SOSIAALITUTKIMUKSEN TILASTOLLISET MENETELMÄT 15.1.–26.2.2013. Jakso: Regressioanalyysi. Luennoi: yliopistonlehtori Pekka Pere.

1 Regressio odotusarvoa kohti

Francis Galtonin ensimmäinen regressio vuodelta 1877 on hajontakuviossa ohessa ("herneen siemen -vanhemmat" ja "herneen siemen -jälkipolvi").[1] Oleellisesti sama ilmiö on seuraavassa kuviossa, joka kuvaa Galtonin havaitsemaa vastaavaa yhteyttä vanhempien pituuksien (mahdollisesti painotetun) keskiarvon (midparent; "keskipituus") ja heidän lastensa pituuden (children) välillä.[2] Kuvion tulkinnassa on hyvä huomioida, että toistasataa vuotta sitten ihmissukupolvien pituudessa ei ollut (mainittavaa) eroa. 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).
- keskimääräistä 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* tai *regression to the mean.*) Pitkien vanhempien lapset ovat keskimääräistä pidempiä ja lyhyempien vanhempien lapset keskimääräistä lyhyempiä, mutteivät yhtäpaljon pidempiä tai lyhyempiä keskipituuteen nähden kuin vanhempansa.

Ohessa on myös Galtonin alkuperäinen kuvio vuodelta 1886. Se on paljon epäselvempi ja esitetään kuriositeettina.[3]

Mieleen saattaisi tulla — kuten Galtonille aikoinaan — että regressiosta keskipituutta kohti seuraisi sukupolvi sukupolvelta pituuden vaihtelun pieneneminen niin, että lopulta kaikki olisivat keskipituisia. Niin ei käy, koska lasten pituuksissa on aina sattumanvaraisuutta, vaikka lasten pituus keskimäärin

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

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

[3] 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.

regressoituukin vanhempiensa pituudesta. Galton havainnollisti asiaa oheisella kuviolla vuonna 1901.[4]

Nykypäivään ja yhteiskuntatieteisiin liittyviä esimerkkejä on helppo keksiä. Verrataan sosiaalitukia saavien (tai rikoksentekijöiden, avioerojen jne.) lukumäärää suomalaisissa kaupungeissa vuosina 2013 ja 2012 (lukumäärät kussakin kaupungissa vuonna 2012 x-akselilla ja vuonna 2013 y-akselilla), kun molempien poikkeamat (vakioksi oletetusta yhteisestä) odotusarvostaan johtuvat vain satunnaisvaihtelusta eli kun sosiaalitukia saavien lukumäärässä ei ole trendiä. Tällöin poikkeuksellisen suuri sosiaalitukea saavien määrä tietyssä kaupungissa tasoittuu lähemmäksi odotusarvoa seuraavana vuonna. Vastaavasti tavanomaista pienemmästä sosiaalitukea nauttivien lukumäärästä vuonna 2012 mahdollisesti ilahtuneet kaupunginjohtajat joutuvat tyypillisesti pettymään, kun sosiaalitukea haetaan vuonna 2013 edellistä vuotta enemmän.

Edellä vertailtiin kahden satunnaismuuttujan yhteyttä, kun ne ovat samoinjakautuneita ja niiden hajontakuvioon piirretyn muuttujien välisen systemaattisen komponentin summeeraavan suoran kulmakerroin on (itseisarvoltaan) alle yksi. Regressioanalyysissa (luku 3) voidaan sallia useampia muuttujia, jotka voivat olla erilailla jakautuneita, eikä systemaattisen selityksen suuruutta tarvitse rajoittaa edelliseen tapaan. (Tällöin ei voida puhua regressiosta odotusarvoa kohti aivan samassa mielessä kuin edellä.) Regressioanalyysillä pyritään selvittämään systemaattisuus yhden muuttujan ja muiden muuttujien välillä. Aina regressioanalyysissa on kuitenkin kyse pohjimmiltaan samasta ilmiöstä kuin edellä eli että osa havaintojen käyttäytymisestä on systemaattista ja osa sattumaa. Sattuman vaikutus tulisi regressoida "pois" muuttujien välistä yhteyttä arvioitaessa. Esimerkiksi Galtonin tutkimusaineistossa lasten ja vanhempien pituuksien suhteella on geneettinen (systemaattinen) selitys, mutta osin lasten pituudet johtuvat (tutkijan näkökulmasta) sattumanvaraisista tekijöistä kuten lapsen saaman ruoan ravinnepitoisuudesta tai sairastamista taudeista, äidin pituudesta, ylipäänsä vanhemmiltaan perimistään geeneistä, kellonajasta, jolloin lapsi on mitattu (aamulla lapsi on pidempi) ja niin edelleen.

