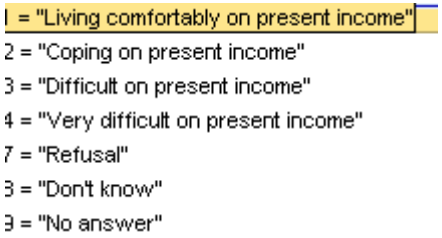
Surveymetodiikka 2013
Seppo
Esimerkki tulotiedon puuttumisesta ESS:ssä

Olemme havainneet että tulotiedossa on puuttuvuutta, mutta vähemmän kuin ESS:n alkuaikoina. Silti sitä esiintyy melko paljon.
Puuttuva tieto on tietysti kiusallinen, erityisesti jos sitä on paljon. Tulojen osalta sitä on siis enemmän kuin useiden muiden muuttujien osalta.
Olisi kiva saada 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.

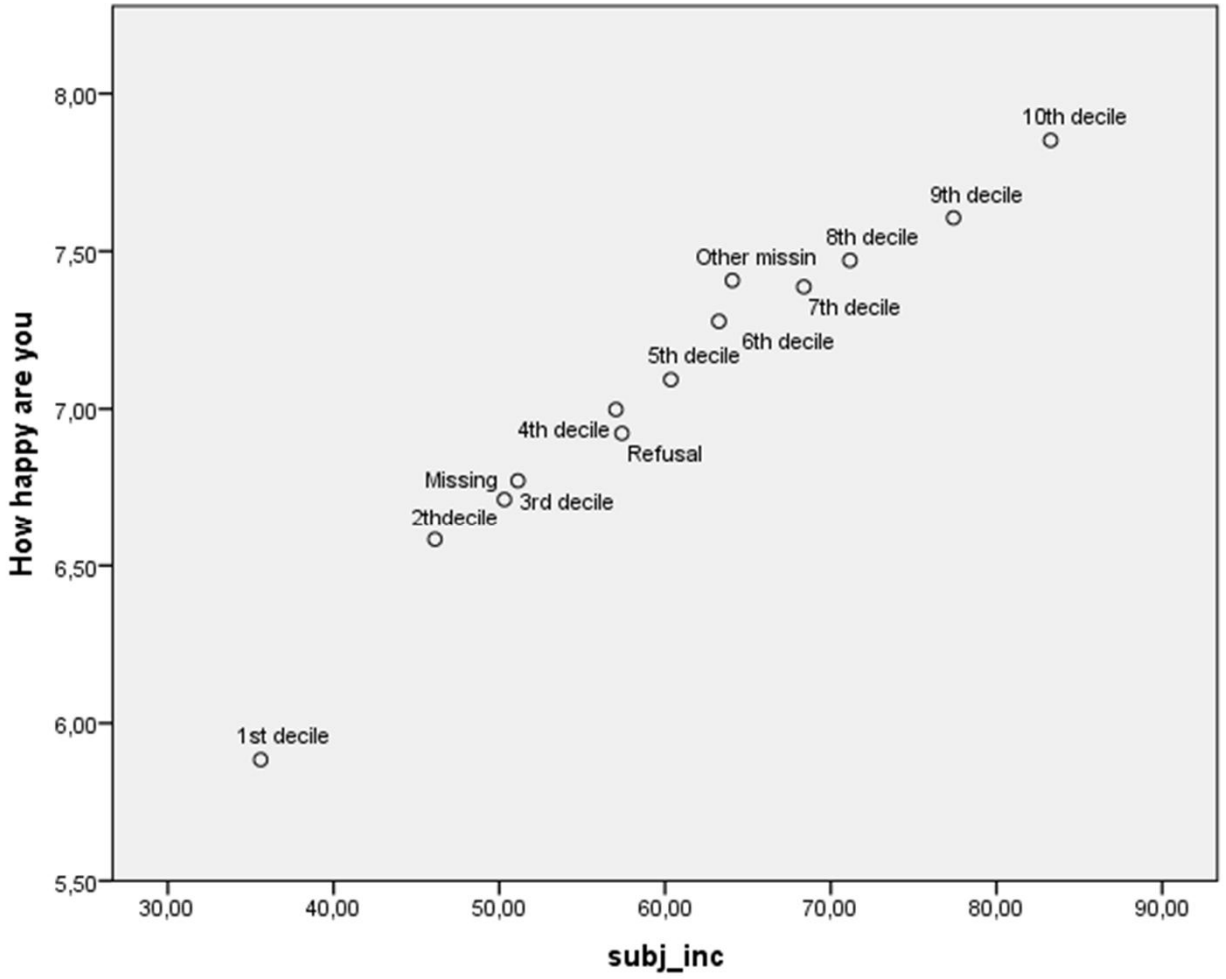
Esimerkki tulotiedon puuttumisesta ESS:ssä

