the selected variable 'School ID 7-digit (region ID + stratum ID...'. There are blue arrows pointing from the 'Variables' list to this box.
- Sample Weight:** An empty text box for selecting a sample weight variable.
- Stage Label:** A text box with the label 'Stage Label:' and an empty input field.

At the bottom of the window, there are five buttons: '< Back', 'Next >', 'Finish', 'Cancel', and 'Help'. The 'Finish' button is currently disabled.

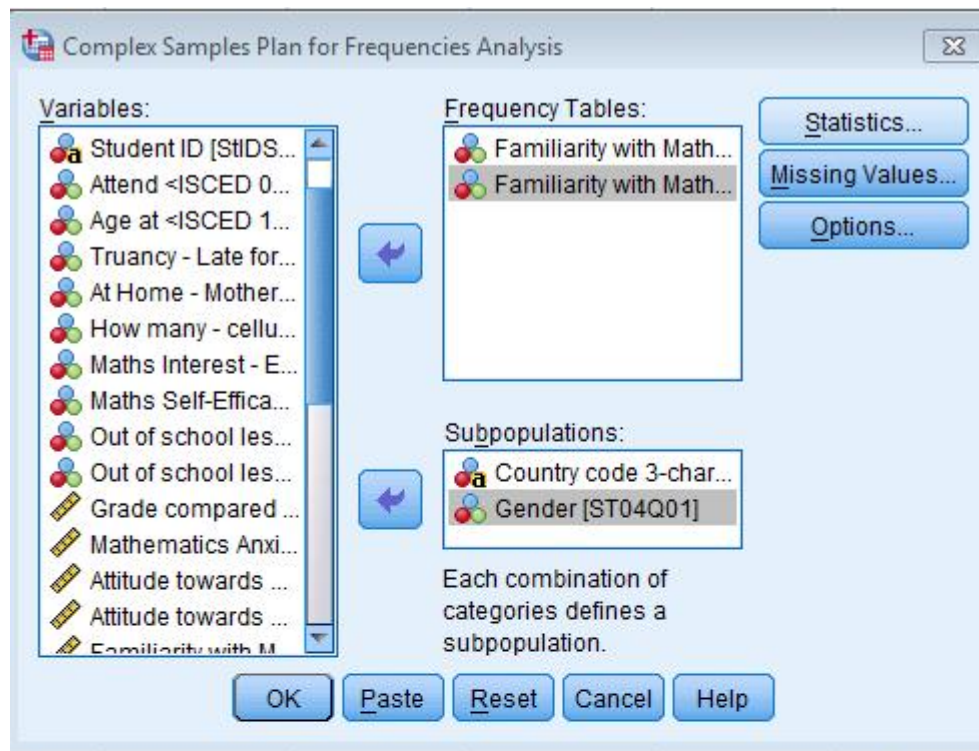
The screenshot shows the SPSS interface with a data list on the left and the 'Complex Samples' menu open on the right. The data list has columns for 'CNT', 'STRATUM', and 'S'. The menu options include 'Select a Sample...', 'Prepare for Analysis...', 'Frequencies...', 'Descriptives...', 'Crosstabs...', 'Ratios...', 'General Linear Model...', 'Logistic Regression...', 'Ordinal Regression...', and 'Cox Regression...'.

	CNT	STRATUM	S
1	DEU	DEU9797	000
2	DEU	DEU9797	000
3	DEU	DEU9797	000
4	DEU	DEU9797	000
5	DEU	DEU9797	000
6	DEU	DEU9797	000
7	DEU	DEU9797	000
8	DEU	DEU9797	000
9	DEU	DEU9797	000
10	DEU	DEU9797	000
11	DEU	DEU9797	000
12	DEU	DEU9797	000
13	DEU	DEU9797	000
14	DEU	DEU9797	000
15	DEU	DEU9797	000
16	DEU	DEU9797	000
17	DEU	DEU9797	000
18	DEU	DEU9797	000
19	DEU	DEU9797	000
20	DEU	DEU9797	000
21	DEU	DEU9797	0000001
22	DEU	DEU9797	0000001
23	DEU	DEU9797	0000001
24	DEU	DEU9797	0000001
25	DEU	DEU9797	0000001
26	DEU	DEU9797	0000002
27	DEU	DEU9797	0000002
28	DEU	DEU9797	0000002

