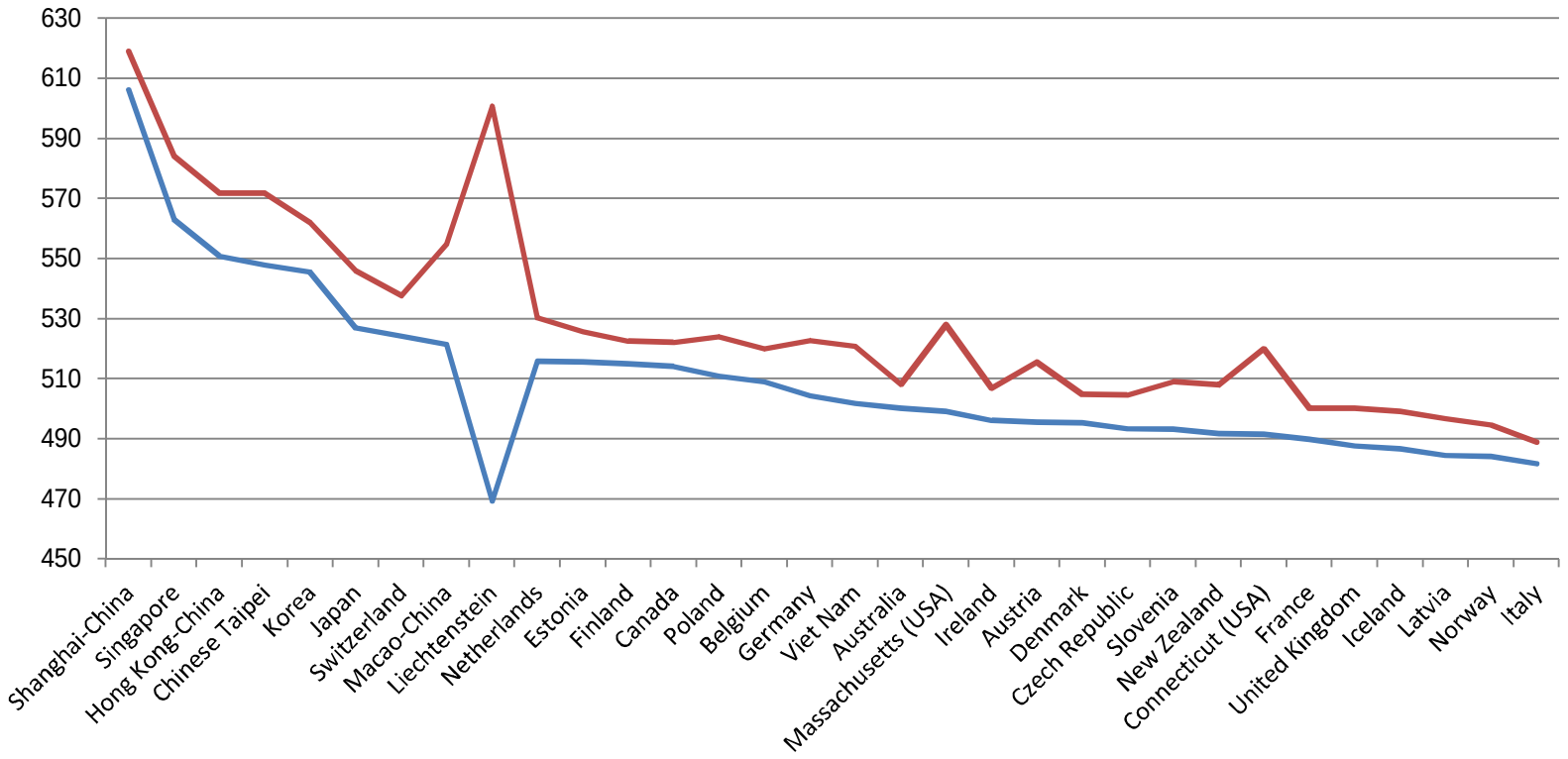
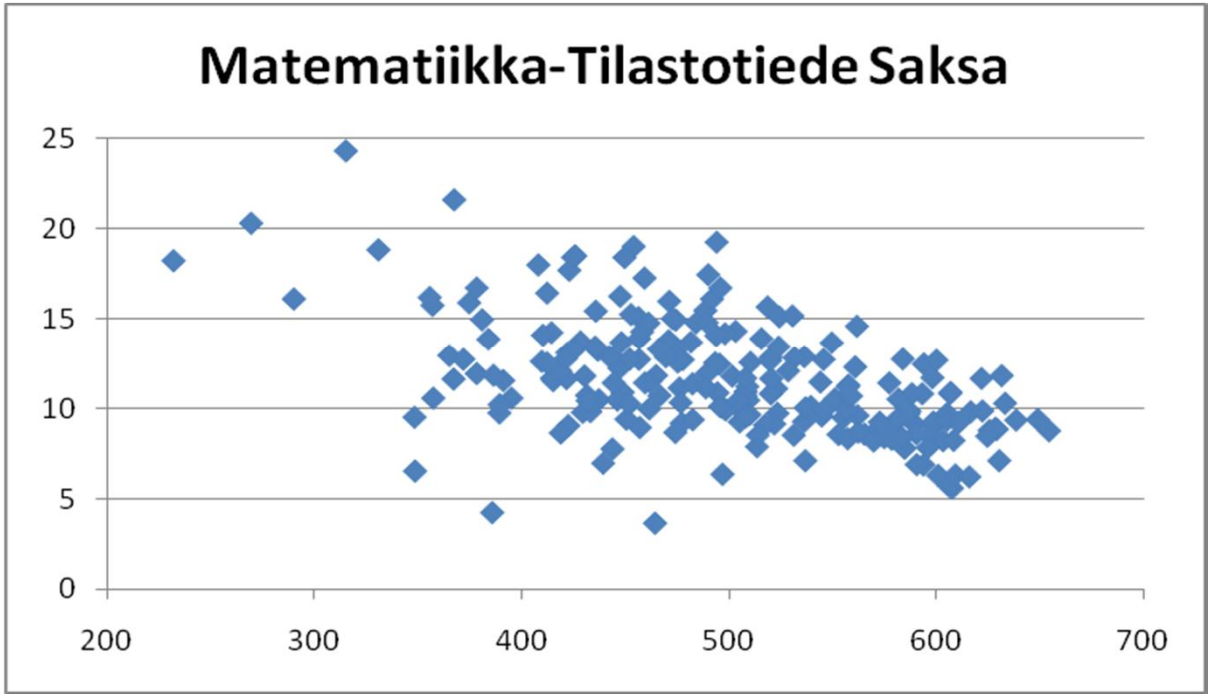
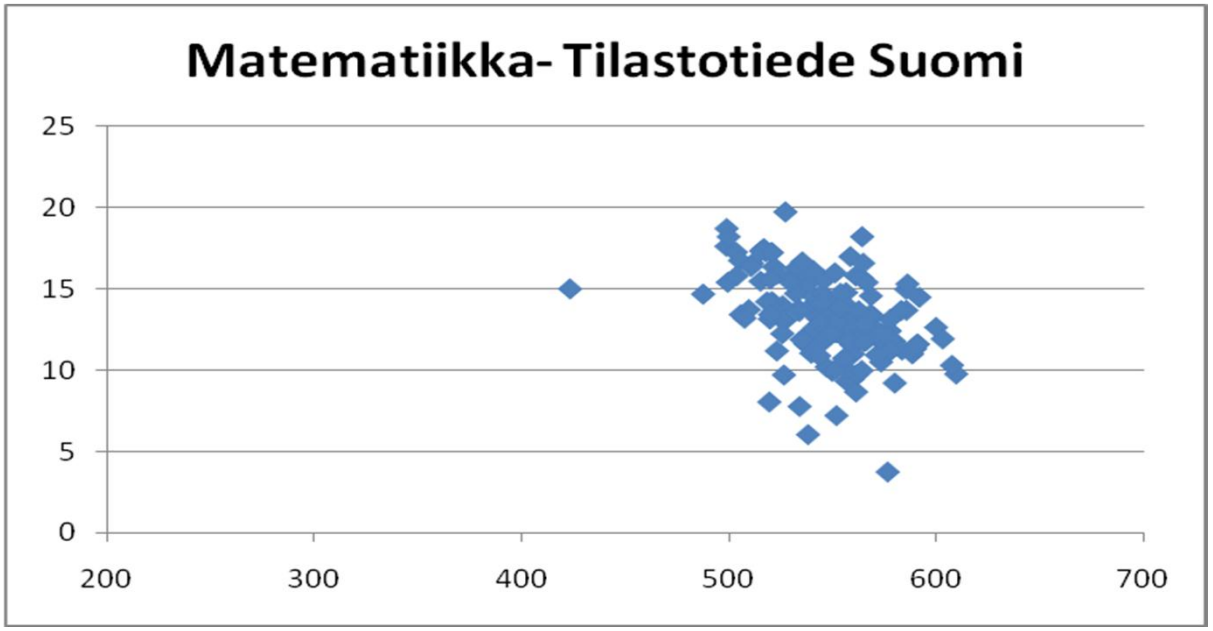
Tässä lisäkekalvopaketissa on muutama esimerkki,
aluksi luottamusvälistä ja sitten puuttuvuudesta
On sekä Pisa- että ESS-esimerkkejä