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

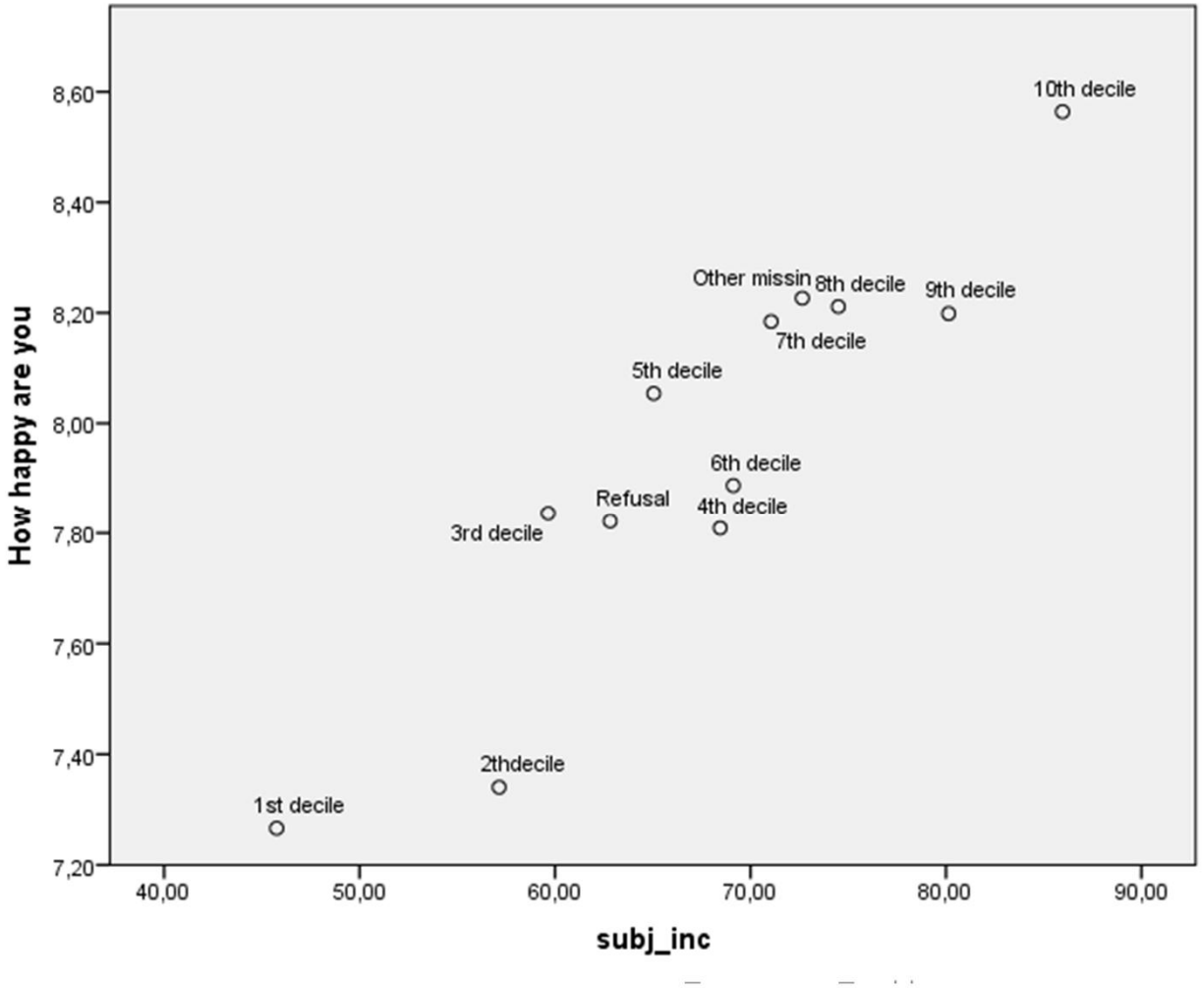


Kuten saatat arvata, tein subjektiiviselle tulolle = Subj_inc muunnoksen välille [0, 100], jolloin siis ensimmäisen muuttujan kategoria = 1 tulee 100:ksi jne.
Nyt on helppo laskea tuloaggregaateille keskiarvot kummastakin muuttujasta.
Jos innostut, voit kokeilla itsekin.
Jotta tulos havainnollistuisi paremmin, kannattaa tehdä sirontakuvio (scatter plot) siten kuin on kahdella seuraavalla sivulla.
Huomaa että noissa apumuuttujissa on hieman puuttuvuutta eli aggregaatissa ne arvot jäävät pois.