[4] Kuvio on artikkelista Warren Gilchrist (2012): Galton — A Victorian Worth Celebrating, Significance Web Exclusive (http://www.significancemagazine.org/details/webexclusive/1497449/Galton—A-Victorian-worth-celebrating.html; viitattu 3.2.2013).

2 Regressiovirhepäätelmä

Intuitiivisimmillaan regressio odotusarvoa kohti on kahden samoinjakautuneen muuttujan tilanteessa, kun niiden yhteyden summeeraavan suoran kulmakerroin on alle yhden. Ääritilanteessa muuttujien välillä ei ole mitään yhteyttä: Kuvitteelliseen kuvioon piirretyn summeeraavan sovitteen kulmakerroin on nolla, ja poikkeamat odotusarvosta pyrkivät keskimäärin "korjaantumaan" täysin seuraavassa havainnossa. Kun tällaisessa ilmiössä kuvittelee hahmottavansa kausaalisen selityksen havainnoille, on kyse regressiovirhepäätelmästä (*regression fallacy* tai *Galton's fallacy*).

Yhteiskuntatieteilijät ovat joskus hahmottavinaan kausaalisuutta tilanteista, joissa sitä ei ole.[5] Kuuluisa esimerkki on Northwesternin yliopiston tilastotieteen(!) professori Horace Secrist. Hän julkaisi 1933 massiivisen empiirisen tutkimuksen amerikkalaisten yritysten liikevoittojen kehityksestä 1920–1930. Hän havaitsi, että yritysten, jotka pärjäsivät parhaimmin tai huonoimmin 1920, liikevoitot olivat lähestyneet 1930 kaikkien yritysten liikevoittojen keskiarvoa. Secrist päätteli, että yritykset "keskiarvoistuvat" ajan myötä. Löytönsä korostamiseksi Secrist antoi kirjalleen nimeksi *The Triumph of Mediocracy in Business*. Todellisuudessa yritysten liikevoittojen jakauma ei ollut muuttunut, ja Secristin havainnot selittyvät regressiolla odotusarvoa kohti.[6]

Vaikka ongelma on tunnettu, edelleen julkaistaan vastaavia tutkimuksia. Kahneman (2011; s:t 204–208)[7] kritisoi business-kirjallisuutta, jossa perehdytään menestyneiden yhtiöiden strategioihin, yrityskulttuureihin ja johtamistapoihin. Esimerkkinä hän mainitsee Collinsin ja Porras'aan (2000) kirjan.[8] Kirjan viesti on, että jokaisen toimitusjohtajan, johtajan ja yrittäjän tulisi lukea se, jotta muutkin yritykset osaisivat noudattaa menestyneiden yritysten toimintamalleja ja pärjäisivät. Kahnemanin mukaan Collinsin ja Porras'aan ylistämät yritykset eivät pian tutkimuksen julkaisemisen jälkeen enää pärjänneet juurikaan kilpailijoitansa paremmin. Kahneman viittaa muihin vastaaviin tapauksiin, joissa tutkimuksessa hehkutettujen yritysten kukoistus lopahtaa sen julkaisemisen jälkeen. Regressio odotusarvoa kohti on luonteva tulkinta tällaisille tapahtumille. Ihaillut yritykset olivat erityisen menestyviä tutkimushetkellä lähinnä sattumasta johtuen.

Kahneman (mts. 174) kertoo konkreettisen ja opastavaisen esimerkin, kuinka ihmiset voivat kuvitella kausaalisuutta siellä, missä on pelkkää sattumaa (lyhennetty käännös luennoitsijan):

---

[5] Howard Wainer (2005) kertoo kirjassaan *Graphic Discovery* (Princeton University Press, luku10) lisää esimerkkejä. Ks. myös Milton Friedman (1992): Do Old Fallacies Ever Die? *Journal of Economic Literature*, 30, 2129–2132.