Graafinen esitystapa jonkalaisia myös Pisa-raportointi käyttää
Pisa 2012: Matemaattis-tilastollisen lukutaidon kärki
Ylempi viiva: 95% luottamusvälin yläraja
Alempi viiva: 95% luottamusvälin alaraja
Järjestys keskiarvon mukaan mikä on yleinen tapa
Tehtävä: Kuinka MONES on Suomi?



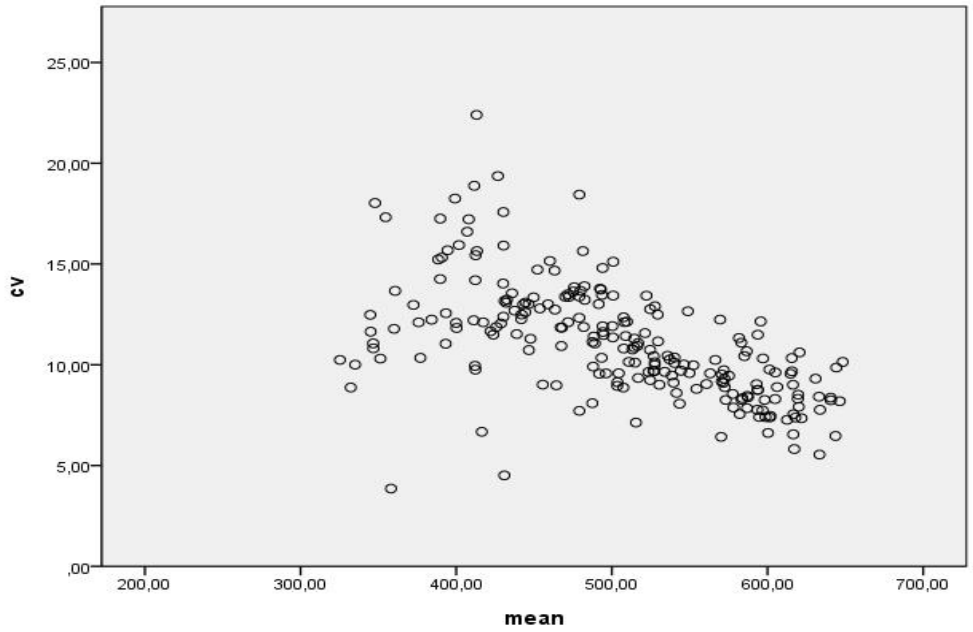


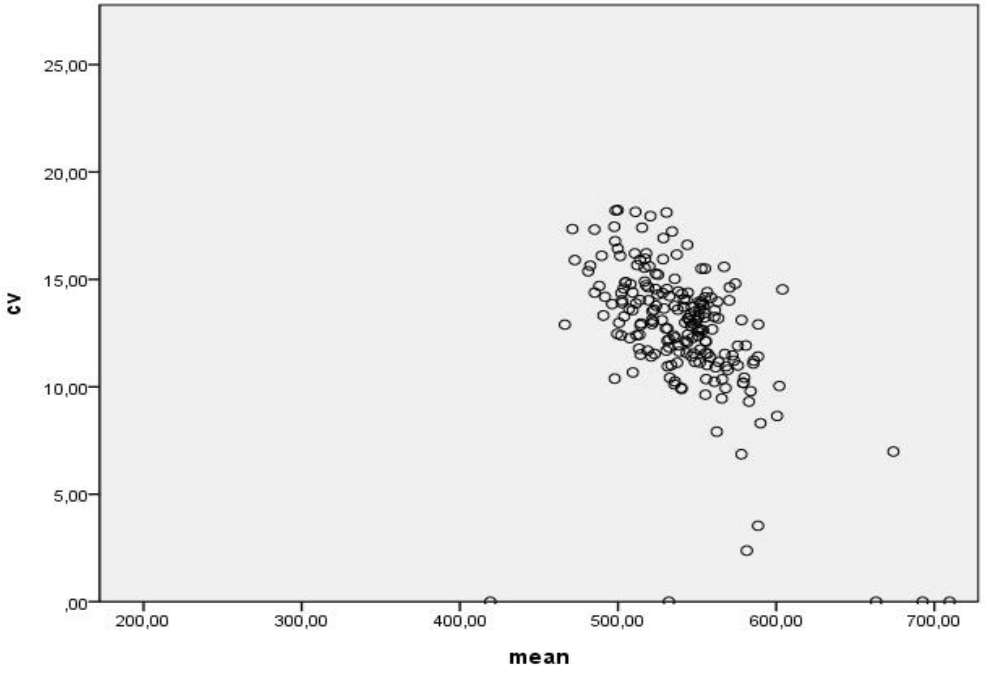


Sirontakuvioesimerkki PISA 2006sta (kirjastani):

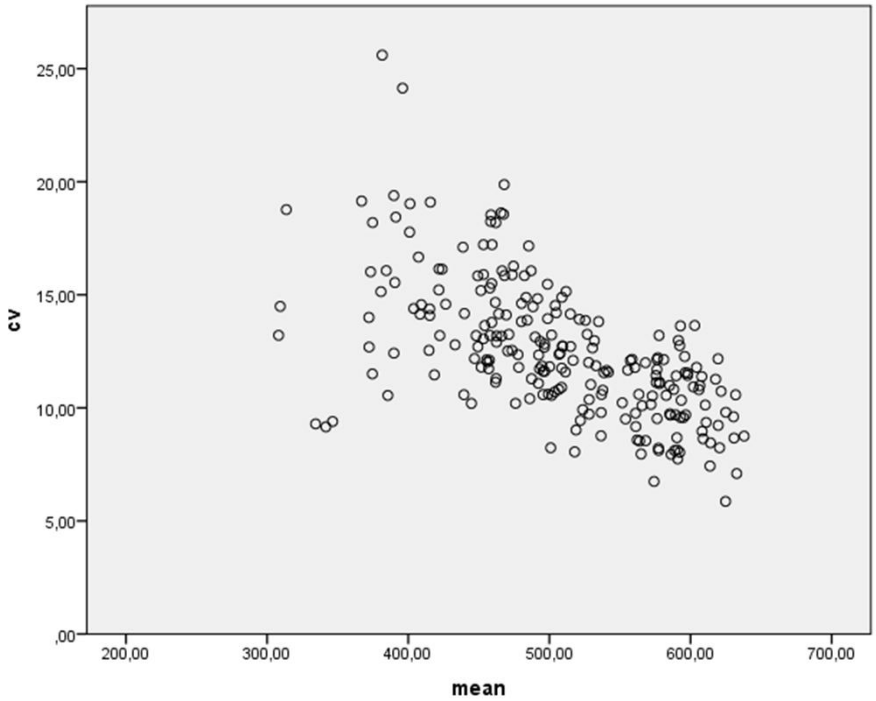
Kuvioissa vaaka-akseli mittaa matemaattis-tilastollista lukutaitoa ja pystyakseli suhteellista keskihajontaa eli vaihtelukerrointa *CV*.

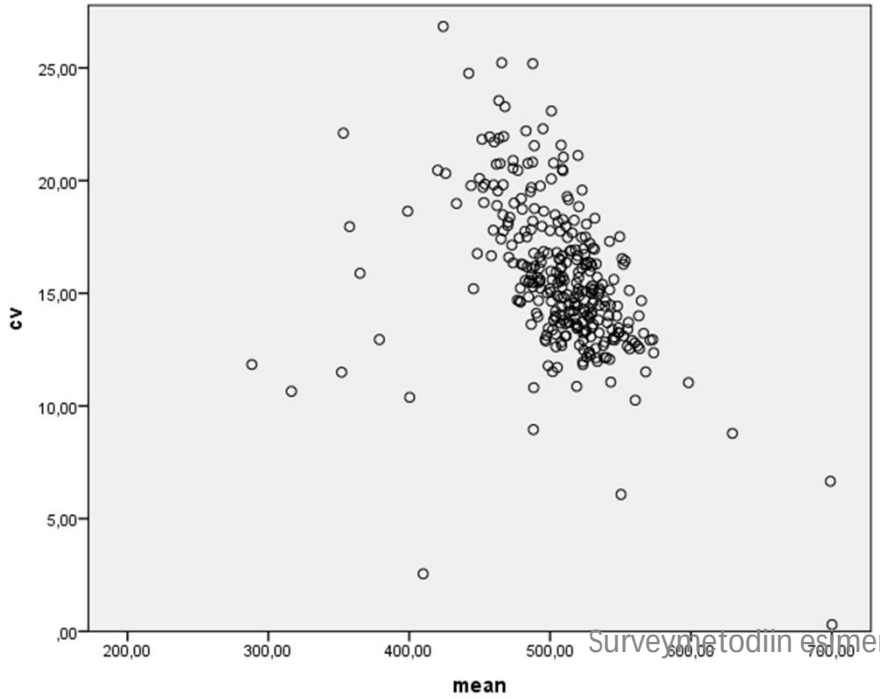
Tulkitse kun kerron että havainnot ovat PISA:n otoskouluja. Siron-takuvion hyödyllisyys näkyy suoraankin mutta voidaan myös tuottaa mittareita.