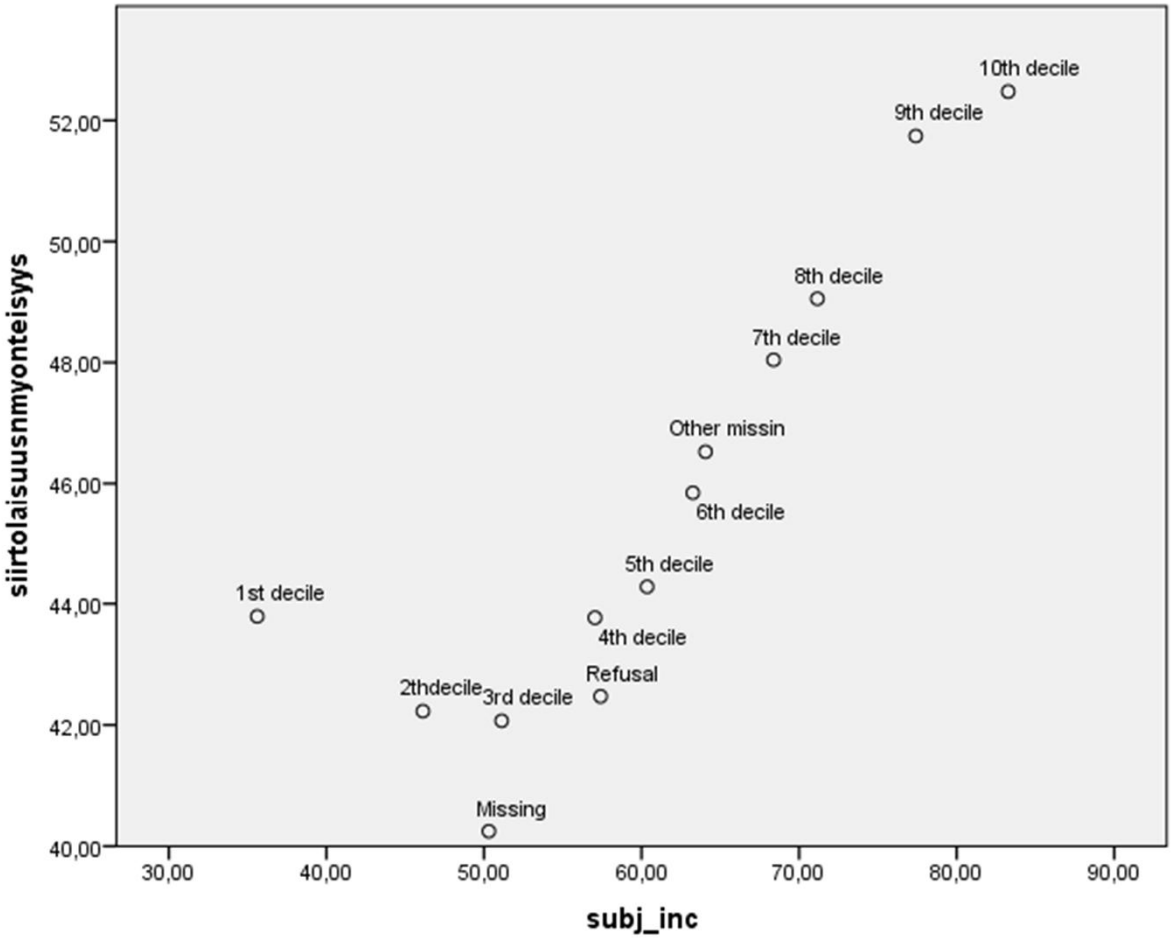
Tässä on sirontakuvio kaikkien maiden aineistolle. Katso erityisesti puuttuvien tietojen koodeja ja mieti minkälaisia tuloja niiden enemmistöllä on?



