SOSIAALITUTKIMUKSEN TILASTOLLISET MENETELMÄT 17.1.–28.2.2012. Jakso Regressioanalyysi ja trendimäisten muuttujien muunnokset. Luennoi: yliopistonlehtori Pekka Pere.

Olin pimeässä työkaluvajassa. Ulkona paistoi aurinko, ja oven yläreunan halkeamasta tuli auringonsäde. Paikasta missä seisoin valonsäde ja siinä leijuvat tomuhiukkaset näyttivät ehdottomasti huomionarvoisimmalta asialta koko tilassa. Kaikki muu oli melkein pikimustaa.

Sitten siirryin niin, että säde osui silmiini. Välittömästi kaikki äsken näkemäni hävisi. En nähnyt työkaluvajaa enkä (ennenkaikkea) sädettä. Sen sijaan näin vihreitä lehtiä puun oksilla heilumassa ja sen takana yhdeksänkymmenen miljoonan mailin päässä auringon oven yläreunan halkeamalla kehystettynä. Sädettä pitkin katsominen ja säteen katsominen ovat aivan eri kokemuksia.[1]

Regressio odotusarvoa kohti

Francis Galtonin ensimmäinen regressio vuodelta 1877 on kuviossa ohessa ("herneen siemen -vanhemmat" ja "herneen siemen -jälkipolvi").[2] Oleellisesti sama ilmiö on seuraavassa kuviossa, joka kuvaa Galtonin havaitsemaa vastaavaa yhteyttä vanhempien pituuksien (mahdollisesti painotetun) keskiarvon (midparent[3]; "keskipituus") ja heidän lastensa pituuden (children) välillä.[4] Kuviosta nähdään, että vanhempien keskipituus on ollut keskimäärin runsas 68 tuumaa. "Midparents"-suora kuvaa vanhempien keskipituuden poikkeamaa keskimääräisestä pituudesta (suoran kulmakerroin on yksi). Kuvion mukaan keskimääräistä

- pidempien vanhempien lapsi on myös keskimääräistä pidempi muttei yhtä paljon kuin vanhempansa (suoran "children" kulmakerroin on 0:n ja 1:n välillä).
- lyhyempien vanhempien lapsi on myös keskimääräistä lyhyempi muttei yhtä paljon kuin vanhempansa.
- pituus "regressoituu" (palautuu, taantuu) eli pyrkii palaamaan kohti odotusarvoansa (yllä runsas 68 tuumaa). ("Regression toward the mean".)

[1] C.S Lewis: Meditation in a Toolshed. Kirjassa *Essay Collection.* S. 607. Lainattu kirjassa Michael Ward (2008): Planet Narnia. Oxford University Press. (S. 17.) Suomennos luennoitsijan.

[2] Kuvio on artikkelista Nicholas Gillham (2009): Cousins, Charles Darwin, Sir Francis Galton and the Birth of Eugenics. *Significance*, 6, 132–135.

[3] Midparent-käsitteen määritelmä löytyy Wikipediasta (http://en.wikipedia.org/wiki/Midparent; viitattu 16.4.2011).

[4] Kuvio on kirjasta Michael O. Finkelstein (2009): *Basic Concepts of Probability and Statistics in the Law.* Springer (s. 128). Kuviossa kahdesti esiintyvä "mild-parent" on painovirhe.

Ohessa on myös Galtonin alkuperäinen kuvio vuodelta 1886. Se on paljon epäselvempi ja esitetään kuriositeettina.[5]

Regressio odotusarvoa kohti ei rajoitu perinnöllisyyteen liittyviin tilanteisiin. Ilmiö on hyvä pitää mielessä aina, kun verrataan samantapaisia muuttujia edellisten kuvioiden tapaan. Esimerkiksi oletetaan, että verrataan sosiaalitukia saavien (tai rikoksentekijöiden tai avioerojen jne.) lukumäärää suomalaisissa kaupungeissa vuosina 2011 ja 2010 (lukumäärät kussakin kaupungissa vuonna 2010 mitattuna x-akselilla ja vuonna 2011 y-akselilla) ja että molempien poikkeamat (vakioksi oletetusta yhteisestä) odotusarvostaan johtuvat vain satunnaisvaihtelusta. Tällöin poikkeuksellisen suuri sosiaalitukea saavien määrä tietyssä kaupungissa tasoittuu lähemmäksi odotusarvoa seuraavana vuonna. Vastaavasti tavanomaista pienemmästä sosiaalitukea nauttivien lukumäärästä vuonna 2010 mahdollisesti ilahtuneet kaupunginjohtajat joutuvat tyypillisesti huomaamaan, että sosiaalitukea haetaan vuonna 2011 edellistä vuotta enemmän. Tämä on äärimmäinen tilanne regressiosta odotusarvoa kohti: Kuvitteelliseen kuvioon piirretyn regressiosuoran kulmakerroin olisi nolla. Muuttujilla ei ole mitään yhteyttä. Poikkeamat odotusarvosta pyrkisivät keskimäärin "korjaantumaan" täysin seuraavana vuonna. Kun tällaista aineistoa tutkiva kuvittelee näkevänsä regressiota on kyse regressiovirhepäätelmästä (regression fallacy). Yhteiskuntatieteilijät edelleenkin joskus ovat hahmottavinaan regressiota tilanteista, joissa sitä ei ole.[6]

Regressioanalyysissa on aina kyse ainakin jossain määrin viimeksi mainitusta tilanteesta: Osa muuttujien välisestä suhteesta johtuu pelkästä sattumasta. Esimerkiksi Galtonin tutkimusaineistossa lasten ja vanhempien pituuksien suhteella on ilmeinen geneettinen selitys, mutta osin lasten pituudet lienevät johtuneet kuitenkin sattumanvaraisista tekijöistä kuten lapsen saamasta poikkeuksellisen ravinteikkaasta ruoasta, kellonajasta, jolloin lapsi on mitattu (aamulla lapsi on pidempi) jne.

Regressioanalyysillä pyritään selvittämään ja erottamaan systemaattinen komponentti muuttujien välisessä suhteessa. Systemaattisen komponentin ei tarvitse olla kulmakertoimeltaan nollan ja yhden välillä tai rajoittua yhteen selittävään tekijään Galtonin esimerkkien tapaan.

[5] Galton kertoi tyttölasten pituudet aineistossaan 1,08:lla, mistä kerrotaan kuviossa englanniksi. Lähde: A. Wachsmuth, L. Wilkinson ja G.E. Dallal (2003): Galton's Bend: A Previously Undiscovered Nonlinearity in Galton's Family Stature Regression Data. *American Statistician*, 57, 190–192.

[6] Esim. Milton Friedman (1992): Do Old Fallacies Ever Die? *Journal of Economic Literature*, 30, 2129–2132.