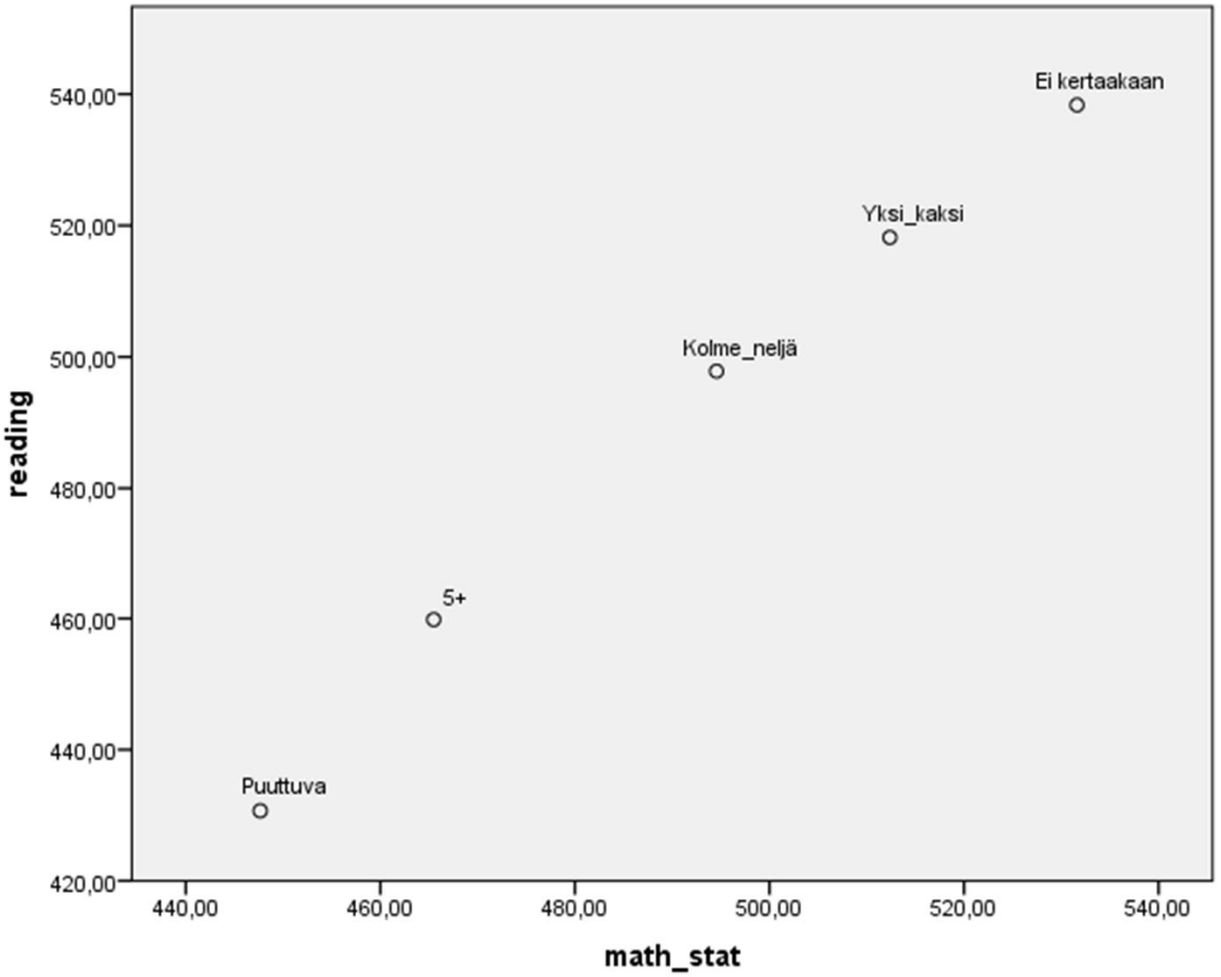
Sirontakuvioesimerkki PISA 2009sta: Ylempi on Saksa, alempi Suomi





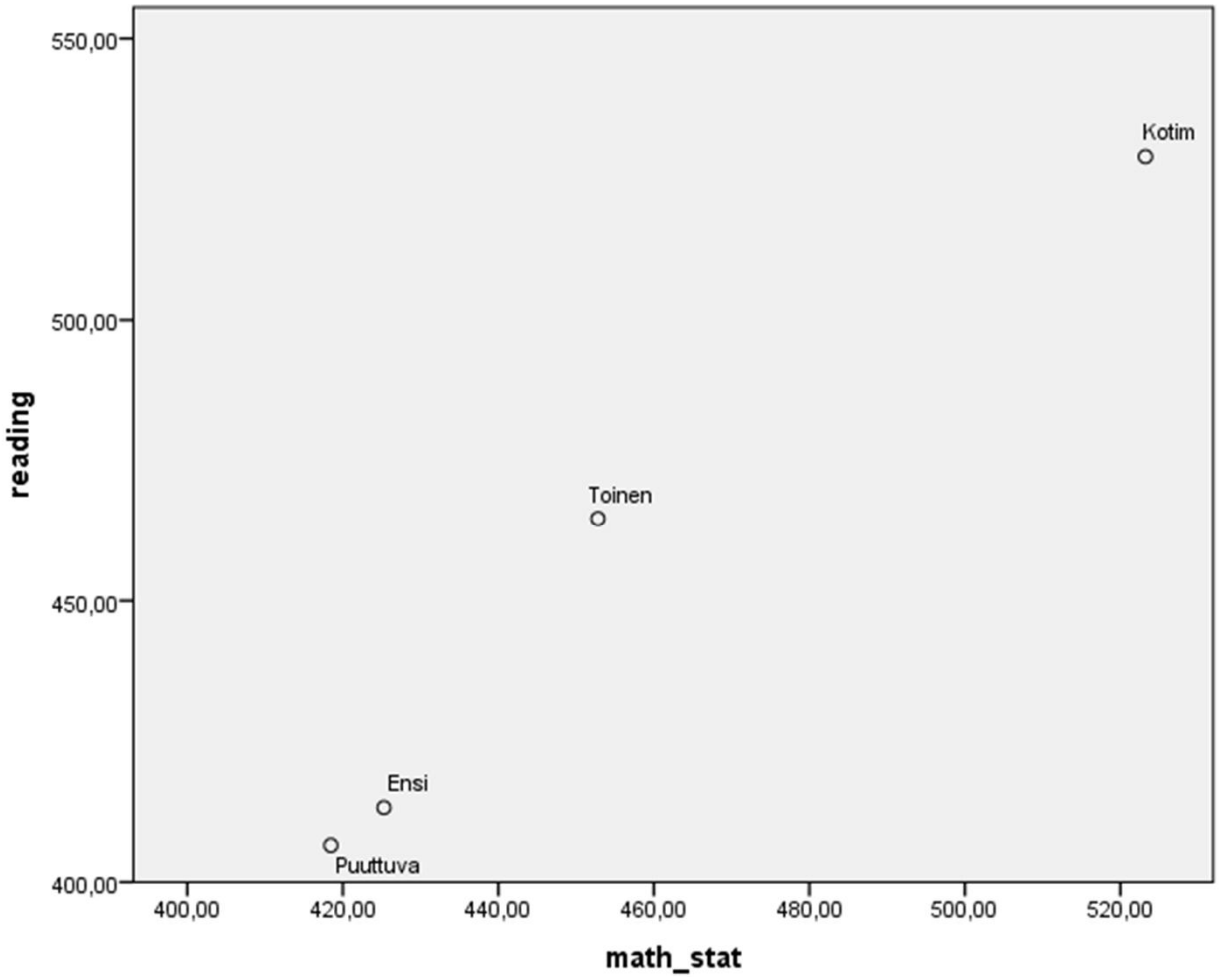
Nyt olemme jo uusimmassa Pisassa vuodelta 2012.
Samat sirontakuviot kuin edellä ja asteikotkin samat.
Mitä muutoksia havaitset?