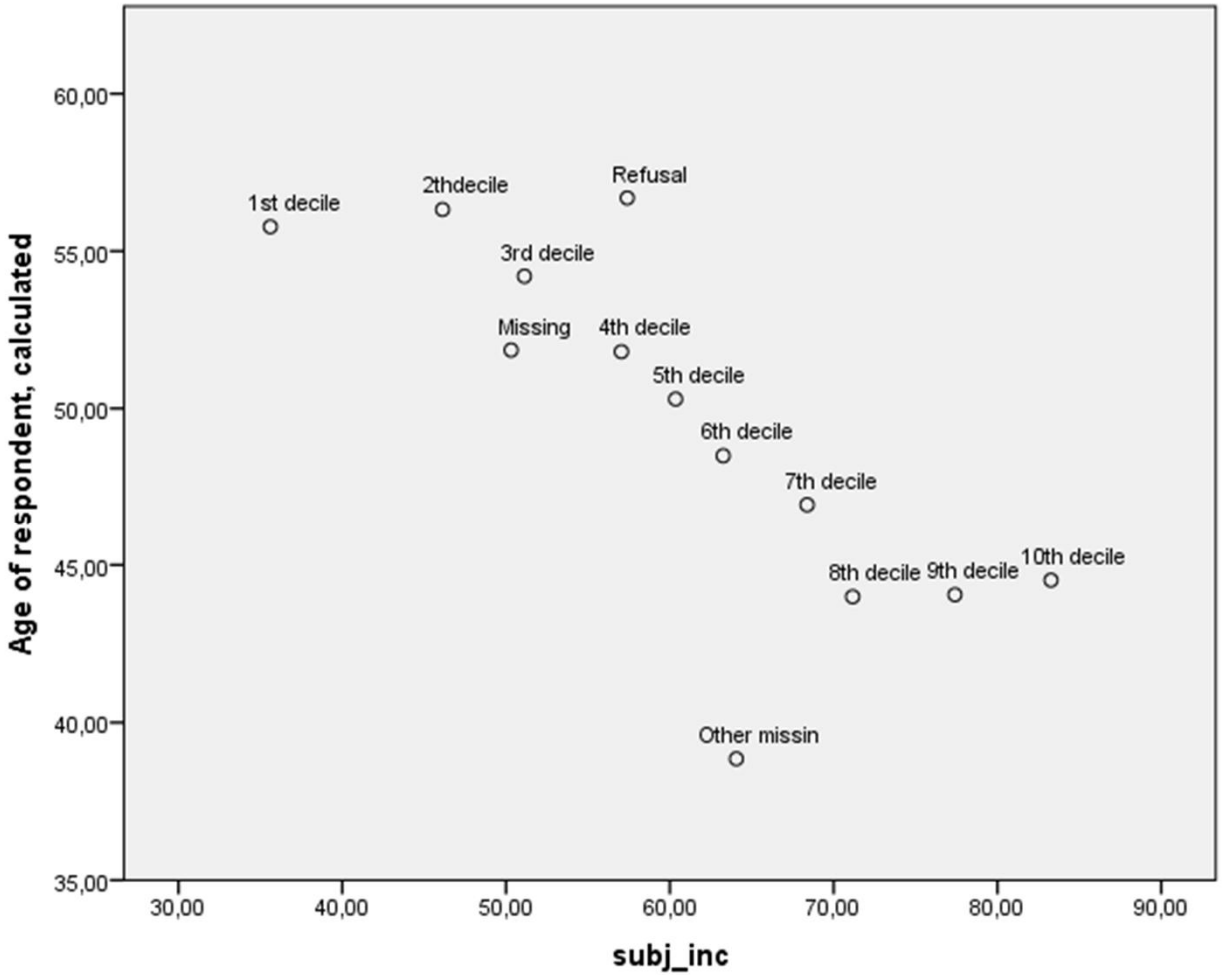
Tässä taas on sirontakuvio Suomen aineistolle.
Molempien perusteella nähnet, että subjektiivinen tulo erityisesti ala- ja yläpäässä on vahvasti objektiiviseen tuloon korreloituva.
Onnellisuuden osaltakin korrelaatio on saman suuntainen.



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.



Tämä esimerkistö on siis johdantona Imputointiin.

Siinähän rakennetaan imputointimalli, ja yksi vaihtoehto mallille on selittää imputoitavaa muuttujaa eli tuloja (objektiivista) sopivilla apumuuttujilla. Subjektiivinen tulo havaittiin mainioksi. Sen ainoa haittapuoli on, että siinäkin on puuttuvuutta muttei niin paljoa kuin objektiivisessa. Siten se auttaisi ainakin imputoimaan osan havainnoista. Lisäksi kannattaisi ottaa muita muuttujia kuten esimerkeissä. Kun mahdollisimman hyvä imputointimalli on saatu, voidaan toteuttaa itse Imputointi (Imputointitoiminto). En tässä sitä tarkemmin esitä. Katso kirjasta ainakin lähestymistapa vaikket kaikkia yksityiskohtia.