[6] Stephen Stigler (1999) kertoo Secristin tutkimuksesta tarkemmin kirjassaan *Statistics on the Table. The History of Statistical Concepts and Methods* (Harvard University Press, luku 8).

[7] Daniel Kahneman (2011): *Thinking, Fast and Slow.* Penguin Books. Kahneman on psykologi ja taloustieteen nobelisti vuodelta 2002.

[8] Jim Collins ja Jerry I. Porras (2000): *Built to Last: Succesful Habits of Visionary Companies.* Random House Business Books.

Sain yhden elämäni tyydyttävimmistä eureka-kokemuksistani opettaessani Israelin ilmavoimien lentokouluttajille tehokkaan opettamisen psykologiaa. Olin kertonut kouluttajille, kuinka hyvän suorituksen palkitseminen toimii paremmin kuin virheistä rangaitseminen. Yksi vanhemmista kouluttajista arveli, että hyvästä suorituksesta palkitseminen sopii ehkä linnuille muttei hävittäjälentäjäkadeteille: "Olen monesti kehunut kadetteja puhtaasta suorituksesta vaikeassa lentoliikkeessä. Seuraavalla kerralla he järjestään suoriutuvat samasta liikkeestä huonommin. Toisaalta olen monesti huutanut kadetin korvakuulokkeeseen haukkuen häntä huonosta suorituksesta. Ylipäänsä haukkumani kadetit pärjäävät seuraavalla yrityksellä paremmin. Olkaa siis hyvä, älkääkä kertoko meille, että kehuminen toimii ja rangaistus ei, koska asia on juuri päinvastoin."

Edellä opitun perusteella on helppo hahmottaa, että vanhemman kouluttajan kokemukset selittyvät sattumalla: Erityisen hyvin pärjänneen kadetin suoritus regressoitui seuraavalla lennolla kohti odotusarvosuoritustaan ja erityisen heikosti suoriutuneen kadetin suoritus samoin. Kouluttaja virheellisesti liitti muutoksiin kuvittelemansa syy-seuraus -suhteen kehuistaan ja karjumisistaan.

Edelläkuvatunkaltainen ilmiö on "voittajan kirous": Kun suuresta joukosta esimerkiksi työpaikan tai urheilujoukkueen jäsenyyden hakijoita poimitaan suorituksiltaan paras, ei valittu yllä aivan aiempien suoritustensa mukaiseen tulokseen. Mikäli hakija on nuori ja kehittyvä, hän ei paranna tuloksiaan aivan samassa määrin kuin muut kehittyvät samassa ajassa.

Tehtävä! Taulukossa on neljästä kuvitteellisesta helsinkiläisestä sosiaalitoimistosta jaetun tuen määrä euroissa viime vuonna. Sosiaalitoimistot ovat yhtäsuuria monilla muilla kriteereillä mitattuina (henkilökunnan lukumäärä, pinta-ala jne.). Erot jaettujen tukien määrässä johtuvat muun muassa toimistojen sijainnista (esim. kuinka hyvin on julkisten kulkuyhteyksien kautta saavutettavissa), vaihteluista sosiaalisista ongelmista kaupunginosittain ja muista sattumanvaraisista tekijöistä. Tänä vuonna sosiaalitoimistojen jakaman tuen tiedetään kasvavan 10 prosentilla — esimerkiksi yleisen valtakunnallisen taloustilanteen takia. Onko järkevää ennustaa, että myös kunkin sosiaalitoimiston jakama tuki kasvaa tänä vuonna 10 prosenttia? Perustele.[9]

sosiaali-toimisto	**jaettu tuki viime vuonna (milj. e)**	**jaettu tuki tänä vuonna (milj. e)**
1	11	
2	23	
3	18	
4	29	
yhteensä	81	89,1

[9] Tehtävä on mukaelma Kahnemanin esittämästä (mts. 184). Alkuperäislähde olisi Max Bazermanin kirja *Judgement in Managerial Decision Making.*