PISA 2012 jossa on kysytty myös poissa oloa eli usein ’lintsaamista’ koulusta. Aika monet vastasivat tähän, sillä puuttuva tieto on noin 1,6 prosentilla
Ei kertaakaan lintsanneita oli noin 56 prosenttia. Tutkipa miten lintsaaminen vaikuttaa osaamiseen ja miten osaavia ovat vastaamatta jättäneet?



Tätä kuten seuraaviakin kuvauksia voi kutsua aggregaatti-tason imputoinniksi.

PISA 2012: Tässä samat osaamistiedot kuin edellä mutta kysytty ’juuria’ eli onko kotimainen (native) puhtaasti vai ensimmäisen tai toisen polven maahan muuttaja. Mukana myös puuttuvaa tietoa 1,3 prosentilla: Mistähän se voisi johtua? Valtaosalla eli 95 prosentilla on kotimaiset juuret. Tulkitse tuloksia.

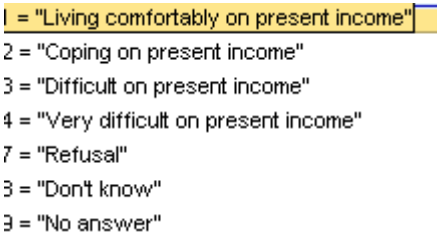


Esimerkki tulotiedon puuttumisesta ESS:ssä

Monet kyselyjen tiedot saatetaan kokea herkiksi eikä niihin vastata ainakaan totuudenmukaisesti. Jos kyselyn merkitsee ylös haastattelija, olkoon se lähellä tai puhelimessa, halutaan antaa itsestä ehkäpä parempi kuva kuin todellisuus on, tai ei haluta esittää liian kriittisiä näkökohtia. Jotkut asiat voivat olla myös inhottavia kertoa kuten vaikkapa väkivallan kokeminen. Jopa tulojen kertominen epäilyttää ihmisiä. Entä sinua? Jotkut ehkä eivät halua kertoa olevansa kovin köyhiä tai kovin rikkaita. Esiintyy siis puuttuvaa tietoa tai väärää tietoa. Puuttuva tieto on tietysti kiusallinen, erityisesti jos sitä on paljon. Olisi kiva saada esimerkiksi puuttuvien tuloista jokin käsitys kuitenkin. Tähän tarjoaa mahdollisuuden imputointi. Sen edellytyksenä on hyvien aputietojen olemassa olo eli sellaisten muuttujien joissa ei ole tai on melko vähän puuttuvuutta. Toiseksi näillä muuttujilla olisi hyvä olla yhteyksiä (korrelaatiota) puuttuviin. Imputointia en nyt kokeile mutta havainnollistan puuttuvuuden yhteyksiä aggregatiivisesti. Tällöin otetaan sopivia apumuuttujia ja katsotaan näiden aggregaatteja sekä tiedetyissä tuloryhmissä että myös puuttuvissa. Yksinkertaisin aggregaatti on keskiarvo joka on esimerkissä käytössä.