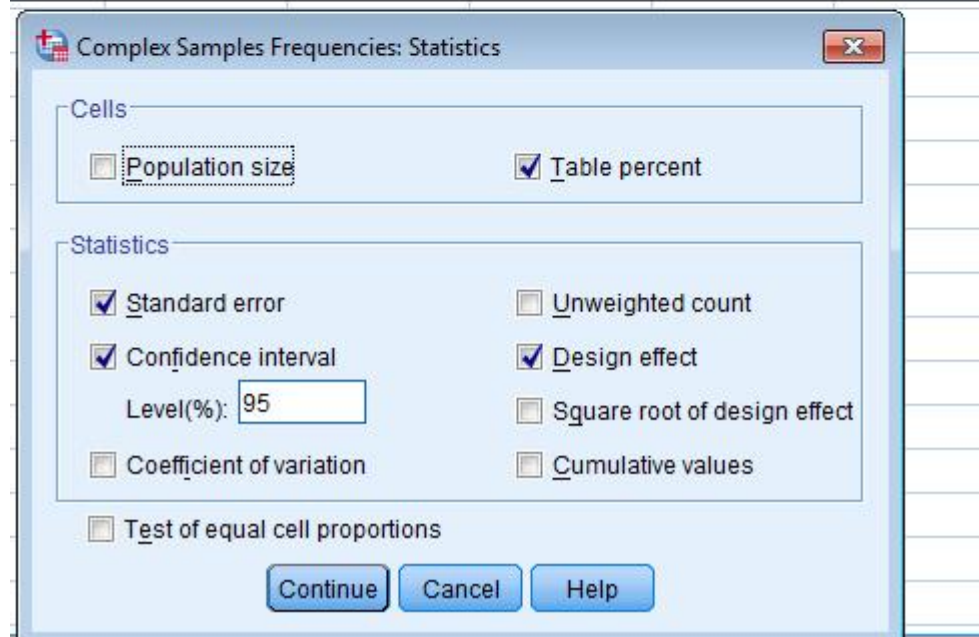
Tässä olen asettanut 'liipasimelle' ohjelman jolla saa mm. keskiarvoja estimoitua ja niille keskivirheet, luottamusvälit ja DEFF-luvut esimerkiksi. Ohjelma ehdottaa analyysiin edellä luotua otanta-asetelmatiedostoa. Jos haluat vaihtaa sen, vaihda. Kun olet tyytyväinen paina 'Continue' ja sulle avautuu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

The dialog box is titled 'Complex Samples Plan for Descriptives Analysis'. It has a 'Plan' section with a 'File:' field containing 'Z:\kurssi2014\pisa2014.csaplan' and a 'Browse...' button. Below this is a text box: 'If you do not have a plan file for your complex sample, you can use the Analysis Preparation Wizard to create one. Choose Prepare for Analysis from the Complex Samples menu to access the wizard.' The 'Joint Probabilities' section has a text box: 'Joint probabilities are required if the plan requests unequal probability WOR estimation. Otherwise, they are ignored.' There are two radio buttons: 'Use default file (Z:\kurssi2014\pisa2014.sav)' which is selected, and 'An open dataset'. Below the second radio button is a text box containing 'Kurssipisa.sav [DataSet1]'. At the bottom, there is a 'Custom file' radio button, a 'File:' field, and a 'Browse...' button. At the very bottom are 'Continue', 'Cancel', and 'Help' buttons.

Tätä keskiarvo-ohjelmaa en vie loppuun vaan annan jatkossa frekvenssi-ohjelman. Molemmat ovat samantapaisia.



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 frekvensseja kannattaa laskea vain kategorisille muuttujille. Jos haluat kategorisoida osaamismuuttujia niin valitse jokin taitekohta, kuten esimerkiksi Jos osaaminen  $\geq 600$  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 vaihtoehdoisen 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 tulotasoja 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 mieltä 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<sup>a</sup>

Parameter	Estimate	Std. Error	Design Effect	Perherakenne
(Intercept)	410,710	8,148	5,367	FAMSTRUCT kolme kategoriaa
[FAMSTRUC=1]	37,126	7,914	4,155	1 (Single Parent) 37,126
[FAMSTRUC=2]	70,824	8,481	4,500	2 (Nuclear) 70,824
[FAMSTRUC=3]	,000 <sup>b</sup>	.	.	3 (Mixed) ,000 on vertailuryhmä muille
[IMMIG=1]	33,049	4,869	7,026	Vastaavasti kolmelle muulle kategoriselle Tulkitse ne itse
[IMMIG=2]	2,497	5,136	5,596	
[IMMIG=3]	,000 <sup>b</sup>	.	.	
[gender=1]	33,825	1,711	5,465	Ja lopussa on jatkuva muuttuja Asenne koulua kohtaan jonka arvo on positiivinen eli positiivinen asenne auttaa osaamisessa.
[gender=2]	,000 <sup>b</sup>	.	.	
[ST08Q01=1]	-13,689	4,786	3,560	
[ST08Q01=2]	,000 <sup>b</sup>	.	.	
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.

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