Esimerkki tulotiedon puuttumisesta ESS:ssä

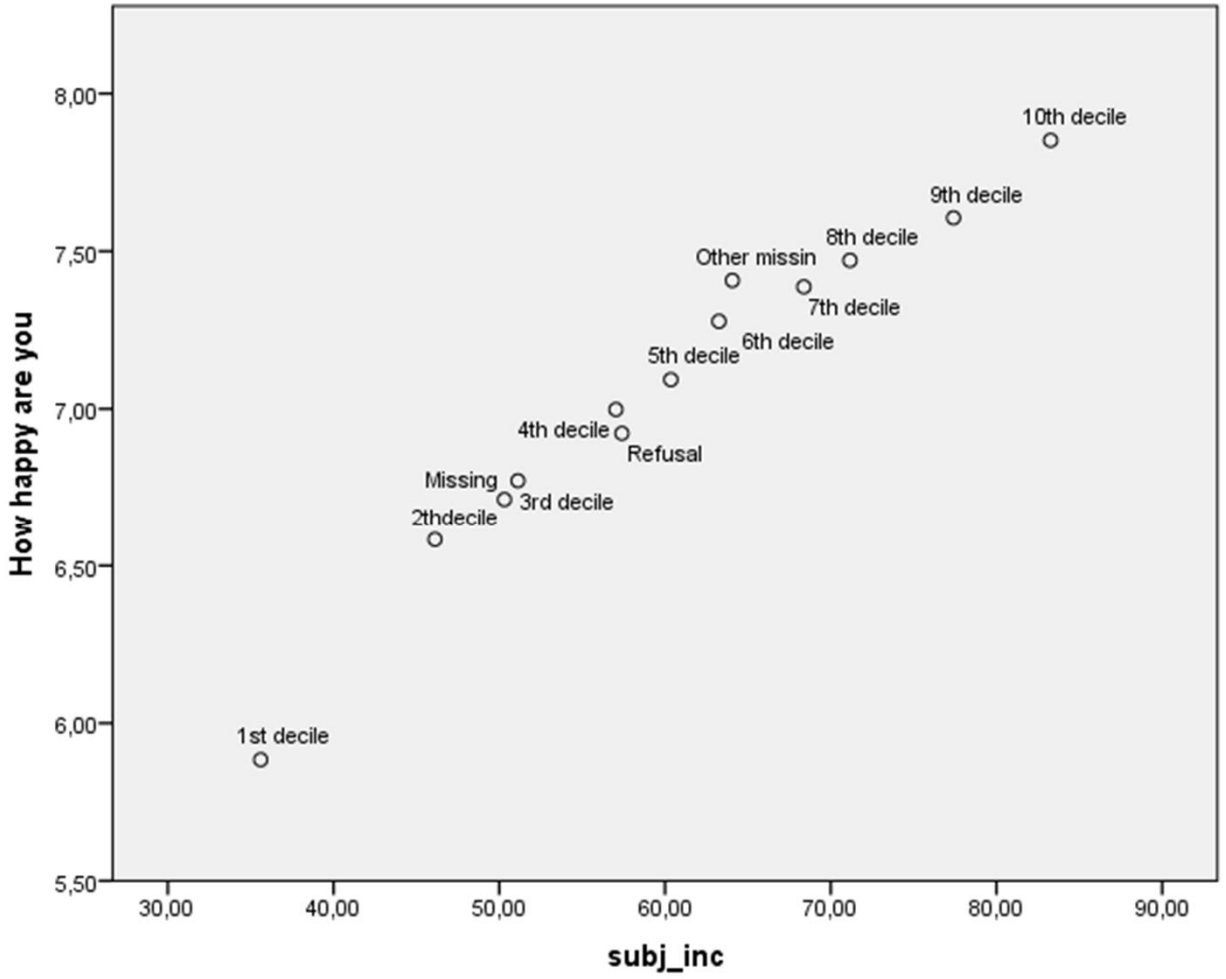
Otan ensin seuraavat kaksi apumuuttujaa samasta ESS-aineistosta:
- Subjektiivinen tulo (Feeling about household's income nowadays) jonka kategoriat ovat:
- Onnellisuus asteikolla [0, 10]



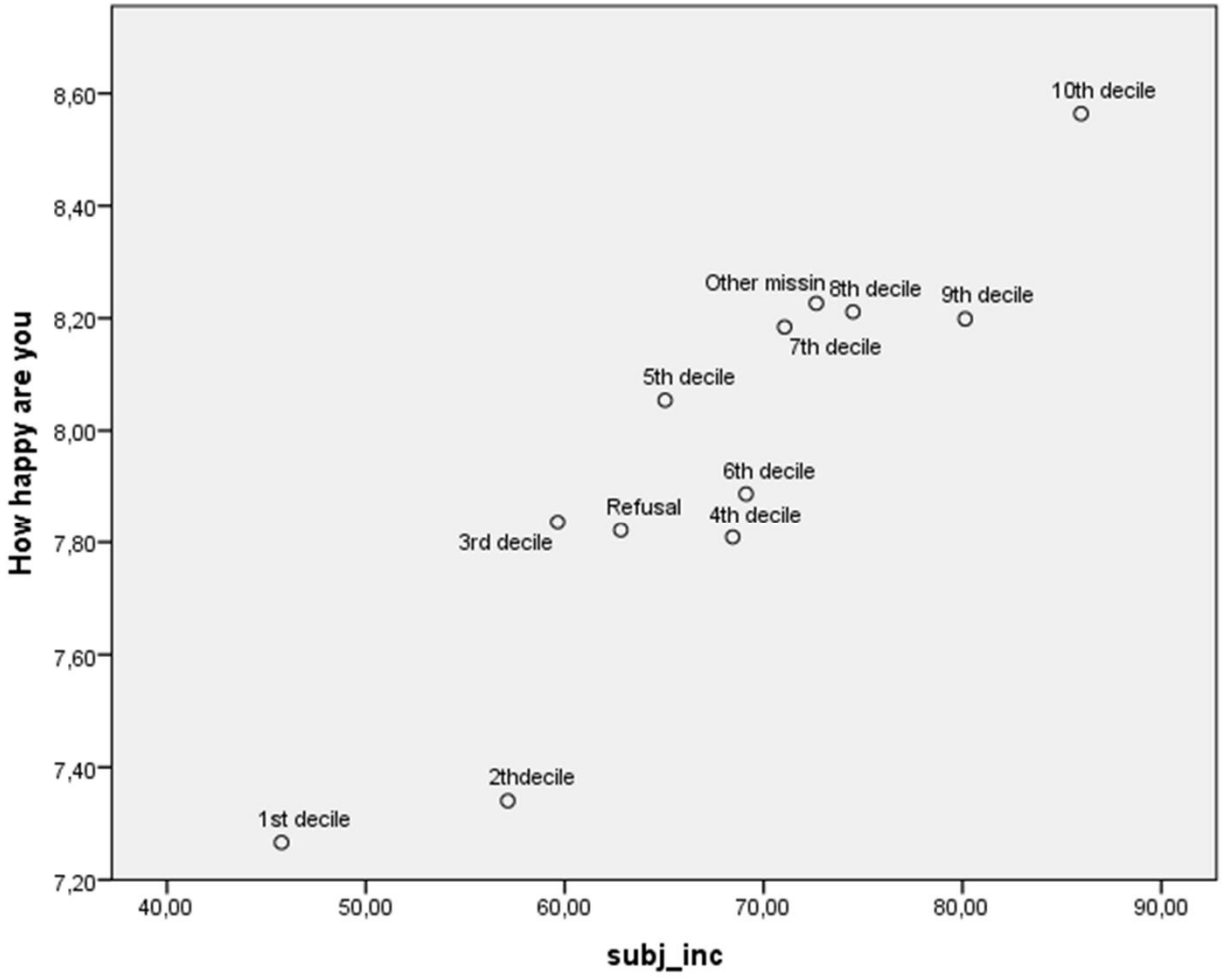
Tein subjektiiviselle tulolle = Subj_inc muunnoksen välille [0, 100], jolloin siis ensimmäisen muuttujan kategoria = 1 tulee 100:ksi jne. Kiinnostavaa on, että subjektiivisen tulon kertominen ei ole yhtä herkkää kuin objektiivisen eli siinä on vähemmän puuttuvuutta.
Nyt on helppo laskea tuloaggregaateille keskiarvot kummastakin muuttujasta.

Jotta tulos havainnollistuisi paremmin, kannattaa tehdä sirontakuvio (scatter plot). Huomaa että noissa apumuuttujissa on hieman puuttuvuutta eli aggregaatissa ne arvot jäävät pois. Ennen tätä kuitenkin maakeskiarvot subjektiiviselle tulolle.

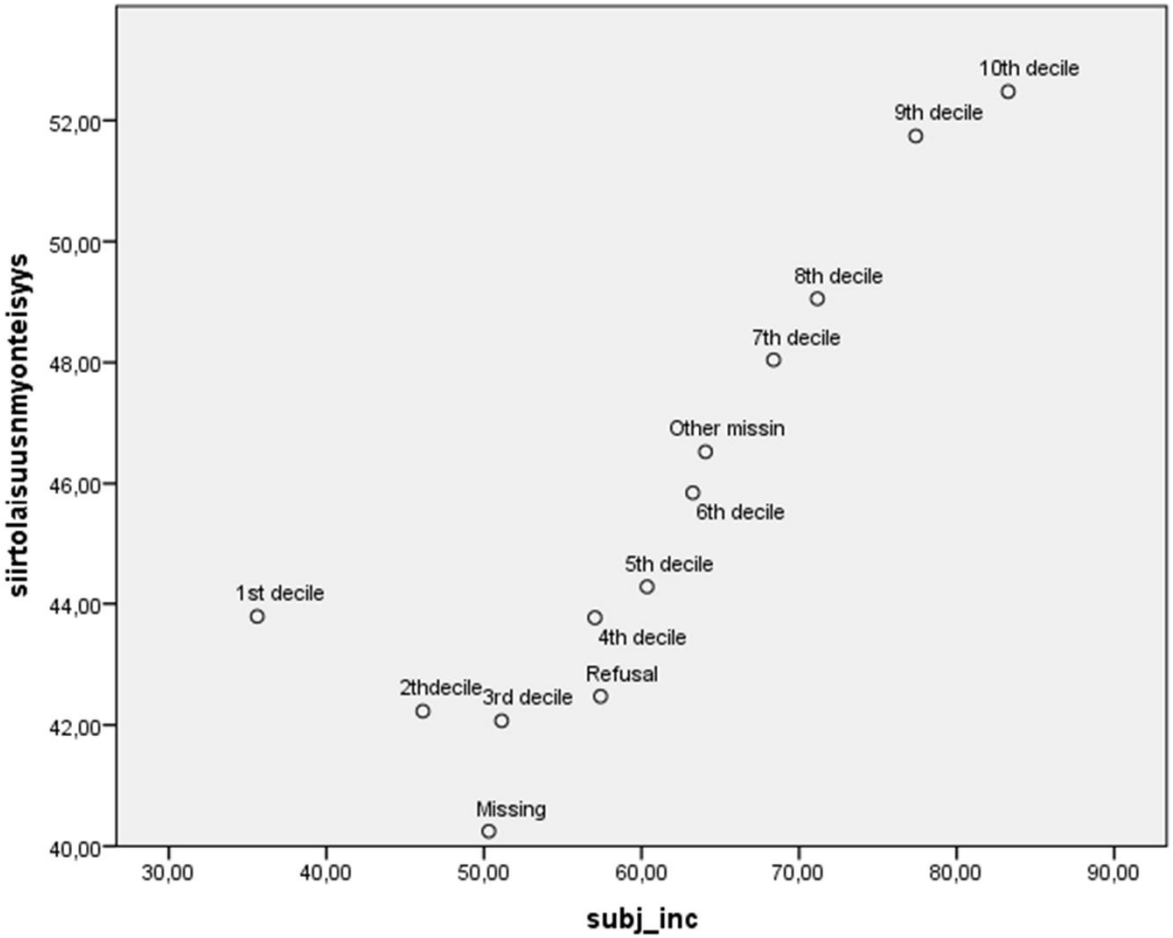
Tässä on sirontakuvio kaikkien maiden aineistolle. Katso erityisesti puuttuvien tietojen koodeja ja mieti minkälaisia tuloja niiden enemmistöllä on?



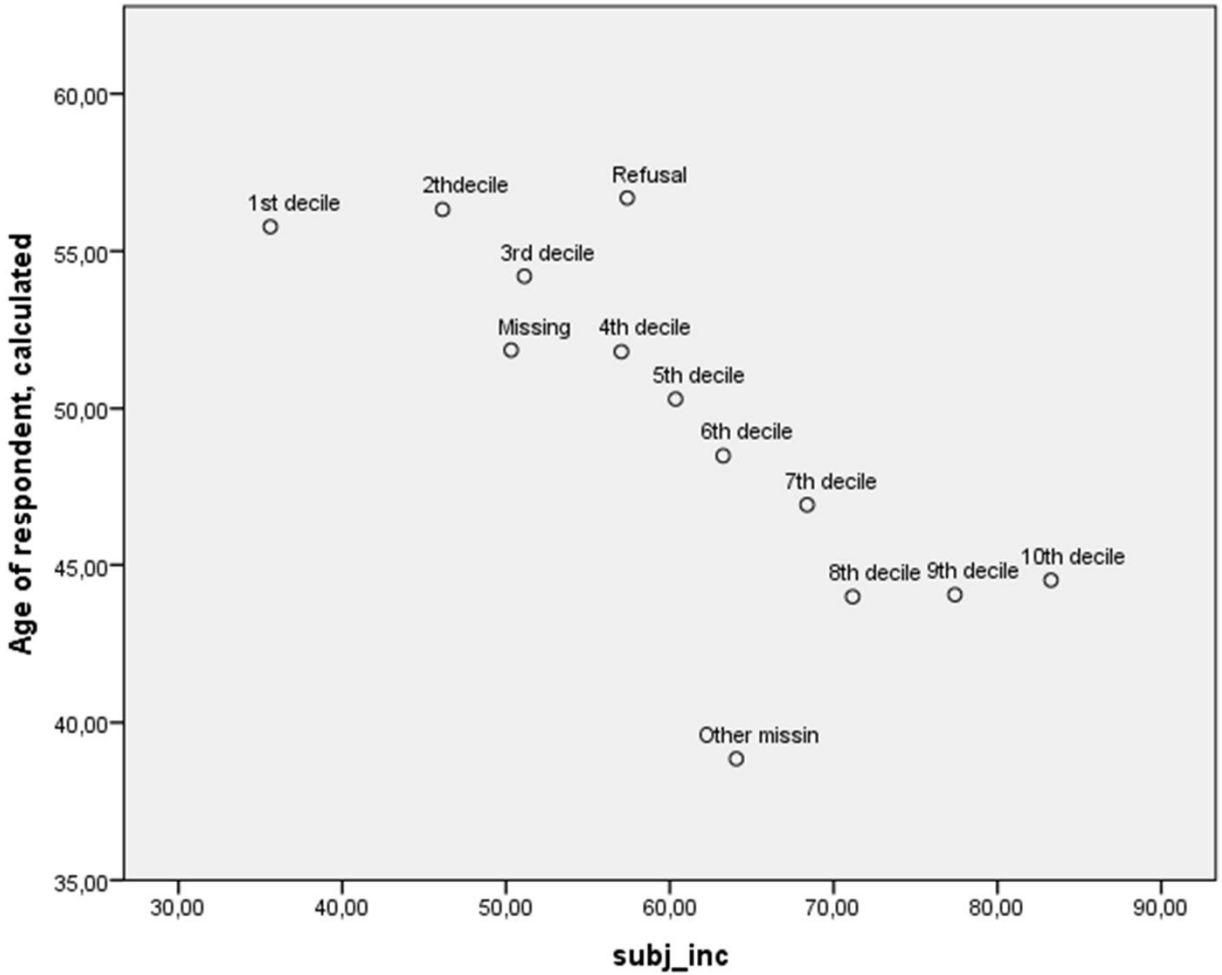
Tässä taas on sirontakuvio Suomen aineistolle.
Molempien perusteella nähnet, että subjektiivinen tulo erityisesti ala- ja yläpäässä on vahvasti objektiiviseen tuloon korreloituva.
Onnellisuuden osaltakin korrelaatio on saman suuntainen.



Jotta asiaan saataisiin vielä toisenlaistakin valoa, niin otin toiseksi muuttujasta siirtolaisuusmyönteisyyden. Subjektiivinen tulohan on vahvasti objektiiviseen liittyvä, mutta muita muuttujia olisi hyvä etsiä myös. Tämä on kaikille maille.



Jotta asiaan saataisiin vielä kolmannenlaista valoa, niin otin toiseksi mukaan ikäkeskiarvon. Tämä on kaikille maille. Se paljastaa ainakin sen, että Other missing –porukka on melkoisen nuorta, mutta Refusal sekä alimmat desiilit taas aika vanhaa.



Noita viimeksi mainittuja näkökulmia ’laajentamalla’ päästään imputointiin jossa puuttuva tieto tuotetaan yksittäisille havainnoille.
Jos halutaan jatkuvan muuttujan tutkimista, on vain kaksi vaihtoehtoa:

- Jättää puuttuvat pois tai
- Imputoida

Jos tarvitaan kategorinen muuttuja täydellisenä, se voidaan imputoida mutta ei aina ole helppoa. Sen sijaan helppoa on tehdä puuttuvista yksi tai useampia kategoria, mikä on tavallaan imputointia. Silloin voi esimerkiksi regressiomallin tehdä tällainen muuttuja selittäjänä, ja yrittää tulkita mitä tulos tarkoittaa. Usein saadaan melko helposti kohtuullinen tulkinta. Seuraavalla sivulla esimerkki ESS:stä.

Selitettävänä on Onnellisuus = Happy ja selittäjänä myöskin maa mutta ei tulosteessa

gndr 1	-0.127372999	0.01651464	-7.71	<.0001
gndr 2	0.000000000	.	.	.
agea	-0.057564802	0.00243735	-23.62	<.0001
agea2	0.000514351	0.00002453	20.97	<.0001
income 1	-1.064678039	0.03830525	-27.79	<.0001
income 2	-0.703135415	0.03696476	-19.02	<.0001
income 3	-0.270233898	0.03712170	-7.28	<.0001
income 4	-0.196200684	0.03747956	-5.23	<.0001
income 5	-0.134556279	0.03789157	-3.55	0.0004
income 6	0.090481208	0.03880170	2.33	0.0197
income 7	0.133832047	0.03919103	3.41	0.0006
income 8	0.203103336	0.04030674	5.04	<.0001
income 9	0.259530311	0.04181735	6.21	<.0001
income 10	0.496656603	0.04165576	11.92	<.0001
income Other missing	-0.166133572	0.03932490	-4.22	<.0001
income Refusal	0.000000000	.	.	.
eisced	0.062963940	0.00504154	12.49	<.0001

Tässä tunnet muut muuttujat paitsi ’eisced’ joka on koulutusaste ja tässä jatkuvana muuttujana jona toimii kohtuullisesti koska asteita on 7.

Tämä on SAS-tuloste